Christian Urff, Pädagogische Hochschule Weingarten · Konzeptbeitrag, Stand September 2026

Ob generative KI dem Lernen nützt oder schadet, hängt von der Passung ab: Wie viel Lenkung braucht diese Person bei dieser Aufgabe, damit sie mit der KI denkt, statt das Denken abzugeben? Das SKILL-Modell koppelt didaktische Lenkung an Kompetenz. Der Name leitet sich aus dieser Kopplung ab: Kompetenz („skill“) und Lenkung sind nicht getrennt voneinander, sondern miteinander verwoben zu denken. Lenkung wird über drei Hebel beschrieben (Scaffolding, Leitplanken, Kontextualisierung) und kann vom Werkzeug wie von der Lehrkraft ausgehen; drei Stufen (Stark gelenkt, Begleitet, Selbstgesteuert) markieren Orientierungsmarken auf einem Kontinuum. Lehrkräfte können damit Werkzeuge auswählen und Einsatzformen planen, Entwicklerinnen und Entwickler können Lernanwendungen verorten, und für die Forschung formuliert das Modell eine prüfbare Grundannahme.

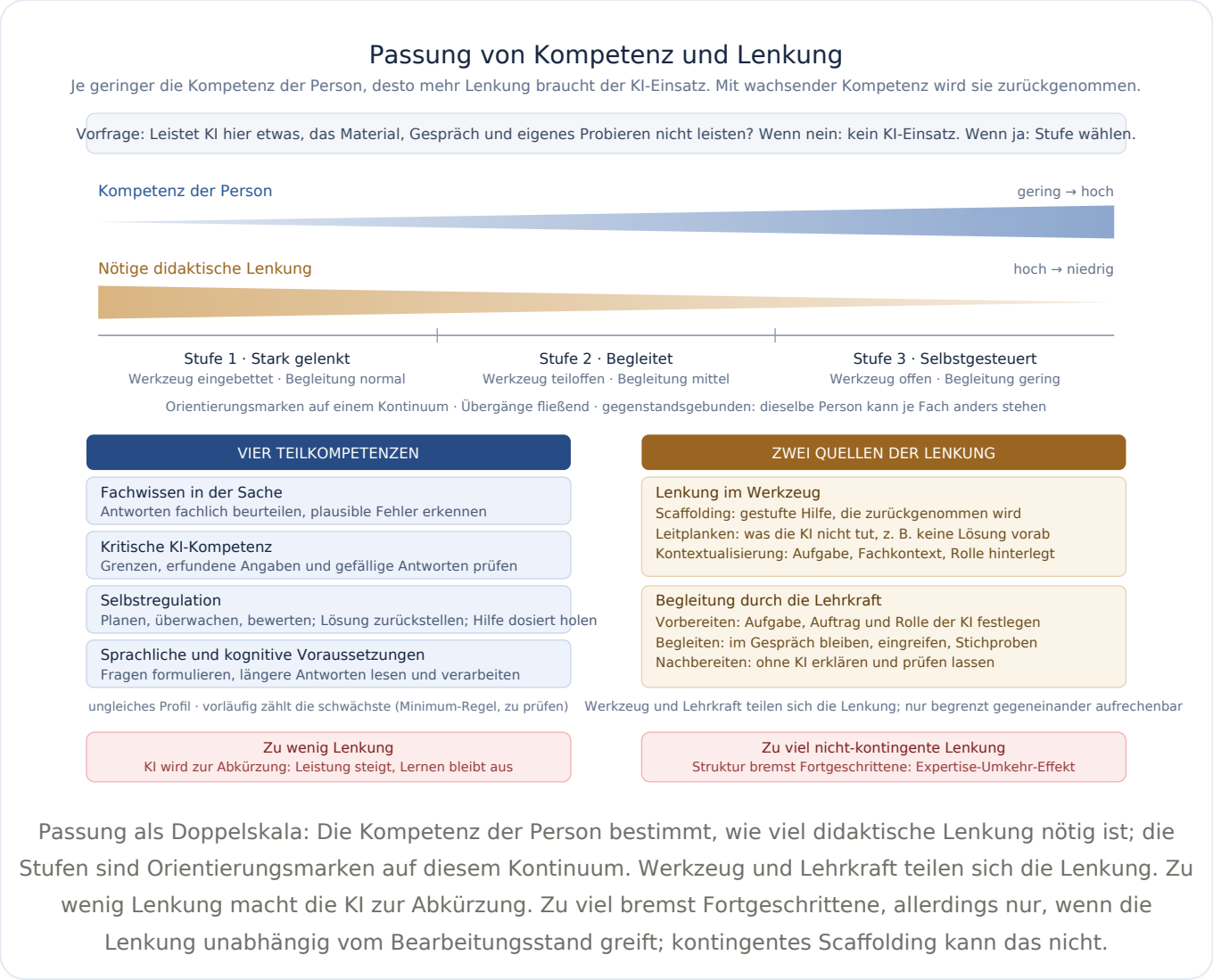
> **Grundidee:** Je geringer die Kompetenz einer Person, KI in einer konkreten Lernsituation lernwirksam zu nutzen, desto mehr didaktische Lenkung braucht der Einsatz: durch Scaffolding, Leitplanken und Kontextualisierung im Werkzeug, durch die Begleitung der Lehrkraft oder durch beides. Mit wachsender Kompetenz wird vor allem die Lenkung zurückgenommen, die unabhängig vom Bearbeitungsstand greift, bis die Person die KI selbstgesteuert als Denkpartner nutzen kann. Gemeint ist dabei nicht Fachwissen allein: Die Fähigkeit, das eigene Verstehen einzuschätzen und den Zugriff auf die Lösung zurückzustellen, entwickelt sich langsamer und ist nicht allein fachgebunden; hohes Fachwissen in einer Sache legt eine höhere Stufe nahe, begründet sie aber nicht. Maßstab bleibt, ob die Person beim Einsatz selbst denkt; der Leitsatz des Modells lautet „Denken mit KI unterstützen, nicht

ersetzen“.

Praktische Bedeutung: Wenn KI eingesetzt werden soll, ist zu klären, welche Stufe für wen bei welcher Aufgabe passt. Die drei Stufen *1 Stark gelenkt, 2 Begleitet, 3 Selbstgesteuert* und die vier Planungsfragen machen die Entscheidung konkret. Die Stufen beschreiben die Passung für eine Person; Werkzeuge heißen eingebettet, teiloffen oder offen. Wer nur planen will, liest die vier Planungsfragen und die Übersicht am Ende.

Geltungsbereich: Das Modell empfiehlt keinen KI-Einsatz. Sein erster Schritt ist die Vorfrage, ob KI in dieser Lernsituation überhaupt gebraucht wird; „kein KI-Einsatz“ ist eine reguläre Antwort des Modells. Häufig ist der Verzicht die bessere Wahl, gerade dort, wo Grundlagen entstehen.

Drei Ebenen: Der Beitrag trennt die theoretische Grundannahme (Kompetenz und Lenkung wirken zusammen auf das Lernen, gemessen ohne KI), die praktische Heuristik (vier Teilkompetenzen, Minimum-Regel, drei Stufen, Planungsfragen) und das Forschungsprogramm (wie Kompetenz und Lenkung zu messen sind und ob die vorhergesagte Wechselwirkung besteht). Die Heuristik ist aus der Grundannahme abgeleitet, aber nicht mit ihr identisch; sie kann sich als verbesserungsbedürftig erweisen, ohne dass die Grundannahme fällt.



Passung als Doppelskala: Die Kompetenz der Person bestimmt, wie viel didaktische Lenkung nötig ist; die Stufen sind Orientierungsmarken auf diesem Kontinuum. Werkzeug und Lehrkraft teilen sich die Lenkung. Zu wenig Lenkung macht die KI zur Abkürzung. Zu viel bremst Fortgeschrittene, allerdings nur, wenn die Lenkung unabhängig vom Bearbeitungsstand greift; kontingentes Scaffolding kann das nicht.

Methode. Der Beitrag ist ein theoretischer Konzeptartikel. Das Modell wurde in einer narrativen, nicht systematischen Synthese entwickelt und an Erfahrungen aus der Entwicklung mathematikdidaktischer Lernanwendungen rückgebunden. Die Literatur wurde bis September 2026 über wissenschaftliche Suchdienste und Rückwärtsrecherche in den Referenzlisten zentraler Arbeiten zusammengetragen. Aufgenommen wurden Metaanalysen und experimentelle Studien zu instruktionaler Lenkung und Scaffolding, experimentelle und quasi-experimentelle Studien zu generativer KI mit einem Lernmaß ohne verfügbare KI, entwicklungspsychologische Arbeiten zu Monitoring und Quellenbewertung sowie einschlägige Rahmenpapiere. Preprints sind als solche

gekennzeichnet und entsprechend gewichtet. Es handelt sich um eine gezielte, nicht um eine erschöpfende Auswahl; das Modell ist empirisch nicht geprüft.

KI-Einsatz. Bei Recherche und Redaktion kamen KI-Werkzeuge zum Einsatz; Auswahl, Prüfung der Quellen und Verantwortung für den Inhalt liegen beim Autor.

Zitiervorschlag. Urff, C. (2026). *Das SKILL-Modell: Wie viel Lenkung braucht Lernen mit KI? Ein Orientierungsrahmen für die kompetenzabhängige Gestaltung des KI-Einsatzes in Schule und Grundschule* (Version 1.0) [Working Paper]. Zenodo. https://doi.org/10.5281/zenodo.22520845 · Als PDF: Urff_2026_SKILL-Modell.pdf

Empirische Ausgangslage

Ein Feldexperiment an einer großen Schule in der Türkei mit knapp tausend Schülerinnen und Schülern der Klassen 9 bis 11 macht das Problem deutlich (Bastani et al., 2025). Klassen wurden per Zufall einer von drei Bedingungen zugeteilt und übten in vier 90-minütigen Sitzungen, die etwa 15 Prozent des im Halbjahr behandelten Mathematikstoffs abdeckten. Wer dabei eine Chatbot-Oberfläche ohne didaktische Einschränkungen nutzen durfte (Bedingung „GPT Base“), schnitt in der Übungsphase um 48 Prozent besser ab als die Kontrollgruppe (bezogen auf deren Leistungsniveau, nicht in Prozentpunkten). In der anschließenden Prüfung ohne KI schnitt dieselbe Gruppe um 17 Prozent schlechter ab als die Kontrollgruppe. Eine zweite Variante desselben Sprachmodells, die per Prompt angewiesen war, von Lehrkräften gestaltete gestufte Hinweise zu geben und Endergebnisse zurückzuhalten, lag in der Übungsphase noch deutlicher über der Kontrollgruppe (etwa 127 Prozent), erreichte in der Prüfung ohne KI aber nur deren Niveau. Beide KI-Gruppen arbeiteten mit demselben Sprachmodell; unterschiedlich war das Antwortverhalten. Das Bündel aus gestuften Hinweisen und gesperrtem Endergebnis verhinderte den Einbruch; welcher Bestandteil dafür ausschlaggebend war, trennt das Design nicht, und die Autorinnen und Autoren

verweisen zusätzlich auf unterschiedliche Gelegenheiten zur Ergebniskontrolle. Einen Lernvorteil gegenüber dem Üben ohne KI belegt die Studie nicht.

Der Befund steht nicht allein. Lehmann et al. (2025) finden in zwei präregistrierten Laborexperimenten keinen Gesamteffekt von Sprachmodellen auf das Lernergebnis. In explorativen Analysen und einer begleitenden Felderhebung hängt die Wirkung davon ab, wie genutzt wird: Wer die KI Lernaktivitäten abnehmen lässt, bearbeitet mehr Themen und versteht das einzelne Thema schlechter; wer sie ergänzend als Tutor nutzt, versteht besser. Wie stark der Effekt ausfällt, hängt vom Vorwissen ab; in der Sprache der Forschung moderiert das Vorwissen den Zusammenhang. Die Arbeit ist ein Preprint, die Befunde zur Moderation sind explorativ.

Ähnlich beim Schreiben. Fan et al. (2025) verglichen bei 117 Studierenden das Schreiben mit ChatGPT, mit einer menschlichen Fachperson, mit einer aus Schreibanalytik abgeleiteten Checkliste und ohne Unterstützung. Die ChatGPT-Gruppe verbesserte ihre Texte am stärksten, unterschied sich im Wissenszuwachs und im Transfer aber nicht von den anderen Gruppen; Häufigkeit und Abfolge ihrer selbstregulativen Prozesse verschoben sich deutlich. Die Autorinnen und Autoren sprechen von metakognitiver Trägheit.

Metaanalytisch fällt die Bilanz zunächst positiv aus. Deng et al. (2025) berichten auf Basis von 69 Artikeln mit experimentellem oder quasi-experimentellem Design positive Effekte von ChatGPT auf die Lernleistung, auf affektiv-motivationale Zustände und auf Denkfähigkeiten höherer Ordnung, einen verringerten mentalen Aufwand und keinen Effekt auf die Selbstwirksamkeit; zugleich weisen sie auf kurze Interventionsdauern, seltene verzögerte Nachtests und methodische Schwächen der Primärstudien hin. Der verringerte Aufwand ist zweideutig: Er kann Entlastung bedeuten oder das Ausbleiben der Anstrengung, die dauerhaftes Behalten erst ermöglicht. Weidlich et al. (2025) zeigen darüber hinaus, dass in dieser Literatur der Beitrag der Technologie regelmäßig mit dem Beitrag der zusätzlich eingeführten

Instruktionsmethode vermischt ist, sodass Sammeleffekte nur begrenzt interpretierbar sind. Eine der derzeit umfangreichsten Metaanalysen findet auf Basis von 68 Studien und 337 Effektstärken einen mittleren Gesamteffekt (SMD = 0,45) bei sehr hoher Heterogenität; stärkere Effekte zeigen sich in kurzen Interventionen mit kleineren Stichproben (Han et al., 2025), das übliche Muster einer Verzerrung durch kleine Studien.

Wo verzögert gemessen wird, kippt das Bild. Barcaui (2025) ließ 120 Studierende ein Thema mit freiem ChatGPT-Zugang oder herkömmlich erarbeiten und prüfte 45 Tage später unangekündigt; in die Auswertung gingen 85 Personen ein. Die ChatGPT-Gruppe erreichte 57,5 Prozent richtige Antworten gegenüber 68,5 Prozent in der Vergleichsgruppe (d = 0,68). Der Ausfall von 35 der 120 randomisierten Personen bis zum Nachtest schränkt die Aussagekraft ein; ob er sich zwischen den Gruppen unterschied, wird nicht berichtet. Effektstärken gelten nach der geläufigen Konvention ab etwa 0,2 als klein, ab 0,5 als mittel und ab 0,8 als groß (Cohen, 1988); für randomisierte Feldstudien im Bildungsbereich sind diese Schwellen zu streng (Kraft, 2020).

Umgekehrt berichten Kestin et al. (2025) aus einem randomisierten Crossover-Experiment mit Zuweisung auf Ebene der Arbeitsgruppen, an dem von 233 Kursteilnehmenden 194 Physikstudierende einer selektiven Universität teilnahmen, für einen gezielt nach Lehr-Lern-Prinzipien gestalteten KI-Tutor höhere Lernzuwächse in kürzerer Zeit als für aktivierenden Präsenzunterricht. Der Tutor gab gestufte Hinweise und keine fertigen Lösungen; getestet wurde jeweils unmittelbar nach der Sitzung, ein verzögerter Behaltenstest fehlt. Die KI-Bedingung fand zu Hause und im selbst gewählten Tempo statt, die Vergleichsbedingung im Präsenzunterricht; Lernumgebung, Tempokontrolle und Medium variieren mit, sodass der Effekt nicht allein dem Tutordesign zuzurechnen ist.

Für die Zuordnung zu einer Stufe gibt die Studie von Kestin et al. wenig her. Ein

themenbezogener Vortest wurde erhoben, geht aber nur in die Berechnung der Lernzuwächse ein und wird nicht als Moderator berichtet; da das Thema für alle neu war, liegen die Werte nahe am Boden. Maße, die sich als Stufenindikator im Sinne dieses Modells nutzen ließen, werden nicht berichtet, weder ein aussagekräftiges Vorwissensmaß noch Selbstregulation oder kritische KI-Kompetenz. Institution und Studienfach taugen dafür nicht: Sie sind bereichsübergreifende Stellvertreter, während das Modell die Kompetenz gegenstandsgebunden bestimmt, und im neu zu erarbeitenden Physikthema war das Fachwissen der Teilnehmenden gering. Der Befund ist mit dem Modell vereinbar, prüft es aber nicht; ob Lenkung Fortgeschrittenen in der Sache schadet, bleibt hier offen. Nach der Instruktionsforschung bremst Lenkung dann, wenn sie unabhängig vom Bearbeitungsstand ausgelöst wird, vorhandenes Wissen redundant wiederholt oder eigene Strategien überschreibt; gestufte Hilfe auf Anforderung fällt nicht darunter.

Zu unterscheiden ist die Menge der Lenkung von ihrer Kontingenz, also davon, ob sie auf den aktuellen Stand der Person reagiert. Damit diese Unterscheidung die Grundannahme nicht gegen Widerlegung abschirmt, wird sie im Abschnitt „Grenzen und offene Fragen“ vorab operationalisiert. Ob ein KI-Einsatz nützt, entscheidet sich an der Gestaltung des Werkzeugs, an der Person, die es nutzt, und daran, was gemessen wird.

Die bildungspolitische Diskussion hat diesen Punkt aufgenommen. Die Ständige Wissenschaftliche Kommission der Kultusministerkonferenz (SWK, 2024) empfiehlt für die Grundschule und den Beginn der Sekundarstufe I weitgehende Zurückhaltung gegenüber offenen Sprachmodellen und eine Konzentration auf die basalen Kompetenzen; die Bildungsministerkonferenz (2024) betont die pädagogische Rahmung jeder KI-Nutzung. Das SKILL-Modell nimmt diese Empfehlungen auf und begründet sie über Kompetenz statt über das Alter. Die Empfehlungen selbst bleiben altersbezogen; die zugrunde liegende Logik ist es nicht. Sie greift überall dort, wo die Voraussetzungen für eine lernwirksame KI-

Nutzung erst entstehen, also auch bei Erwachsenen in einem für sie neuen Fachgebiet. Das Alter behält daneben eigenes Gewicht, wie der Abschnitt zu den theoretischen Grundlagen zeigt.

Unterscheidung von Leistung und Lernen

Lernende können mit KI besser abschneiden und trotzdem weniger lernen. Yan et al. (2025) trennen dafür zwei Größen, die im Alltag leicht verschmelzen: die Leistung während der Nutzung und das, was ohne das Werkzeug bleibt. Lernen zeigt sich erst im zweiten Fall. Salomon et al. (1991) haben dafür früh die Unterscheidung zwischen Effekten *mit* einer Technologie und Effekten *der* Technologie vorgeschlagen: Was die Leistung im Zusammenspiel hebt, muss nicht als Fähigkeit zurückbleiben. Die Unterscheidung ist in der Lernpsychologie alt. Bjork und Bjork (2011) sprechen von wünschenswerten Erschwernissen: Bedingungen, die die aktuelle Leistung senken, aber das dauerhafte Behalten und die Übertragung auf neue Aufgaben (Transfer) erhöhen, weil sie eigene Abrufversuche, Fehler und Umwege erzwingen. Kapur (2016) systematisiert die Befundlage zum produktiven Scheitern und argumentiert, dass ein zunächst erfolgloses eigenes Ringen mit einer Aufgabe das spätere Verstehen fördert, sofern es anschließend im Unterricht aufgegriffen wird. Diese Phase überspringt eine KI, die auf Anfrage die Lösung liefert.

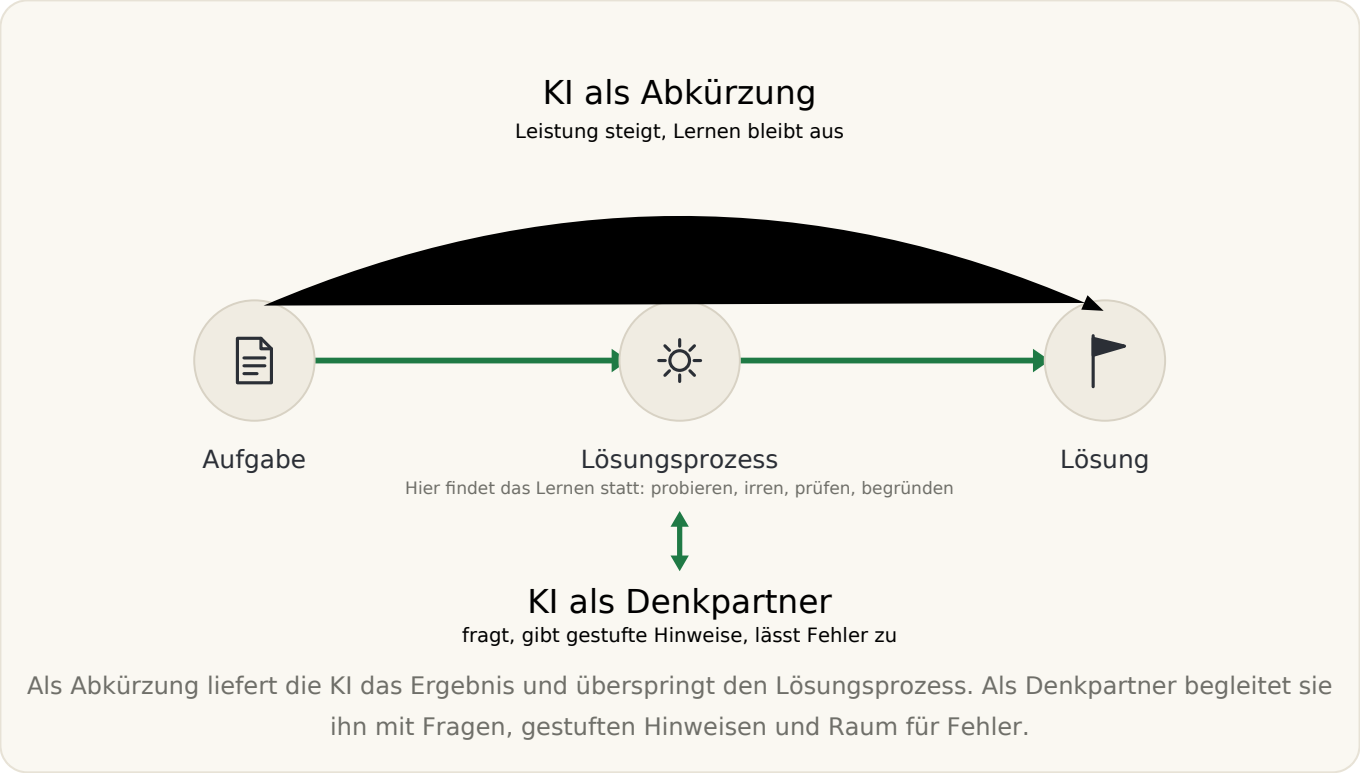
Das Auslagern kognitiver Arbeit ist dabei an sich nichts Neues; Menschen nutzen Notizen, Taschenrechner und Suchmaschinen, um ihr Arbeitsgedächtnis zu entlasten (Risko & Gilbert, 2016). Auch Kinder verfügen ab etwa vier Jahren über solche Strategien und setzen sie bei steigender Aufgabenschwierigkeit vermehrt ein; sie tun dies jedoch weniger selektiv und weniger selbstinitiiert als Erwachsene und lagern teils zu viel, teils zu wenig aus (Armitage & Gilbert, 2025). Problematisch wird es, wenn das Ausgelagerte selbst das Lernziel ist. Eine KI, die den Lösungsweg schreibt, entlastet vom Lernen selbst.

Hinzu kommt ein zweiter Mechanismus: Flüssig formulierte Erklärungen erzeugen ein Gefühl von Verstehen, das mit dem tatsächlichen Können wenig zu tun haben muss. Rozenblit und Keil (2002) zeigen an erwachsenen Stichproben eine Illusion der Erklärungstiefe: Menschen halten ihr Verständnis für vollständiger, als es der Versuch offenlegt, den Zusammenhang selbst auszuformulieren. Die Illusion ist spezifisch für Erklärungswissen und fällt für Fakten, Prozeduren und Erzählungen schwächer aus. Beide Mechanismen bilden eine Kette: flüssige Erklärung, Gefühl des Verstehens, ausbleibende eigene Anstrengung, schwächere Leistung ohne KI.

Für einen Zwischenschritt dieser Kette liegt Evidenz vor: Fisher et al. (2015) zeigen in neun Experimenten, dass bereits das Nachschlagen einer Erklärung im Internet die Einschätzung des eigenen Wissens erhöht; Zugang zu Information wird mit eigenem Verstehen verwechselt, auch dann noch, wenn die Quelle nicht mehr verfügbar ist. Offen ist weniger der Mechanismus als seine Reichweite: ob eine dialogisch erzeugte, auf die eigene Frage zugeschnittene KI-Erklärung stärker wirkt als ein Suchergebnis und wie stark er bei Kindern ausfällt; eine Illusion der Erklärungstiefe zeigen für Geräte und ihre Funktionsweise bereits Zweit- und Viertklässler, nicht dagegen Kindergartenkinder und nicht für Handlungsabläufe, und sie erkennen deren Unzutreffen erst nach dem eigenen Erklärversuch (Mills & Keil, 2004). Belegt ist der Endpunkt der Kette, schwächere Leistung ohne KI (Bastani et al., 2025) und verschobene Selbstregulation ohne Wissenszuwachs (Fan et al., 2025).

Erste neurophysiologische Befunde deuten in dieselbe Richtung: Beim Schreiben mit einem Sprachmodell war die Gehirnaktivität schwächer vernetzt, und die Teilnehmenden konnten kurz danach eigene Sätze nicht korrekt wiedergeben und berichteten die geringste Autorschaft an ihrem Text (Kosmyna et al., 2025). Die Arbeit ist ein nicht begutachteter Preprint mit 54 erwachsenen Teilnehmenden; eine methodische Erwiderung kritisiert Stichprobengröße, EEG-Auswertung und Reproduzierbarkeit (Stanković et al., 2026). Die EEG-Maße erlauben keine Aussage

über Lernergebnisse; der Befund ist als Hinweis zu lesen, nicht als Beleg.



Daraus ergibt sich die Grundsatzfrage, an der sich jeder KI-Einsatz beim Lernen messen lassen muss: Übernimmt die KI einen kognitiven Prozess, oder unterstützt sie die Lernenden dabei, ihn selbst auszuführen? Übernahme ist nicht an sich ein Fehler; sie wird es, wenn der übernommene Prozess das ist, was gelernt werden soll. Die Grenze verläuft dabei nicht erst beim Endergebnis: Auch ein Hinweis, der den entscheidenden Schritt vorwegnimmt, übernimmt den Prozess.

Kompetenz als Bezugsgröße: theoretische Grundlagen

Das SKILL-Modell richtet den Grad der Lenkung an der Kompetenz aus, KI in einer konkreten Sache lernwirksam zu nutzen, und nicht am Alter oder an der Klassenstufe.

Die erste Wurzel dieser Entscheidung ist das Scaffolding. Vygotsky (1978) beschreibt Lernen als Bewegung durch die Zone der nächsten Entwicklung: Was ein Kind heute mit Unterstützung kann, kann es morgen allein. Unabhängig davon prägten Wood et al. (1976) das Bild des Gerüsts, das nur so lange steht, wie es gebraucht wird; beide Traditionen wurden erst später aufeinander bezogen. Van de Pol et al. (2010) nennen drei Merkmale, die ein Gerüst von bloßer Hilfe unterscheiden: Es ist kontingent, reagiert also auf den aktuellen Stand; es wird zurückgenommen (Fading); und es überträgt Verantwortung an die Lernenden.

In der Unterrichtsforschung ist dies als „gradual release of responsibility“ bekannt (Pearson & Gallagher, 1983): Verantwortung wandert schrittweise von der Lehrkraft zu den Lernenden. Rowe (2026) hat dieses Prinzip mit dem GRAIT-Modell auf KI übertragen und unterscheidet die schrittweise Freigabe kognitiver Verantwortung von der schrittweisen Freigabe der Technik, gestaffelt nach nachgewiesener Bereitschaft in der Sache.

Für das schrittweise Zurücknehmen von Hilfe liefern Renkl und Atkinson (2003) die klassische empirische Grundlage: Der Übergang vom ausgearbeiteten Beispiel zum eigenen Lösen gelingt am besten, wenn Lösungsschritte nach und nach ausgeblendet werden. Für computergestütztes Scaffolding in den MINT-Fächern belegt die Metaanalyse von Belland et al. (2017) auf Basis von 144 Studien einen mittleren Effekt ($g = 0{,}46$). Eine Größenordnung für gestufte Hilfe insgesamt liefert VanLehn (2011): Tutorsysteme mit schrittweiser Hilfe erreichen gegenüber keiner Tutorierung etwa $d = 0{,}76$, menschliche Tutorierung etwa $d = 0{,}79$; die lange kolportierte Überlegenheit menschlicher Tutorierung um zwei Standardabweichungen bestätigt sich nicht. Kulik und Fletcher (2016) berichten in ihrer Metaanalyse einen Median von 0,66 Standardabweichungen und zeigen zugleich, dass lokal entwickelte Tests größere Effekte ergeben als standardisierte.

Zweitens das Zusammenspiel von Vorwissen und Lenkung. Kirschner et al.

(2006) fassen in einer viel diskutierten Bestandsaufnahme die Befunde zusammen, nach denen Anfängerinnen und Anfänger von minimal gelenktem Lernen selten profitieren; die Erwiderung von Hmelo-Silver et al. (2007) hält dem entgegen, dass problem- und forschungsbasierte Ansätze in der Regel selbst stark gerüstet sind und die Gleichsetzung mit unangeleitetem Entdecken deshalb nicht haltbar ist. Die Metaanalyse von Lazonder und Harmsen (2016) stützt diese Einschätzung: Über 72 Studien zum forschenden Lernen hinweg wirkt Lenkung mittel auf die Lernergebnisse (d = 0,50) und mittel bis stark auf Lernaktivitäten (d = 0,66) und Aufgabenerfolg (d = 0,71). Zu Altersunterschieden kommen Lazonder und Harmsen zu keinem konsistenten Befund und halten die Zahl der Studien ausdrücklich für zu gering, um sie abschließend zu beurteilen. Für das SKILL-Modell folgt daraus nicht, dass Alter irrelevant wäre, sondern dass es ein grober Stellvertreter für den Stand der Kompetenzen ist.

Zugleich gilt die Umkehrung, in der Lehr-Lern-Forschung als Expertise-Umkehr-Effekt bekannt: Eine enge Lenkung, die Anfängerinnen und Anfängern hilft, behindert Fortgeschrittene, weil die dargebotene Struktur mit dem Wissen konkurriert, das die Lernenden bereits geordnet abgespeichert haben, und Arbeitsgedächtnis bindet (Kalyuga et al., 2003; Kalyuga, 2007; Sweller et al., 2019). Eine Metaanalyse über 60 Studien bestätigt den Effekt, zeigt aber zugleich, dass das Geben von Hilfe für Anfängerinnen und Anfänger stärker wiegt als ihr Vorenthalten bei Fortgeschrittenen (Tetzlaff et al., 2025); für die Grundschule sowie für Geistes- und Sprachwissenschaften ist die Befundlage dort weniger klar, und für die starke Fassung des Modells ist das eine Mahnung zur Vorsicht. Für Tutorsysteme haben Koedinger und Aleven (2007) dieses Spannungsfeld als „assistance dilemma“ beschrieben: Wann ist es lernförderlicher, Informationen zu geben, und wann, sie zurückzuhalten, damit die Lernenden sie selbst hervorbringen? Beides kann schaden. Zu frühe Hilfe verhindert die Eigenkonstruktion, ausbleibende Hilfe lässt Lernende in unproduktiven Sackgassen stecken. Der Ort des Optimums verschiebt sich mit der Kompetenz. Diese Verschiebung ist die Grundbewegung des SKILL-

Spektrums: Mit wachsender Kompetenz sinkt die nötige Lenkung.

Drittens die Entwicklung von Metakognition und epistemischer Wachsamkeit. Ob jemand eine KI als Denkpartner nutzen kann, setzt voraus, das eigene Verstehen einschätzen und fremde Aussagen prüfen zu können. Beides entwickelt sich über Jahre. Zu unterscheiden sind das Einschätzen des eigenen Verstehens (Monitoring) und die daran anschließende Steuerung des eigenen Vorgehens (Kontrolle). Das Monitoring kann vorausschauend erfolgen, vor der Bearbeitung, oder rückblickend, nach ihr (Nelson & Narens, 1990); beide entwickeln sich unterschiedlich schnell. Innerhalb des Monitorings unterscheidet Schraw (2009) fünf Gütemaße: absolute und relative Genauigkeit, Verzerrung, Streuung und Diskrimination. Hier sind zwei Größen zentral. Absolute Genauigkeit und Verzerrung zusammengenommen, im Folgenden abkürzend Kalibrierung genannt, ein Sammelbegriff, den Schraw selbst nicht führt, beschreiben, ob die Einschätzung im Mittel zutrifft; die Diskrimination beschreibt, ob eine Person zwischen eigenen richtigen und falschen Antworten unterscheidet. In den vorliegenden Untersuchungen überschätzen sich Kinder im frühen Grundschulalter deutlich und unterscheiden gleichwohl über dem Zufall zwischen richtigen und falschen Antworten (Kolloff et al., 2023; im Überblick Roebers, 2017). Für das Modell folgt daraus, dass die Lenkung auf Stufe 1 weniger die Urteilsfähigkeit im Einzelfall ersetzen muss als die globale Entscheidung, wie viel Aufwand eine Aufgabe noch verlangt.

Destan und Roebers (2015) teilten 93 Sechsjährige nach der Genauigkeit ihrer Gesamteinschätzung in Kinder, die sich unterschätzten, realistisch einschätzten oder überschätzten: Wer sich unterschätzte, unterschied in den Konfidenzurteilen deutlich besser zwischen richtigen und falschen Antworten als Kinder, die sich überschätzten; die Gruppen unterschieden sich auch in nachgelagerten Kontrollprozessen; ein Zusammenhang der globalen Einschätzungsgenauigkeit mit den erhobenen exekutiven Funktionen (kognitive Flexibilität, Inhibition,

Arbeitsgedächtnis) zeigte sich dagegen nicht. Die Studie ist querschnittlich angelegt und zeigt Zusammenhänge, keine Wirkrichtung; die Annahme des Modells, dass Selbsteinschätzung und Impulskontrolle gemeinsam die Stufe bestimmen, findet in ihr auf der Ebene individueller Unterschiede keine Stütze.

Eine quer- und längsschnittliche Studie mit 305 Kindern der Klassen 2 und 4 (im Mittel 7,6 und 9,6 Jahre), von denen 274 ein Jahr später erneut untersucht wurden, zeigt, dass sich sowohl die rückblickende Einschätzung der eigenen Arbeit als auch die daran anschließende Kontrolle über dieses Jahr verbessern, die vorausschauenden Pendants dagegen nicht, mit Ausnahme der vorausschauenden Kontrolle bei den Zweitklässlern; die Kontrolle blieb auch dort suboptimal, wo sie sich verbesserte, nach Deutung der Autorinnen und Autoren als Folge eines weiterhin überoptimistischen Monitorings. Den Zugewinn der rückblickenden Kontrolle führen sie auf Aufgabenvertrautheit zurück, die Verbesserung des rückblickenden Monitorings auf Entwicklung, gestützt auf den Vergleich mit einer eigenen Querschnittsstichprobe von 144 Kindern und entsprechend vorläufig (Bayard et al., 2021). Kinder erkennen also eher, wie gut sie eine Aufgabe gelöst haben, als sie vorher einschätzen können, wie schwer sie wird; über das Grundschulalter hinaus lässt sich aus dieser Studie nichts ableiten.

Für die anschließende Entwicklung liegt Evidenz aus dem Jugendalter vor: Weil et al. (2013) trennten bei 56 Personen zwischen 11 und 41 Jahren die Aufgabenleistung von der metakognitiven Diskrimination und fanden bei durch ein Staircase-Verfahren konstant gehaltener Leistung einen Anstieg dieser Diskrimination innerhalb der Adoleszenz; zur Kalibrierung sagt die Studie nichts; sie beschreiben den Verlauf als höchsten Stand im späten Jugendalter mit anschließendem Plateau; diese Deutung stützen ihre Daten nur teilweise, denn die Erwachsenen lagen im Mittel unter den Jugendlichen. Der Alterseffekt beruht auf einer einzigen Korrelation in der Jugendgruppe ($r = .38$, $n = 28$), die Stichprobe ist klein und querschnittlich, enthält zwischen 17 und 19 Jahren niemanden, innerhalb der Erwachsenengruppe

stieg die Genauigkeit mit dem Alter nicht weiter an, und die Aufgabe ist eine Wahrnehmungsaufgabe ohne Fachbezug. Der Befund ist deshalb ein Hinweis auf eine über das Grundschulalter hinausreichende Entwicklung, keine belastbare Verlaufskurve.

Roebers (2017) ordnet Monitoring und exekutive Funktionen, also Planen, Aufmerksamkeitskontrolle und Impulshemmung, einem gemeinsamen Rahmen kognitiver Selbstregulation zu; sie arbeitet Gemeinsamkeiten wie verbleibende Unterschiede beider Forschungstraditionen heraus und schlägt die Integration als Rahmen vor, nicht als belegte Identität. Roebers selbst nimmt an, dass exekutive Funktionen der metakognitiven Kontrolle früh vorausgehen und für sie notwendig, aber nicht hinreichend sind; sie hält die empirische Grundlage jedoch für dünn: Nur wenige Arbeiten prüfen die Verbindung auf latenter Ebene, es liegen Nullbefunde vor, und Intelligenz wurde bislang nicht kontrolliert. Die Wirkrichtung ist damit nicht entschieden.

Wer das eigene Verstehen ungenau einschätzt, kann kaum entscheiden, wann eine KI-Hilfe angemessen ist und wann sie das Lernen ersetzt. Für Tutorsysteme ist die entsprechende Fehlregulation dokumentiert: Lernende holen Hilfe zu früh, zu spät oder zu ausgiebig (Aleven et al., 2016).

Für KI-gestützte Systeme zeigen Darvishi et al. (2024) in einer achtwöchigen Untersuchung mit 1625 Studierenden zur Qualität von Peer-Rückmeldungen in der Hochschullehre, dass verfügbare KI-Hilfe die Selbstregulation untergräbt und die Leistung einbricht, sobald die Hilfe zurückgenommen wird; eine Checkliste zur Selbstbeobachtung, die anstelle der KI-Hinweise eingesetzt wurde, fing den Einbruch nur teilweise auf und blieb hinter der KI-Unterstützung zurück; die Kombination aus beidem war der KI allein nicht überlegen. Die Autorinnen und Autoren folgern, dass die Studierenden sich auf die KI verließen, statt von ihr zu lernen. Das macht die Forderung dringlicher, das vom Werkzeug übernommene

Monitoring auf Stufe 1 im Unterricht zum Thema zu machen, zeigt aber auch, dass eine begleitende Checkliste dafür nicht genügt. Zielgröße war dort die Qualität von Rückmeldungen, nicht fachliches Lernen; für generative Tutoren im Unterricht steht die Prüfung aus.

Ähnliches gilt für das Prüfen von Informationen. Sperber et al. (2010) fassen unter epistemischer Wachsamkeit Mechanismen, die zwei Dinge prüfen: die Vertrauenswürdigkeit der Quelle, also ihre Kompetenz und ihr Wohlwollen, und die Plausibilität und Kohärenz des Inhalts. Über den ersten dieser beiden Mechanismen verfügen Kinder früh: Sie gewichten bisherige Zuverlässigkeit, Expertise und die Übereinstimmung mehrerer Auskunftspersonen; zugleich stellen Kinder bei unerwarteten Auskünften eigene Überzeugungen zurück und folgen Auskunftspersonen mitunter aus sozialen, nicht aus epistemischen Gründen (Harris et al., 2018). Zugleich ist die Prüfhaltung nicht der Regelfall: Kinder setzen ihre Prüffähigkeiten unregelmäßig und situationsabhängig ein und übernehmen Auskünfte kooperativer Erwachsener im Zweifel (Mills, 2013). Eine einflussreiche, an Säuglingsbefunden entwickelte Theorie deutet diese Bereitschaft als Anpassung an die Weitergabe verallgemeinerbaren Wissens (Csibra & Gergely, 2009); ihre Annahme, ostensive Hinweisreize wirkten spezifisch, hielt zumindest für das Blickfolgen sechs Monate alter Säuglinge einer Prüfung mit zusätzlichen, nicht-ostensiven Aufmerksamkeitsbedingungen nicht stand (Szufnarowska et al., 2014), und die Theorie ist auch konzeptuell bestritten (Heyes, 2016). Auch die inhaltliche Prüfung ist Kindern nicht fremd: Sie erkennen Unwissen, Inkompetenz und Täuschung und widersprechen auffällig unplausiblen Aussagen, doch entwickelt sich diese kritische Haltung langsam und bleibt stark davon abhängig, wie deutlich eine Unstimmigkeit hervortritt (Mills, 2013).

Diese Prüfmechanismen sind an menschlichen Quellen entstanden, und die Merkmale, an denen sie ansetzen, sind bei einem Sprachmodell nicht informativ; seine Fehler treten gerade nicht auffällig hervor, sondern sind in flüssige, gut

geformte Sprache eingebettet. Ein Sprachmodell täuscht nicht absichtlich; seine Fehler entstehen ohne Motiv. Trainings- und Produktentscheidungen bringen aber systematische Neigungen in die Ausgabe. Die bekannteste ist die Tendenz, der fragenden Person zuzustimmen, über mehrere Systeme hinweg nachgewiesen und auf das Training an menschlichen Präferenzurteilen zurückzuführen (Sharma et al., 2024). Flüssigkeit, Freundlichkeit und Selbstsicherheit sind daher von der Richtigkeit des Inhalts entkoppelt, und die Ausgabe hat keine erkennbare Herkunft, an der sich Zuverlässigkeit ablesen ließe.

Eine bereichsabhängige Selektivität wenden Kinder auch auf technische Quellen an: In zwei Studien mit 4- bis 5-Jährigen und 7- bis 8-Jährigen wählten die älteren Kinder (7 bis 8 Jahre) bei Sachfragen häufiger den Sprachassistenten als die jüngeren (4 bis 5 Jahre), während beide Altersgruppen bei persönlichen Auskünften über eine anwesende Person den Menschen bevorzugten (Girouard-Hallam & Danovitch, 2022); die Befunde zeigen nicht alle in dieselbe Richtung, denn unter 3- bis 5-Jährigen folgten die Jüngeren einem ungenauen menschlichen Informanten eher als einem ungenauen Roboter, die Älteren behandelten beide gleich (Li & Yow, 2024; verkörperter Roboter, Benennung vertrauter Objekte, N = 120); die Stichprobe stammt aus überwiegend weißen US-Familien der oberen Mittelschicht, und geprüft wurde ein eigens konstruierter Sprachassistent mit vorgegebenen Antworten, kein generatives Modell. Gerade in dem Bereich, in dem ein Sprachmodell plausibel klingende Fehler produziert, wächst das Vertrauen in die maschinelle Quelle mit dem Alter. Die Last liegt damit auf der inhaltlichen Prüfung, die Fachwissen voraussetzt. Erfahrung mit dem Werkzeug bringt diese Prüfkompetenz nicht von selbst mit: Nutzungsroutine und Verständnis der Funktionsweise fallen bei Kindern auseinander (Danovitch, 2019).

Diese Entwicklungsbefunde machen plausibel, warum jüngere Lernende in der Regel mehr Lenkung brauchen. Sie belegen es nicht: Anders als beim Vorwissen, für das die Expertise-Umkehr eine geprüfte Wechselwirkung von Personenmerkmal und

Instruktionsform liefert, prüfen erst wenige Studien individuelle Unterschiede im Monitoring oder in der Selbstregulation als Moderator des Nutzens von Lenkung, und sie fallen überwiegend negativ aus: Herold und Seufert (2025) fanden bei 102 Studierenden eine Moderation durch das Vorwissen, und diese nur für das Verstehen, nicht für Behalten und Transfer, dagegen keine Moderation durch die metakognitive Kompetenz; Hofer und Reinhold (2025) fanden bei 332 Sechstklässlerinnen und Sechstklässlern Wechselwirkungen von Personenmerkmal und Scaffolding nur in einer der beiden untersuchten Schulformen. Beide Prüfungen stammen nicht aus dem Grundschulalter, für das das Modell die stärkste Lenkung vorsieht.

Näher an der Sache liegen erste Arbeiten zur Metakognition im Umgang mit generativer KI: Stadler et al. (2024) finden bei Studierenden gegenüber der Websuche geringere kognitive Belastung bei zugleich flacherer Argumentation; umgekehrt verbesserte in einem randomisierten Experiment mit Studierenden eine modellgestützte Vorhersage der eigenen Testleistung mit Reflexionsimpulsen die Kalibrierung der zuvor überkonfidenten Teilnehmenden, ein kleiner Subgruppeneffekt (Lee et al., 2025). Ein Werkzeug kann Monitoring also auch trainieren statt ersetzen; das ist die prüfbare Alternative zur Annahme, die Übernahme des Monitorings auf Stufe 1 sei nur eine Entlastung auf Zeit. Die Übertragung auf Monitoring und Selbstregulation ist damit nicht nur ungeprüft, sondern von den vorliegenden Prüfungen bislang unbestätigt; sie steht und fällt mit der im Abschnitt zu den Grenzen skizzierten Untersuchung.

Wirksam ist dabei der Stand der Kompetenzen, nicht das Alter. Der Schluss zurück, wer viel weiß, sei auch bereit für eine höhere Stufe, gilt allerdings nur eingeschränkt. Fachwissen ist gegenstandsgebunden und kann in einem Gebiet rasch wachsen; darauf stützt sich ein Teil der inhaltlichen Prüfung unmittelbar.

Die Genauigkeit der Selbsteinschätzung verbessert sich dagegen nach den

vorliegenden, noch schmalen Hinweisen über einen langen, bis in die Adoleszenz reichenden Zeitraum (Weil et al., 2013). Bei den exekutiven Funktionen verlaufen die Komponenten unterschiedlich. Die deutlichsten Zuwächse der Inhibition liegen im Vorschulalter, doch verbessern sich auch einfache Hemmaufgaben bis in die Adoleszenz weiter, während Arbeitsgedächtnis und kognitive Flexibilität über Kindheit und Jugend eher stetig zunehmen. Die Befundlage gilt den Autoren selbst als uneinheitlich (Best & Miller, 2010). Für das Modell einschlägig ist eine Variante, die die üblichen neutralen Verfahren nur unvollständig abbilden: eine verfügbare Belohnung zugunsten eines späteren Ziels zurückzustellen. Zelazo und Carlson (2012) ordnen sie den „heißen“ exekutiven Funktionen zu und halten für diese einen eigenen, möglicherweise verzögerten Verlauf für wahrscheinlich, verweisen aber zugleich auf Arbeiten, die keine getrennten Faktoren für heiße und kühle exekutive Funktionen finden. Der eigenständige Verlauf ist damit eine begründete Annahme, kein gesicherter Befund.

Ob diese Verläufe auf Reifung, auf kumulierte Erfahrung oder auf beides zurückgehen, ist mit überwiegend querschnittlichen Daten nicht zu entscheiden (siehe oben zu Bayard et al., 2021). Für das Modell genügt der deskriptive Befund: Durch Fachwissen in einem einzelnen Gebiet lassen sich diese Anteile nicht aufwiegen. Ein Kind, das in einer Sache viel weiß, kann darum plausible Fehler in dieser Sache durchaus finden und dennoch nicht die Selbststeuerung mitbringen, die Stufe 3 verlangt.

Das Alter setzt den Selbstregulationsanteilen deshalb eine Erwartung, keine Grenze: Es macht bestimmte Ausprägungen unwahrscheinlich, sagt über die einzelne Person aber wenig und taugt nicht als Zuordnungsregel. Die genannten Befunde beschreiben Gruppenmittel mit erheblicher Streuung zwischen Personen und Aufgaben; sie begründen die Reihenfolge der Stufen, ersetzen aber keine individuelle Einschätzung.

Abgrenzung zu verwandten Modellen

Verwandte Modelle liegen vor. Rowe (2026) staffelt im GRAIT-Modell den KI-Zugang nach nachgewiesener Bereitschaft in der Sache (Wissen, Fertigkeit, epistemisches Verständnis, KI-Kompetenz) und unterscheidet drei Phasen: unabhängige Kognition ohne KI, geteilte Kognition und schließlich ausgelagerte und umverteilte Kognition. Die Leitfrage ist dieselbe wie hier.

Praktisch am weitesten verbreitet ist die AI Assessment Scale (Perkins et al., 2024), die den zulässigen KI-Einsatz in fünf Stufen von „kein KI-Einsatz“ bis „vollständiger KI-Einsatz“ staffelt und vor allem der Transparenz und Prüfungsintegrität dient; in der überarbeiteten Fassung (Perkins et al., 2025) sind die Stufen neu benannt und ausdrücklich nicht mehr hierarchisch gemeint. Bezugsgröße bleibt in beiden Fassungen die Aufgabe oder das Prüfungsformat, nicht die Person; dieselbe Einstufung gilt für alle Lernenden einer Kohorte.

Molenaar (2022b) beschreibt unter dem Begriff der hybriden Intelligenz, wie sich Erkennen, Diagnostizieren und Handeln zwischen Lehrkraft und Technik verteilen, und staffelt diese Verteilung in sechs Grade der Automatisierung; in einer zweiten Arbeit (Molenaar, 2022a) macht sie daraus eine normative Forderung: Adaptive Lernumgebungen übernehmen die Regulation von 10- bis 14-Jährigen, was die Entwicklung ihrer Selbstregulation gefährdet, weshalb die Kontrolle schrittweise zurückzugeben ist. Diesen Gedanken teilt das SKILL-Modell; Molenaars Arbeiten sind ihm am nächsten. Drei Unterschiede bleiben: Molenaar bestimmt die Rückgabe der Kontrolle über den Automatisierungsgrad des Systems, SKILL über eine in vier Teilkompetenzen aufgeschlüsselte, gegenstandsgebundene Kompetenz der Person; Molenaar betrachtet die adaptive Lernumgebung, SKILL das Zusammenspiel von Werkzeug und Lehrkraft als gemeinsam zu verteilende Größe; und Molenaars Stufung ist beschreibend, während SKILL daraus eine Auswahlregel für Werkzeuge ableitet. Die Verteilung von Unterstützung auf mehrere Instanzen ist zudem als

verteiltes Scaffolding beschrieben (Puntambekar & Hübscher, 2005), und Tabak (2004) hat mit dem Begriff der Synergie gefasst, dass Software und Lehrkraft dieselbe Lernbedürftigkeit adressieren und sich wechselseitig verstärken können.

Das ICAP-Modell (Chi & Wylie, 2014) ordnet Lernaktivitäten nach dem Grad der Beteiligung von passiv über aktiv und konstruktiv bis interaktiv und sagt in dieser Reihenfolge steigende Lernergebnisse voraus; eine KI, die fertige Lösungen liefert, verschiebt Lernende nach dieser Ordnung von konstruktiv nach passiv, wobei diese Übertragung auf generative KI eine Ableitung ist und nicht geprüft. TPACK (Mishra & Koehler, 2006) und das SAMR-Modell (in der Beschreibung von Hamilton et al., 2016) beschreiben das Wissen der Lehrkraft beziehungsweise den Grad der technologischen Veränderung einer Aufgabe; beide bestimmen den Lenkungsgrad eines Werkzeugs nicht. SAMR liegt keine begutachtete Primärpublikation zugrunde, und es wird wegen fehlender Kontextsensitivität, seiner Hierarchieannahme und seiner Produktorientierung kritisiert (Hamilton et al., 2016).

Das SKILL-Modell setzt an drei Stellen anders an. Die Bezugsgröße ist die Kompetenz, KI lernwirksam zu nutzen, aufgeschlüsselt in vier Teilkompetenzen und teilweise gegenstandsgebunden: Dieselbe Person kann in einem vertrauten Fach die Obergrenze ausschöpfen, die ihre bereichsübergreifenden Anteile setzen, und in einem neuen Gebiet auf Stufe 1 zurückfallen, während der umgekehrte Sprung nicht vorgesehen ist. Gegenüber aufgabenbezogenen Rahmen wie der AI Assessment Scale bleibt entscheidend, dass die Zuordnung personenbezogen erfolgt und innerhalb einer Kohorte variiert. Lenkung wird zudem als Zusammenspiel von Werkzeug und Lehrkraft gefasst und an eine Stufenentscheidung gekoppelt.

Die Unterscheidung der beiden Quellen ist nicht neu: Saye und Brush (2002) trennen fest im Werkzeug verankerte Gerüste (hard scaffolds), die vorab geplant sind und für alle gleich wirken, von situativ erzeugten Gerüsten der Lehrkraft (soft

scaffolds), die diagnostisch reicher, aber knapp sind; das verteilte Scaffolding und Tabaks Synergiebegriff beschreiben ihr Zusammenwirken. Neu ist die normative Wendung: Fehlende Lenkung im Werkzeug muss in der Situation von der Lehrkraft hergestellt werden, und weil personelle Lenkung nur punktuell verfügbar ist, begrenzt sie, welche Werkzeuge in welchem Setting vertretbar sind. Aus einer Beschreibung wird eine Auswahlregel. Und Lenkung wird in drei Hebel zerlegt, sodass sich Werkzeuge beschreiben lassen, die etwa vollständige Kontextualisierung mit geringer Hilfedichte verbinden; eine Stufenlogik ohne Hebel kann diese Kombination nicht abbilden. Die Trennung ist analytisch und in einer Richtung asymmetrisch: Kontextualisierung ohne Leitplanken ist ohne Weiteres herstellbar, belastbare Leitplanken setzen dagegen eine Einbettung auf Anwendungsebene voraus (Abschnitt „Begriffsklärung“). Ob sich die drei Hebel empirisch unabhängig variieren lassen, ist offen und im Forschungsteil als eigene Frage geführt.

GRAIT staffelt dagegen entlang erreichter Meilensteine im Curriculum und bezieht ein, ob Lernende die Ausgabe schon bewerten können, reguliert aber den Zugang zur Technik als Ganzes. Molenaars Automatisierungsgrade lassen sich innerhalb der SKILL-Stufen als Beschreibung der Regulationsverteilung nutzen. Was das Modell nicht leistet: Es enthält kein Diagnoseinstrument für die Stufenzuordnung und ist empirisch nicht geprüft.

Teilkompetenzen

Die Kompetenz, KI lernwirksam zu nutzen, ist situations- und gegenstandsgebunden. Vier Teilkompetenzen bestimmen sie; sie lassen sich unabhängig vom Lernergebnis erfassen und halten die Bezugsgröße des Modells damit prüfbar:

- **Fachwissen in der Sache.** Wer wenig über das Thema weiß, kann die fachliche

Qualität einer Antwort nicht beurteilen und übersieht plausible Fehler (Kalyuga, 2007; Kirschner et al., 2006). Wer die schriftliche Division noch nicht sicher beherrscht, sieht einer sauber gesetzten, aber falschen Rechnung die Fehlstelle nicht an. Erfassbar über Vortests, unabhängig vom KI-Einsatz.

- **Kritische KI-Kompetenz.** Grenzen des Werkzeugs, erfundene Angaben und gefällige Antworten erkennen und Ausgaben systematisch prüfen. Die Kompetenzrahmen der UNESCO (2024) sowie von OECD und Europäischer Kommission (2026) führen diese Fähigkeit als wesentlichen Bestandteil von KI-Bildung; Ng et al. (2021) unterscheiden Kennen, Anwenden, Bewerten und ethisches Reflektieren. Erfassbar vor dem Einsatz an standardisiertem Material, und zwar in zwei getrennten Anteilen: fachliche Fehlererkennung in einem Gebiet, dessen Vorwissen separat erhoben und statistisch kontrolliert wird, sowie quellen- und formbezogene Prüfindikatoren, die kein Fachwissen verlangen, etwa das Erkennen zustimmender Anpassung an die Fragestellung (Sharma et al., 2024), erfundener Belegangaben und fehlender Herkunft der Ausgabe. Ohne diese Trennung misst der Indikator im Wesentlichen Fachwissen; die Abgrenzung gegenüber dem Vorwissenstest ist vor der konfirmatorischen Analyse zu berichten.
- **Selbstregulation.** Wer selbstreguliert lernt, bereitet sich vor, beobachtet sich beim Arbeiten und schaut anschließend zurück (Zimmerman, 2002), gestützt auf Zielsetzung und Selbstwirksamkeit. Für den KI-Einsatz heißt das: entscheiden können, wann die KI als Denkpartner sinnvoll ist und wann nicht. Dazu gehört die Impulskontrolle, die Lösung nicht sofort abzurufen. Sie greift auf inhibitorische Kontrolle zurück, die zu den exekutiven Funktionen zählt; exekutive Funktionen sind früh angelegt, verstärken sich aber über Kindheit und Jugend hinweg weiter, wobei ihre Komponenten unterschiedlich verlaufen (Best & Miller, 2010). Anspruchsvoll ist hier weniger das Unterdrücken eines Impulses in einer neutralen Aufgabe, wie es die klassischen Verfahren erfassen, als das Zurückstellen einer verfügbaren und unmittelbar erfolgreichen Handlung

zugunsten eines Lernziels, das erst später sichtbar wird; diese Anforderung enthält eine motivationale Komponente, die die üblichen Maße exekutiver Funktionen nicht abbilden. Dazu gehört die adaptive Hilfesuche, von Newman (2002) als Strategie der Selbstregulation gefasst, die auf künftige Selbstständigkeit zielt: einen Hinweis statt der Lösung anfordern, und nur so viel wie nötig; dass Lernende Hilfe erst nach einem eigenen Versuch abrufen sollten und dies regelmäßig verfehlen, ist für Tutorsysteme dokumentiert (Aleven et al., 2016). Fan et al. (2025) zeigen, dass sich gerade diese Prozesse unter freier KI-Nutzung verschieben. Erfassbar über verhaltensnahe Prozessdaten aus KI-fernen Aufgaben, etwa Zeitverteilung, Wiederholungswahl und Hilfeanforderung nach einem eigenen Versuch, nicht am Verlauf des laufenden Einsatzes; Selbstauskünfte kommen ergänzend hinzu, genügen allein aber nicht, weil sie mit verhaltensnahen Maßen nur schwach übereinstimmen (Rovers et al., 2019).

- **Sprachliche und kognitive Voraussetzungen.** Eine Frage präzise formulieren, eine längere Antwort lesen und im Arbeitsgedächtnis halten, Kontext bereitstellen. Eine lange KI-Antwort belastet das Arbeitsgedächtnis zusätzlich zur Aufgabe; bei geringem Vorwissen fehlen die Schemata, die diese Last auffangen (Sweller et al., 2019). Ein Kind der dritten Klasse tippt „warum ist das falsch“, bekommt acht Sätze zurück und liest ab dem dritten nicht mehr mit. Die Antwort war fachlich richtig, hat dem Kind aber nicht geholfen. Sprachliche Fertigkeiten und Arbeitsgedächtnis werden hier gebündelt, weil sie im KI-Dialog denselben Engpass bilden, die Verarbeitung der Antwort; für die Modellprüfung sind sie getrennt zu erheben. Erfassbar über Lese- und Formulierungsproben sowie eine Arbeitsgedächtnisspanne.

Wer in einem oder mehreren dieser Bereiche am Anfang steht, braucht mehr Lenkung. Diese Lernenden stellen die nötige Struktur nicht selbst her: Sie geben der KI keine Rolle vor und liefern ihr keinen Fachkontext, weil sie noch nicht wissen, dass beides den Unterschied macht. Mit wachsender Kompetenz kehrt sich die

Anforderung um. Dann stört redundante Struktur, und das Werkzeug muss sich öffnen lassen. Für die Stufenzuordnung in der Praxis legt dieser Beitrag vorläufig die niedrigste Teilkompetenz zugrunde (Minimum-Regel): Eine einzelne fehlende Voraussetzung, etwa geringe Impulskontrolle bei hohem Fachwissen, wird dann nicht durch die übrigen aufgewogen. Die Regel ist eine konservative Entscheidungsheuristik, kein Bestandteil der Grundannahme. Ob die vier Teilkompetenzen tatsächlich jede für sich notwendig sind oder einander teilweise ersetzen können, ist offen und im Forschungsteil als eigene Frage benannt. Konservativ ist die Regel, weil sie im Zweifel mehr Lenkung vorsieht. Ihr Preis ist, dass eine kleine Veränderung in einer einzelnen Teilkompetenz die Zuordnung kippen kann, obwohl Kompetenz und Lenkung stetige Größen sind; in der Praxis wird das abgefedert, indem die Hebel innerhalb der Stufe unterschiedlich eingestellt werden. Wie sich die vier Teilkompetenzen ökonomisch erfassen lassen, ist offen (Abschnitt „Grenzen und offene Fragen“).

Vorfrage: ob KI in dieser Lernsituation gebraucht wird

Das Modell hat zwei Schritte. Der erste ist die Vorfrage, ob KI in dieser Lernsituation überhaupt einen Beitrag leistet; erst wenn sie bejaht wird, folgt als zweiter Schritt die Wahl der Stufe. Die Vorfrage ist damit Bestandteil des Modells, nicht sein Vorspann, und „kein KI-Einsatz“ ist eine seiner regulären Antworten. Ohne diese Stellung könnte das Modell so gelesen werden, als sei der Einsatz gesetzt und nur noch die Dosierung offen. Viele Lernziele werden aber ohne KI besser erreicht: weil die Tätigkeit, die die KI übernehmen würde, das ist, was geübt werden soll, oder weil Material, Gespräch und eigenes Probieren ausreichen. Ein Kind, das Zahlen zerlegt, braucht Plättchen und eine Frage, keinen Chatbot. Die Vorfrage ist der Stufenlogik deshalb vorgelagert: Nur wenn sie bejaht wird, stellt sich die Frage nach der passenden Stufe. Wird sie bejaht und sind die Voraussetzungen gering, führt das zu Stufe 1; häufig fällt die Vorfrage aber schon negativ aus. International

kommt die OECD (2026) in ihrer Synthese der vorliegenden Evidenz zu demselben Schluss: Generative KI unterstützt Lernen, wenn klare didaktische Prinzipien sie führen; wird die Aufgabe ohne pädagogische Rahmung an sie ausgelagert, steigt die Leistung, ohne dass Lernen entsteht.

Das gilt altersunabhängig, hat aber in der Grundschule besonderes Gewicht. Dort entstehen die Grundlagen, an denen sich später jede KI-Antwort messen lassen muss: Lesen, Schreiben, Zahlvorstellungen, erste Erfahrungen mit dem eigenen Denken und seinen Fehlern. Die Entwicklungsbefunde zu Monitoring und epistemischer Wachsamkeit stützen die Zurückhaltung, die die SWK (2024) für diese Phase empfiehlt. Auf Stufe 1 bleibt der Einsatz auf wenige, klar umrissene Situationen begrenzt: solche, in denen ein eingebettetes Werkzeug etwas leistet, das anders nicht zu haben ist, etwa geduldige Rückmeldung zum eigenen Lösungsweg, sprachliche Entlastung oder Zugänge für Kinder mit Beeinträchtigungen. Für die übrigen Lernsituationen genügt der reguläre Unterricht.

Vor der didaktischen Entscheidung steht zudem eine rechtliche. Frei zugängliche Chatbots setzen in ihren Nutzungsbedingungen ein Mindestalter voraus, in der Regel 13 Jahre oder das im Land geltende Mindestalter, für Minderjährige teils nur mit Erlaubnis der Sorgeberechtigten oder über elterlich verwaltete Konten; für einwilligungsbasierte Dienste gilt daneben Art. 8 DSGVO, in Deutschland mit einer Altersgrenze von 16 Jahren. In der Grundschule scheidet die Nutzung mit eigenem Konto damit aus. Für den Einsatz in der Schule gelten die Vorgaben der DSGVO und der Länder zur Auftragsverarbeitung.

Die KI-Verordnung der EU (Europäische Union, 2024) betrifft Schulen in vier Punkten. Erstens ist der Einsatz von KI-Systemen zur Ableitung von Emotionen in Bildungseinrichtungen seit dem 2. Februar 2025 verboten (Art. 5 Abs. 1 Buchst. f); erfasst sind Systeme, die Emotionen oder Absichten aus biometrischen Daten ableiten (Art. 3 Nr. 39), ausgenommen solche aus medizinischen Gründen oder

Sicherheitsgründen. Zweitens müssen Anbieter und Betreiber Maßnahmen ergreifen, um die KI-Kompetenz ihres Personals und der in ihrem Auftrag mit Betrieb und Nutzung befassten Personen zu fördern (Art. 4); die Pflicht adressiert die Schule als Betreiberin, nicht die Lernenden.

Drittens gelten Systeme, die über Zugang, Zulassung oder Zuweisung entscheiden, Lernergebnisse bewerten, das angemessene Bildungsniveau einer Person beurteilen oder Prüfungen überwachen, als Hochrisikosysteme (Anhang III Nr. 3 Buchst. a bis d; Art. 6 Abs. 3 nimmt Systeme ohne erhebliches Risiko aus, Systeme mit Profiling natürlicher Personen gelten aber stets als hochriskant). Die Pflichten für diese Hochrisikosysteme greifen nach der Änderungsverordnung (EU) 2026/1744 vom 8. Juli 2026 erst ab dem 2. Dezember 2027, für Anhang-I-Systeme ab dem 2. August 2028; dieselbe Verordnung hat die Pflicht aus Art. 4 von einem sicherzustellenden Kenntnisstand auf eine Maßnahmenpflicht zurückgenommen (Europäische Union, 2026). Viertens gelten seit dem 2. August 2026 die Transparenzpflichten des Art. 50: Anbieter müssen KI-Systeme, die unmittelbar mit Personen interagieren, so gestalten, dass diese über die Interaktion informiert werden. Die Kennzeichnungspflicht der Betreiber nach Art. 50 Abs. 4 ist enger, als oft angenommen: Sie greift bei Bild-, Ton- und Videoinhalten nur für Deepfakes und bei Texten nur, wenn diese veröffentlicht werden, um die Öffentlichkeit über Angelegenheiten von öffentlichem Interesse zu informieren, und auch dort nicht, wenn der Inhalt redaktionell verantwortet ist. Für KI-gestützt erstelltes Unterrichtsmaterial folgt daraus keine gesetzliche Kennzeichnungspflicht; für den Unterricht heißt Art. 50 vor allem, gegenüber den Lernenden offenzulegen, wenn sie mit einem KI-System arbeiten.

Für das Modell heißt das: Diese Schranken knüpfen am Alter an, nicht an der Kompetenz, und sind der Stufenlogik äußerlich. Wo sie greifen, in der Grundschule und weitgehend in der Sekundarstufe I, ist die eingebettete Lösung häufig die einzige zulässige, weil die Lehrkraft rechtliche Anforderungen nicht durch

Begleitung ersetzen kann. Für Erwachsene, die in einem neuen Fachgebiet auf Stufe 1 stehen, ist die Werkzeugwahl allein eine didaktische.

Begriffsklärung: didaktische Lenkung

Didaktische Lenkung bezeichnet in diesem Beitrag alles, was den Handlungsspielraum der Lernenden im Sinne des Lernziels formt. Der Begriff ist weiter gefasst als der Begriff *guidance* der Instruktionsforschung (Kirschner et al., 2006; Lazonder & Harmsen, 2016), der gegebene Unterstützung meint; er schließt hier zusätzlich das gezielte Zurückhalten von Information ein. Neu ist dieser zweite Anteil nicht: Reiser (2004) unterscheidet innerhalb des Scaffolding das Strukturieren, das Komplexität senkt, vom Problematisieren, das Lernende gezielt auf Schwierigkeiten stößt, statt sie zu umgehen; beide Mechanismen setzen etwas hinzu. Leitplanken arbeiten anders: Sie entziehen eine verfügbare Handlungsmöglichkeit. Sie sind damit die werkzeugseitige Antwort auf das Assistance Dilemma (Koedinger & Aleven, 2007) und auf die Frage, wie sich die Phase des eigenen Ringens schützen lässt, die Kapur (2016) vom Lernergebnis her begründet. Neu ist, dass dieses Zurückhalten als eigener, unabhängig einstellbarer Hebel geführt wird. Die Effektstärken der genannten Metaanalysen beziehen sich gleichwohl auf gegebene Unterstützung und lassen sich auf das Zurückhalten nicht übertragen. Daraus folgt eine Asymmetrie in der Belegbasis: Für Scaffolding liegen metaanalytische Effektschätzungen vor (Belland et al., 2017; Lazonder & Harmsen, 2016). Für Kontextualisierung als eigenständigen Hebel fehlt eine solche Schätzung; sie wird aus der Literatur zu Vorwissen und Lenkung abgeleitet. Für Leitplanken als gezieltes Zurückhalten stützt sich das Modell auf zwei Einzelstudien in je eigenen Kontexten (Bastani et al., 2025; Buçinca et al., 2021). Zwei der drei Hebel sind damit nur schwach belegt, und der Hebel, auf dem Stufe 1 beruht, am schwächsten.

Eine dritte Studie prüft den Kontrast unmittelbar: Bassner et al. (2026) verglichen in einem randomisierten Experiment mit 275 Studierenden eines Programmier-

Grundkurses einen Tutor mit gestuften Hinweisen und zurückgehaltener Lösung, unbeschränktes ChatGPT und eine Bedingung ohne KI. Beide KI-Bedingungen führten zu besseren Übungsergebnissen und weniger Frustration, während sich der KI-freie Wissenstest zwischen allen drei Bedingungen nicht unterschied. Der Befund wiederholt die Trennung von Leistung und Lernen, stützt aber nicht die Erwartung, dass Leitplanken im KI-freien Maß einen Vorteil erzeugen. Er wurde unmittelbar nach einer 90-minütigen Sitzung an einer hochschulischen Stichprobe erhoben, also nicht in dem Kompetenzbereich, für den das Modell den Vorteil vorhersagt; die Beweislast verschiebt er dennoch zulasten der Grundannahme.

In dieselbe Richtung weist ein randomisiertes Experiment mit 334 Studierenden, die einen Lehrbuchtext 25 Minuten lang bearbeiteten und anschließend ohne KI und ohne Text geprüft wurden: Gegenüber der Gruppe ohne Zugang schnitt die Gruppe mit unbeschränktem Zugang zu einem KI-Tutor um 0,34 Standardabweichungen besser ab, während sich der beschränkte Zugang, KI erst nach eigenständigem Lesen, nicht signifikant von der Gruppe ohne Zugang unterschied; der direkte Vergleich beider KI-Bedingungen fiel mit 0,21 Standardabweichungen zugunsten des unbeschränkten Zugangs aus und verfehlte die Signifikanz knapp (Fischer et al., 2025; Preprint). Der Befund spricht nicht gegen Leitplanken, lässt den erwarteten Vorteil aber offen; auch hier fehlt ein verzögerter Nachtest, und die Stichprobe ist hochschulisch. Der Begriff der didaktischen Lenkung ist dem Sammelwort „Vorstrukturierung“ vorzuziehen, weil er drei Dinge unterscheidbar macht, die bei KI-Werkzeugen analytisch getrennt sind:

Hebel	Was er tut	Beispiel
Scaffolding adaptive Hilfe	Gestufte Hinweise, Rückfragen und Rückmeldungen, die sich am aktuellen Stand orientieren und zurückgenommen werden, sobald sie nicht mehr nötig sind (Fading).	Der Tutor fragt erst „Was ist gesucht?“, dann „Welche Angaben brauchst du?“, bevor er einen Rechenweg andeutet.

Hebel	Was er tut	Beispiel
Leitplanken Guardrails	Feste Grenzen dessen, was die KI tut oder nicht tut. Sie schützen den Lösungsprozess, wenn er selbst Lernziel ist.	Die KI gibt keine Endergebnisse aus, bevor ein eigener Versuch vorliegt; sie bleibt beim Thema.
Kontextualisierung Einbettung von Aufgabe und Rolle	Lernziel, Aufgabe, fachlicher Hintergrund und die Rolle der KI sind im Werkzeug hinterlegt und müssen nicht von der Person hergestellt werden.	Ein Sachrechen-Tutor kennt die Aufgabe, die Klassenstufe und typische Fehlvorstellungen.

Bei Leitplanken sind zwei Dimensionen zu trennen: was sie didaktisch verlangen, etwa keine Lösung vor einem eigenen Versuch, und wie verlässlich sie technisch durchgesetzt werden. Die erste Dimension ist der Hebel der Lenkung; die zweite entscheidet, ob ein Werkzeug Stufe 1 ermöglicht. In dieser zweiten Dimension sind Leitplanken unterschiedlich belastbar. Eine im Systemprompt hinterlegte Rolle lässt sich im Gespräch aushebeln: Sprachmodelle sind auf Hilfsbereitschaft trainiert und geben Lösungen preis, wenn Lernende drängen, Autorität behaupten oder das Thema wechseln. Zhao et al. (2026) zeigen das systematisch: Ein eigens auf das Umgehen pädagogischer Vorgaben feinabgestimmter simulierter Lernender entlockt gängigen KI-Tutoren in mehrschrittigen Dialogen die Lösung, mit erheblichen Unterschieden zwischen den Modellen; rein kontextbasierte Versuche ohne solche Feinabstimmung scheitern häufiger, und einfache Schutzmaßnahmen senken die Quote, beseitigen sie aber nicht. Die Studie beziffert die obere Grenze der Angreifbarkeit; wie leicht Lernende im Unterricht diese Grenzen verschieben, ist nicht untersucht.

Verlässlich wirken Grenzen, die außerhalb des Dialogs liegen, etwa eine Freigabe des Endergebnisses erst nach einem eigenen Versuch, ein begrenzter Antwortraum oder eine Prüfung der Ausgabe vor der Anzeige. Buçinca et al. (2021) zeigen in einem Entscheidungsexperiment mit 199 Erwachsenen, dass ein erzwungenes

eigenes Urteil vor der KI-Ausgabe die Übernahme falscher Vorschläge deutlicher verringert als bloße Erklärungen. Für die Übertragung ist der Befund dreifach eingeschränkt. Erstens beruht er auf einer einzelnen Online-Stichprobe an einer Alltagsaufgabe ohne Lernziel. Zweitens bewerteten die Teilnehmenden gerade die Gestaltungen am schlechtesten, die die Übernahme falscher Vorschläge am stärksten senkten. Drittens nützten diese Gestaltungen vor allem Personen mit hoher Bereitschaft zur kognitiven Anstrengung. Ausgerechnet die Gruppe, der Stufe 1 solche Zwänge zumutet, könnte also am wenigsten davon profitieren; ob sich der Befund auf Lernaufgaben und Kinder überträgt, ist ungeprüft.

Anbieterseitige Lernmodi wie der Study Mode in ChatGPT oder Guided Learning in Gemini sollen das Verhalten in die didaktisch gewünschte Richtung verschieben, garantieren die Grenzen aber nicht; für die Gemini-Variante liegt eine Darstellung des Anbieters vor (LearnLM Team, 2024). Zu Guided Learning berichtet Google DeepMind (2026) eine präregistrierte randomisierte Feldstudie über acht Wochen mit 1763 Lernenden der Sekundarstufe I an zwölf Schulen in Sierra Leone und einen Effekt von 0,26 Standardabweichungen auf die Mathematikleistung; nach Angaben des Berichts erreichten 69 Prozent der Lernenden die Nutzungsvorgaben, eine getrennte Angabe von Zuweisungs- und Nutzungseffekt fehlt; die Studie stammt vom Anbieter, ist nicht begutachtet, und der Bericht nennt nicht, was die Vergleichsklassen erhielten; verglichen wurde eine lehrkraftgeführte Lektionsreihe mit regulärem Unterricht, sodass die Studie den hier interessierenden Kontrast zwischen gelenkter und offener Nutzung nicht prüft. Eine anbieterunabhängige Prüfung der Lernwirkung dieser Modi steht aus. Maurya et al. (2025) haben im Benchmark MRBench die Antworten aktueller Sprachmodelle mit denen erfahrener und unerfahrener menschlicher Tutorinnen und Tutoren verglichen und 1596 Einzelantworten aus 192 Dialogen entlang acht pädagogischer Dimensionen bewertet: Die Systeme erkennen Fehler, verorten sie und geben Hinweise, ohne die Lösung preiszugeben, aber sehr unterschiedlich verlässlich, und keines erreicht über alle acht Dimensionen hinweg das Niveau erfahrener Tutorinnen und Tutoren,

wobei die Arbeit auch bei den menschlichen Vergleichsgruppen Lücken feststellt. Ein Lernmodus lässt sich zudem abschalten oder durch einen zweiten Zugang umgehen.

Auf Stufe 1 müssen die Leitplanken deshalb belastbar sein, also außerhalb des Dialogs verankert: entweder auf Anwendungsebene oder dadurch, dass die Lehrkraft den Dialog selbst führt und die Ausgabe prüft, bevor sie die Lernenden erreicht. Eine per Prompt oder Modus gesetzte Rolle, die die Lernenden selbst bedienen, bleibt im Dialog aushebelbar und reicht nur für Stufe 2. Belastbare Leitplanken setzen damit eine Einbettung auf Anwendungsebene oder eine durchgängige personelle Vermittlung voraus; die drei Hebel sind analytisch getrennt, in der Praxis aber nur teilweise unabhängig.

Der Lenkungsgrad eines Werkzeugs ergibt sich aus dem Zusammenspiel der drei Hebel. Die Hebel skalieren Dichte und Reichweite der Hilfen; ihre Kontingenz ist keine Stufengröße, sondern auf allen Stufen Qualitätsmerkmal. Damit diese Festlegung nicht gegen Widerlegung abschirmt, ist Kontingenz vor einer Erhebung an Merkmalen des Werkzeugs zu bestimmen, unabhängig vom Lernergebnis. Kontingenz betrifft dabei allein Umfang und Inhalt der Hilfe: Sie richtet sich nach einem im Dialog beobachtbaren Indikator des Bearbeitungsstands, etwa der Art des Fehlers, der Zahl der Versuche oder der angeforderten Hinweisstufe, nicht nach einem festen Ablauf, und eine Regel für ihre Rücknahme ist hinterlegt und im Protokoll nachweisbar ausgelöst. Ob überhaupt erst nach einem eigenen Versuch geholfen wird, gehört dagegen zum Hebel der Leitplanken und bleibt aus der Kontingenzdefinition ausgeklammert, weil beide Hebel im Forschungsteil gegeneinander gekreuzt werden. Ob eine Bedingung diese Merkmale erfüllt hat, prüft ein Manipulationscheck an den Dialogprotokollen, nicht der Lernerfolg; eine Hilfe, die Fortgeschrittene gebremst hat, darf nicht nachträglich zur nicht-kontingenten erklärt werden. Besteht eine Bedingung den Manipulationscheck und zeigt sich für Fortgeschrittene gleichwohl ein Nachteil, ist das ein Befund gegen die

Auffassung, kontingente Hilfe könne nicht bremsen; er ist als solcher zu berichten. Daraus folgt eine Einschränkung der Achse: Weil Scaffolding hier als kontingente Hilfe definiert ist, kann seine Erhöhung den für Fortgeschrittene erwarteten Nachteil definitionsgemäß nicht erzeugen; dieser geht auf Leitplanken und Kontextualisierung zurück sowie auf Hilfen, die den Modellsinn verfehlen, etwa eine feste Hinweisreihenfolge. Für Werkzeuge sind beschreibende Bezeichnungen klarer als „vorstrukturiert“: *eingebettet* oder *app-integriert* für KI, die in einer Lernanwendung mit hinterlegter Aufgabe und fester Rolle arbeitet; *teiloffen* für ein offenes Werkzeug, in dem Rolle und Auftrag vorbereitet, aber im Dialog veränderbar sind; *offen* für ein Werkzeug ohne hinterlegte Vorgaben. Diese Bezeichnungen betreffen nur das Werkzeug; wie viel personelle Begleitung hinzukommt, bestimmt erst die Stufe des Einsatzes. Die Achse heißt bewusst „Lenkung“ und nicht „Unterstützung“: Auch das Ausgeben einer fertigen Lösung ist eine Unterstützung, obwohl sie dem Lernen schadet.

Drei Größen sind zu unterscheiden und werden im Folgenden sprachlich getrennt gehalten. Das Kompetenzniveau (gering, mittel, hoch) beschreibt die Person. Der Lenkungsgrad eines Werkzeugs (hoch, mittel, niedrig) beschreibt, wie weit die drei Hebel im Werkzeug eingestellt sind: eingebettet, teiloffen oder offen. Der dritte Hebel heißt Kontextualisierung, damit „eingebettet“ dem Werkzeugtyp vorbehalten bleibt. Die Stufe des Einsatzes (1 Stark gelenkt, 2 Begleitet, 3 Selbstgesteuert) bezeichnet den Passungsbereich: das Kompetenzniveau der Person zusammen mit dem Arrangement aus Werkzeuglenkung und personeller Begleitung, das dazu passt. Die Stufe wird über die Person bestimmt; welches Arrangement sie einlöst, zeigen die Tabellen. Dieselbe Stufe ist deshalb mit verschiedenen Werkzeugen erreichbar, solange die Begleitung den Unterschied ausgleicht; wo im Folgenden „Stufe 1“ steht, ist die Stufe des Einsatzes gemeint, nicht der Werkzeugtyp.

Auch der Begriff „KI-Kompetenz“ braucht eine Präzisierung. Gemeint ist die **lernbezogene KI-Kompetenz in der Sache**, also die Fähigkeit, KI bei diesem

Gegenstand so zu nutzen, dass eigenes Können entsteht. Sie ist teilweise gegenstandsgebunden: Fachwissen und, in schwächerem Maß, die kritische Prüfung von KI-Ausgaben variieren bei derselben Person von Fach zu Fach. Selbstregulation und die sprachlich-kognitiven Voraussetzungen gelten üblicherweise als bereichsübergreifender. Die neuere Befundlage bewegt sich allerdings in die Gegenrichtung: Exekutive Kontrolle wird zunehmend als ziel- und kontextgebunden beschrieben (Doebel, 2020), metakognitives Monitoring hängt zwischen Arithmetik und Rechtschreibung bei Zweitklässlern noch nicht zusammen, bei Drittklässlern dagegen schon (Bellon et al., 2020; Kohortenvergleich, kein Längsschnitt), und metakognitive Profile unterscheiden sich auch darin, wie weit sie über Aufgaben reichen (van Loon & Roebers, 2024). Unter der im Abschnitt „Teilkompetenzen“ eingeführten Minimum-Regel setzen die bereichsübergreifenden Anteile eine Obergrenze, die das Fachwissen im einzelnen Gebiet nur unterschreiten, nicht anheben kann. Diese Obergrenzenannahme steht damit nicht nur unbelegt da, sondern quer zu dieser Entwicklung. Sie wird hier als konservative Arbeitsannahme beibehalten, weil sie im Zweifel mehr Lenkung vorsieht; erweisen sich Selbstregulation und Monitoring als weitgehend gegenstandsgebunden, entfällt sie, und die Stufe wäre für jedes Gebiet getrennt zu bestimmen. Diese Alternative ist mit dem Kompetenzmaß der Modellprüfung mitzuprüfen. Im Folgenden heißt sie kurz lernbezogene KI-Kompetenz. Damit die Definition nicht zirkulär wird, bestimmt das Modell die Kompetenz nicht über das Lernergebnis, sondern über die vier Teilkompetenzen. Vollständig unabhängig vom Einsatz erfassbar sind davon Fachwissen sowie sprachliche und kognitive Voraussetzungen; kritische KI-Kompetenz und Selbstregulation werden über das Verhalten in KI-nahen Situationen erschlossen und sind deshalb vor dem Einsatz an standardisiertem Material zu erheben, nicht am Verlauf des laufenden Einsatzes. Der Begriff KI-Kompetenz wird im Beitrag in drei Bedeutungen gebraucht, die auseinanderzuhalten sind: lernbezogene KI-Kompetenz als Bezugsgröße des Modells, kritische KI-Kompetenz als eine ihrer Teilkompetenzen und KI-Kompetenz im Sinne von Artikel 4 der KI-

Verordnung als Förderpflicht der Betreiberinnen und Betreiber.

Lenkung durch Werkzeug und Lehrkraft

Wie viel Lenkung ein Werkzeug mitbringt, beschreibt zunächst nur das Werkzeug. Lenkung kann aber ebenso von Personen kommen, vor allem von der Lehrkraft, die den Einsatz vorbereitet, im Raum begleitet und nachbereitet. Erst beide Quellen zusammen ergeben die Stufe des Einsatzes: Die Stufe beantwortet die Frage, wie viel Lenkung dieses Kind braucht, der Werkzeugtyp die Frage, wie viel Lenkung dieses Programm mitbringt. Sie sind dabei nicht beliebig gegeneinander austauschbar: Lenkung aus dem Werkzeug wirkt kontinuierlich, in jeder Interaktion und auch dann, wenn niemand danebensitzt; Lenkung durch die Lehrkraft ist diagnostisch reicher und flexibler, steht aber nur punktuell und nicht für alle Lernenden gleichzeitig zur Verfügung. Fehlt Lenkung im Werkzeug, muss die Lehrkraft sie in dem Moment herstellen, in dem sie gebraucht wird, und nur so lange, wie sie das leisten kann.

Erste Hinweise stützen diese Annahme. Die laufend fortgeschriebene, nicht begutachtete Metaanalyse von Strohmaier et al. (2026) zu generativer KI im Mathematiklernen findet in ihrer vierten Fassung auf Basis von 27 Studien einen positiven Gesamteffekt ($g = 0{,}49$; Kredibilitätsintervall 0,24 bis 0,75, das bayesianische Gegenstück zum Konfidenzintervall) und keinen Hinweis auf Publikationsverzerrung. Am deutlichsten unterscheiden sich die Studien darin, ob die KI den regulären Unterricht ergänzt oder die Lehrkraft ersetzt; als Ergänzung wirkt sie stärker. Spätere Fassungen können abweichen.

In dieselbe Richtung weist ein randomisierter Feldversuch, in dem nicht die Lernenden, sondern rund 900 Tutorinnen und Tutoren KI-Unterstützung erhielten: Ihre 1800 Lernenden beherrschten Themen um vier Prozentpunkte häufiger, bei den zuvor am schlechtesten bewerteten Tutorinnen und Tutoren um neun (Wang et al.,

2025; Preprint).

Einen ersten Hinweis darauf, dass Stufe 2 in der Praxis funktioniert, liefert ein randomisierter Feldversuch in Nigeria: Ein generisches, nicht eingebettetes Sprachmodell wurde sechs Wochen lang in einem lehrkraftbegleiteten Nachmittagsprogramm eingesetzt und erbrachte einen Gesamteffekt von 0,31 Standardabweichungen, im Zielfach Englisch 0,23 und in der regulären Englischprüfung des dritten Trimesters 0,21 (De Simone et al., 2025); randomisiert wurde auf Schülerebene, mit von den Autorinnen und Autoren berichteten Übertragungseffekten auf die Vergleichsgruppe, und verglichen wurde mit einer Gruppe ohne zusätzliche Lerngelegenheit, sodass der Effekt die KI-Nutzung mit zusätzlicher Unterrichtszeit vermengt, worauf die Autorinnen und Autoren selbst hinweisen. Der Befund zeigt, dass ein offenes Werkzeug mit dichter Begleitung wirken kann, nicht, dass es dem gelenkten überlegen wäre. Alle drei Befunde betreffen die Verteilung der Lenkung; ob sich Werkzeuglenkung und Begleitung in den hier angenommenen Kombinationen gegeneinander aufrechnen lassen, bleibt offen (Abschnitt „Grenzen und offene Fragen“).

In einer Klasse liegen die Stufen in der Regel nebeneinander. Praktikabel ist dann, das Werkzeug an der niedrigsten vorkommenden Stufe auszurichten und die Begleitung dort zurückzunehmen, wo die Lernenden weiter sind, und nicht umgekehrt, weil ein zu offenes Werkzeug für die Unsichersten nicht durch Aufsicht auszugleichen ist.

Auch auf Stufe 1 kann deshalb ein offener Chatbot eingesetzt werden, soweit die rechtlichen Voraussetzungen vorliegen; in der Grundschule regelmäßig nur ohne eigenes Konto der Kinder, also über einen Zugang der Lehrkraft im Plenum. Dann liegt fast die gesamte Lenkung bei der Lehrkraft. Sie wählt Aufgabe und Ausschnitt, stellt die Anweisung an die KI selbst bereit, arbeitet im Plenum oder in kleinen Gruppen, unterbricht, wenn die KI eine Lösung vorwegnimmt, und lässt die Kinder

anschließend ohne KI erklären, was sie herausgefunden haben. Das ist möglich, aber aufwendig, und es funktioniert nur so lange, wie die Lehrkraft tatsächlich präsent ist. Ein eingebettetes Werkzeug leistet dieselbe Lenkung dauerhaft, auch in der Freiarbeit oder zu Hause. Umgekehrt lässt sich ein eingebettetes Werkzeug auf Stufe 3 fast ohne Begleitung nutzen; dort kann seine Struktur allerdings stören. Die Stufe des Einsatzes ergibt sich also aus zwei Fragen: Wie viel Lenkung bringt das Werkzeug mit, und wie viel personelle Begleitung kommt hinzu? Auch hier gibt es keine scharfen Grenzen. Die Begleitung wird Schritt für Schritt zurückgenommen, so wie ein Gerüst nicht auf einmal abgebaut wird.

Personelle Lenkung setzt Einblick voraus. Lehrkraftansichten, die anzeigen, wer feststeckt, wer Hinweise überspringt und wo dieselbe Fehlvorstellung auftritt, können Aufmerksamkeit dorthin verlagern, wo sie gebraucht wird. Holstein et al. (2018) zeigten das für eine an ein Tutorsystem gekoppelte Mixed-Reality-Brille im Mathematikunterricht der Sekundarstufe I: In einem Experiment mit drei Bedingungen lernten 286 Schülerinnen und Schüler in 18 Klassen bei acht Lehrkräften mehr, wenn die Lehrkraft Echtzeitanalysen zu Lernstand, Metakognition und Verhalten erhielt. Für Rückmeldeansichten in generativen KI-Anwendungen steht eine vergleichbare Prüfung aus. Auf Stufe 1 und 2 gehört eine Lehrkraftansicht deshalb konzeptuell zur Lenkung; ohne sie bleibt die Begleitung auf Stichproben angewiesen. Rechtlich ist dabei zu unterscheiden: Eine Ansicht, die Bearbeitungsverläufe sichtbar macht und der Lehrkraft die Deutung überlässt, ist unkritisch; ein System, das den Lernstand selbst bewertet und daraus Zuweisungen oder Leistungsurteile ableitet, fällt unter Anhang III Nr. 3 der KI-Verordnung und unterliegt den Pflichten für Hochrisikosysteme. Rückmeldeansichten sind deshalb so zu gestalten, dass sie Beobachtung stützen, statt Bewertung zu automatisieren.

Stufe des Einsatzes	Eingebettetes Werkzeug (Lenkung hoch)	Teiloffenes Werkzeug (Lenkung mittel)	Offenes Werkzeug (Lenkung niedrig)
1 · Stark gelenkt geringe Kompetenz	Begleitung normal: Einführung, Beobachtung, Aufgreifen im Unterricht. Nutzung auch in Freiarbeit oder zu Hause möglich.	Nicht ausreichend, wenn die Lernenden den Dialog selbst führen: Die per Prompt oder Lernmodus gesetzte Rolle ist aushebelbar und erzeugt eine Sicherheit, die sie nicht einlöst. Führt die Lehrkraft den Dialog, gilt dieselbe Regelung wie für das offene Werkzeug.	Begleitung intensiv: Die Lehrkraft führt den Dialog selbst, im Plenum oder in Kleingruppen; gemeinsame Vorbereitung, feste Regeln, Nachbereitung ohne KI. Nicht für die unbeaufsichtigte Nutzung.
2 · Begleitet mittlere Kompetenz	Begleitung gering: Werkzeug wird geöffnet, Ergebnisse werden besprochen.	Begleitung mittel: klarer Auftrag, vorbereitete Rolle, Stichproben, Reflexion der Nutzung. Der Regelfall dieser Stufe.	Begleitung mittel bis gering: Auftrag und Transparenzregeln, Stichproben, Nutzung reflektieren.
3 · Selbstgesteuert hohe Kompetenz	Kaum Begleitung nötig; eingebaute Struktur kann bremsen.	Kaum Begleitung nötig; vorbereitete Rolle ist entbehrlich.	Kaum Begleitung nötig: Transparenzregeln vereinbaren, gelegentlich reflektieren, Lernen weiterhin ohne KI prüfen.

Das SKILL-Spektrum

Das Spektrum ordnet Werkzeuge und Einsatzformen entlang der Kompetenz. Drei Stufen strukturieren es: *Stark gelenkt*, *Begleitet* und *Selbstgesteuert*. Die Namen beschreiben die Passung für eine Person, nicht das Werkzeug. Sie sind Orientierungsmarken auf einem Kontinuum mit fließenden Übergängen. Eine Person

kann in der Sache schon weit sein, im Prüfen von KI-Antworten aber noch am Anfang stehen. Nach der vorläufigen Minimum-Regel wird sie dann Stufe 1 zugeordnet; die Hebel lassen sich innerhalb dieser Stufe aber unterschiedlich einstellen, sodass die Lenkung dort zurückgenommen wird, wo die Person weiter ist, ohne dass sich die Stufe ändert.

Die Stufen lassen sich an zwei Entwicklungsgrößen ablesen. Stufe 1 gilt, solange die Selbsteinschätzung der eigenen Lösungsgüte systematisch und deutlich über der tatsächlichen Leistung liegt und sich dieser Abstand auch dann nicht erkennbar verringert, wenn sich die Rückmeldung ausdrücklich auf die Einschätzung selbst bezieht und nicht nur auf die Leistung, und solange fremde Aussagen nur bei deutlich hervortretender Unstimmigkeit oder bei vorliegendem Quellenhinweis zurückgewiesen werden; Monitoring und Leitplanken übernimmt dann überwiegend das Werkzeug. Im ersten Schuljahr fällt die Überschätzung so durchgängig aus, dass der Indikator dort kaum zwischen Kindern trennt; er ist im frühen Grundschulalter als Erwartungswert zu lesen und erst danach als individuelles Unterscheidungsmerkmal, dessen Schwelle empirisch zu bestimmen wäre. Reines Leistungsfeedback taugt für diese Prüfung nicht, weil es in den vorliegenden Untersuchungen die Einschätzungsgüte nicht verbessert hat (Kolloff et al., 2023; Buehler et al., 2025).

Dass sich die Selbsteinschätzung im frühen Grundschulalter durch Rückmeldung nicht ohne Weiteres verbessern lässt, ist dokumentiert: Kolloff et al. (2023) gaben 112 Kindern im ersten Schuljahr (7 bis 8 Jahre) über sechs Sitzungen entweder reines Leistungsfeedback oder zusätzlich Feedback über die Übereinstimmung von Leistung und eigenem Urteil; die Kinder unterschieden in allen Sitzungen über dem Zufall zwischen richtigen und falschen Antworten und überschätzten sich zugleich durchgehend, doch verbesserte keine der beiden Varianten diese Unterscheidung. Die Studie führte keine Kontrollgruppe ohne Feedback mit; beide Bedingungen erhielten Rückmeldung. Der Indikator für Stufe 1 bezieht sich auf die Kalibrierung,

der geprüfte Nullbefund auf die Diskrimination. Für die Kalibrierung liegt ein deskriptiver Hinweis vor, weil die Überschätzung über alle sechs Sitzungen bestehen blieb; ein kontrollierter Trainingsvergleich fehlt jedoch, da beide Bedingungen Rückmeldung erhielten.

Buehler et al. (2025) fanden in einem verwandten Paradigma derselben Arbeitsgruppe, ergänzt um eine aktive Kontrollgruppe und ein Prä-Post-Design, dagegen einen kleinen Zuwachs im Unsicherheitsmonitoring, allerdings nur für das metakognitive und nicht für das leistungsbezogene Feedback und nur in gemischten Modellen, nicht in den ergänzend berichteten Varianzanalysen; die Autorinnen und Autoren mahnen selbst zur Vorsicht. Der Zuwachs betrifft damit erneut die Diskrimination, nicht die Kalibrierung, auf die sich der Indikator bezieht. Ein kontrollierter Nachweis, dass sich der Abstand zwischen Selbsteinschätzung und Leistung im frühen Grundschulalter durch Rückmeldung auf die Einschätzung selbst verringern lässt, steht aus; der Indikator setzt ihn voraus und ist insoweit eine Setzung. Er ist deshalb als Muster über mehrere Gelegenheiten hinweg zu lesen, nicht als Ergebnis eines einzelnen Rückmeldeversuchs. Gemeint ist ein Ausmaß, kein Alles-oder-nichts: Eine gewisse Überschätzung des eigenen Verstehens bleibt bis ins Erwachsenenalter bestehen (Rozenblit & Keil, 2002) und schließt die höheren Stufen nicht aus.

Stufe 2 beginnt, wo Fehler in fremden Antworten selbst gefunden werden und Hilfe erst nach einem eigenen Versuch geholt wird; das Monitoring des eigenen Stands wandert zur Person, die Leitplanken hält weiterhin das Werkzeug. Zwei Voraussetzungen kommen hier zusammen: die inhaltliche Prüfung fremder Aussagen (Sperber et al., 2010; Mills, 2013) und die rückblickende Einschätzung des eigenen Ergebnisses, die sich im Grundschulalter als einzige der geprüften metakognitiven Fertigkeiten über ein Jahr hinweg verbesserte (Bayard et al., 2021). Die Reihenfolge lässt sich damit entwicklungspsychologisch plausibilisieren, ohne als Stufenfolge belegt zu sein. Für das Jugendalter, in dem Stufe 2 und 3

typischerweise erreicht werden, liegen erst wenige Arbeiten vor: Brown und Gummerum (2025) verglichen zwei Altersgruppen (11 bis 16 Jahre, n = 153; 16 bis 20 Jahre, n = 222) und fanden bei den Älteren eine ausgeprägtere selektive Nutzung fehlerfreier Online-Quellen, besonders bei inhaltlichen gegenüber orthografischen Fehlern; ein vorgeschalteter Genauigkeitshinweis veränderte diese Auswahl nicht. Der Vergleich ist querschnittlich und erlaubt keinen Verlauf. Für generative Modelle als Auskunftsquelle steht eine entsprechende Untersuchung aus. Aus Mills' Rahmenmodell folgt keine Reihenfolge, nach der die Prüfung der Quelle der Prüfung des Inhalts entwicklungslogisch vorausginge: Beide Mechanismen sind bereits im Vorschulalter nachweisbar, die inhaltliche Prüfung allerdings deutlich schwächer ausgeprägt; was sich entwickelt, ist die Empfindlichkeit für weniger auffällige Unstimmigkeiten und die Regelmäßigkeit, mit der die Prüfung überhaupt einsetzt (Mills, 2013).

Stufe 3 setzt voraus, dass auch subtile, sprachlich gut eingebettete Unstimmigkeiten erkannt werden, für die weder ein Quellenhinweis noch eine auffällige Abweichung vom Erwartbaren einen Anhaltspunkt gibt, und dass der eigene Lernstand ohne Rückmeldung eingeschätzt wird; dass dieser anspruchsvollste Indikator erst hier steht, passt zu dem Befund, dass sich das vorausschauende Monitoring im Grundschulalter über ein Jahr hinweg nicht verbesserte (Bayard et al., 2021).

Die Übernahme des Monitorings durch das Werkzeug auf Stufe 1 ist eine Entlastung auf Zeit, kein Dauerzustand. Selbstreguliertes Lernen ist bereits im Grundschulalter durch Unterricht förderbar: Dignath und Büttner (2008) berichten in getrennten Metaanalysen für Grundschule (49 Studien) und Sekundarstufe (35 Studien) über 357 Effektstärken einen mittleren Gesamteffekt von d = 0,69 über alle Ergebnisvariablen, also Schulleistung, Strategienutzung und Motivation zusammen; für die Schulleistung allein fallen die Werte je Schulstufe niedriger aus, und die Effekte fallen höher aus, wenn die Trainings metakognitive Reflexion einschließen,

also vermitteln, wann, warum und wie eine Strategie einzusetzen ist. Die Effekte stammen überwiegend aus kurzen, von Forschenden durchgeführten Trainings; sie halten jedoch über die Zeit: In einer Metaanalyse von 48 Interventionen fiel der Effekt im Follow-up höher aus als unmittelbar danach ($g = 0{,}63$ gegenüber 0,50; der Zuwachs selbst ist mit $g = 0{,}12$ klein), und Lernende aus sozial benachteiligten Verhältnissen profitierten langfristig am stärksten (de Boer et al., 2018). Ob sich dadurch auch die Genauigkeit der Selbsteinschätzung verbessert, führt keine der beiden Metaanalysen als Ergebnisvariable. Zur Lenkung auf Stufe 1 gehört deshalb, dass das vom Werkzeug übernommene Monitoring im Unterricht zum Thema wird: einschätzen vor der Rückmeldung, Einschätzung und Ergebnis vergleichen, den eigenen Weg ohne KI erklären. Ohne diesen Anteil stabilisiert Stufe 1 die Voraussetzung, die sie überwinden soll.

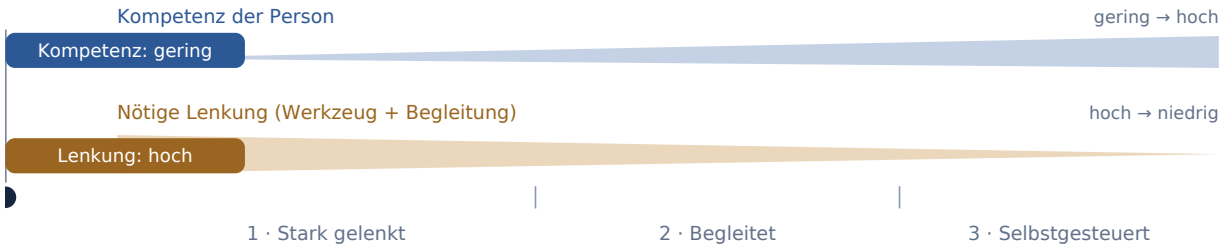
In der interaktiven Darstellung lässt sich der Punkt frei verschieben; alternativ ist eine der drei Stufen wählbar. Die Darstellung zeigt zu jeder Stufe die passende Ausprägung der vier Teilkompetenzen, die Einstellung der drei Lenkungshebel, den nötigen Umfang der Begleitung, die Aufgabe der Lehrkraft und die Anzeichen dafür, dass die nächste Stufe erreicht ist.

Vorfrage: Leistet KI hier etwas, das Material, Gespräch und eigenes Probieren nicht leisten? Wenn nein, beginnt das Spektrum nicht.

Spektrum erkunden

Punkt entlang der Achse ziehen oder eine Stufe wählen.

WO STEHT DIE PERSON?



1Stark gelenkt 2Begleitet 3Selbstgesteuert

WAS FOLGT DARAUS?

Stufe 1 · Stark gelenkt
Person Werkzeug Lehrkraft
Fachwissen in der Sache Kritische KI-Kompetenz Selbstregulation Sprache und Kognition
WERKZEUG ROLLE DER KI HILFEDICHTE · LEITPLANKEN · KONTEXTUALISIERUNG PERSONELLE BEGLEITUNG WAS DIE LEHRKRAFT TUT BEISPIEL AUS DEM UNTERRICHT
Werkzeugwahl auf dieser Stufe: eingebettet oder offen?

Woran man erkennt, dass die nächste Stufe passt
Typisches Risiko auf dieser Stufe

Vorab zu klären: Lernwerkzeug oder Nachteilsausgleich? · Die Stufen sind Orientierungsmarken auf einem Kontinuum. Stufenidee nach Pearson & Gallagher (1983) und Rowe (2026); Abwägen von Geben und Zurückhalten nach Koedinger & Aleven (2007); Teilkompetenzen nach UNESCO (2024) sowie OECD & Europäische Kommission (2026).

Folgerungen für KI im Unterricht

Auch erfahrene Personen können bei einem neuen oder wenig vertrauten Gegenstand über geringes relevantes Vorwissen verfügen. Die offene Nutzung frei zugänglicher Chatbots kann deshalb nicht der Normalfall des Lernens sein; sie ist die dritte Stufe. Für die meisten Lernsituationen in der Schule ist Stufe 1 oder 2 angemessen: KI, die in eine Lernanwendung eingebettet ist und im Hintergrund einer didaktisch gestalteten Umgebung arbeitet, oder ein offenes Werkzeug mit klarem Auftrag, vorbereiteter Rolle und Begleitung durch die Lehrkraft. In beiden Fällen bleibt die Lehrkraft Teil des Arrangements. Die Frage lautet, welche Anteile der Lenkung sie übernimmt und welche das Werkzeug.

Das Modell hat eine Gerechtigkeitsdimension. Wenn lernwirksame KI-Nutzung Vorwissen, sprachliche Sicherheit, Selbstregulation und technisches Verständnis voraussetzt, profitieren zuerst diejenigen, die all das schon mitbringen. Lehmann et al. (2025) berichten eine solche Schere entlang des Vorwissens bei Studierenden; der Befund ist explorativ. Ob sich die Schere auch entlang sozialer Herkunft öffnet, ist damit nicht gezeigt, liegt aber nahe, weil die genannten Voraussetzungen selbst herkunftsabhängig verteilt sind. Eingebettete Werkzeuge senken die Einstiegshürde und lassen sich mit wachsender Kompetenz öffnen. Das Modell erwartet deshalb, dass sie die Ergebnisse gleichmäßiger verteilen als der offene Chatbot, der zwar allen offensteht, den aber vor allem diejenigen produktiv nutzen, die die genannten Voraussetzungen schon mitbringen. Diese Erwartung ist nicht geprüft. Die

Befundlage ist gemischt: Lehmann et al. (2025) berichten explorativ eine Schere entlang des Vorwissens, Fischer et al. (2025) die größten Gewinne bei geringem Vorwissen, und Bastani et al. (2025) fanden in präregistrierten Moderatoranalysen kaum Hinweise auf heterogene Effekte; ihre Streuungsanalyse zeigt eine Kompression der Leistungsspanne nur im gestützten Üben, nicht in der Prüfung ohne KI. Keine dieser Studien prüft den Kontrast zwischen eingebettetem und offenem Werkzeug. Außerhalb des Bildungsbereichs weist ein randomisierter Feldversuch in die erwartete Richtung: Bei Kleinunternehmerinnen und Kleinunternehmern in Kenia verschlechterte der Zugang zu einem KI-Assistenten die Ergebnisse der zuvor schwächeren Hälfte, während die stärkere profitierte (Otis et al., 2026). Auf Lernergebnisse übertragbar ist das nicht. Solange weder differenzielle Effekte für diesen Kontrast noch Streuungsmaße vorliegen, bleibt der Satz eine Hypothese.

Für Lehrkräfte gilt die Logik ebenfalls. Wer fachlich sicher ist und Erfahrung mit Sprachmodellen hat, kann offene Werkzeuge für Planung, Materialerstellung und Organisation produktiv einsetzen. Andere brauchen Werkzeuge, die ohne ausgeprägte Prompting-Kompetenz funktionieren: mit hinterlegten Rollen, fachlich geprüften Wissensbeständen, vorbereiteten Bausteinen und gezielten Rückfragen. Auch hier sollte sich das Werkzeug öffnen lassen, sobald die Souveränität wächst. Der europäische Referenzrahmen für die Digitalkompetenz Lehrender (Redecker, 2017) beschreibt dieselbe Bewegung von der Anwendung vorbereiteter Ressourcen zu deren eigenständiger Gestaltung.

Vier Fragen zur Planung eines KI-Einsatzes

Mit den vier Fragen wird eine einzelne Unterrichtssituation geplant und die Stufe gewählt; mit der Übersicht danach werden die Schülerinnen und Schüler eingeschätzt.

1. **Leistet KI hier etwas, das ich mit Material, Gespräch und eigenem**

Probieren nicht erreiche? Wenn nein, kein KI-Einsatz.

2. **Was sollen die Kinder lernen, und welchen kognitiven Prozess würde die KI dabei übernehmen?** Ist das genau der, den sie lernen sollen, braucht der Lösungsprozess Leitplanken; ist er es nicht, darf die KI ihn übernehmen. Für Kinder mit Beeinträchtigungen wird die KI dann vom Lernwerkzeug zum Mittel der Teilhabe, etwa als Vorlesefunktion beim Lernziel Textverständnis; hier ist Fading nicht das Ziel, die Unterstützung darf bleiben.
3. **Wie sicher sind die Lernenden in der Sache, im Prüfen von KI-Antworten und in der Selbststeuerung?** Je unsicherer, desto eher Stufe 1: KI in einer Lernanwendung mit fester Rolle, oder ein offener Chatbot, den ich selbst intensiv vor- und nachbereite und im Raum begleite. Mittlere Sicherheit: Stufe 2 mit klarem Auftrag und Begleitung. Erst bei hoher Sicherheit Stufe 3. Werkzeuglenkung und meine Begleitung müssen zusammen zur Kompetenz passen. Drei Anhaltspunkte: Findet die Person Fehler in einer KI-Antwort selbst? Fragt sie nach einem eigenen Versuch oder davor? Kann sie nach einer Aufgabe einschätzen, was ihr davon ohne KI wieder gelingen würde (Stufe 2), und vor einer neuen Aufgabe, wo ihr Verstehen endet (Stufe 3)? Dreimal nein spricht für Stufe 1, dreimal ja für Stufe 3; die Übersicht am Ende führt die Indikatoren aus.
4. **Woran erkenne ich nachher, ob gelernt wurde?** An einer Leistung ohne KI, an Erklärungen in eigenen Worten, am Transfer auf eine neue Aufgabe. Nicht an der Qualität des Produkts, das mit KI entstanden ist.

Beispiel: Sachaufgaben in Klasse 4

Klasse 4, Sachaufgaben. Ein Kind liest langsam und bleibt oft schon am Text hängen. Frage 1: Material und Gespräch helfen beim Rechnen, aber nicht bei

der Textmenge; eine KI, die den Text vorliest und nach dem Gesuchten fragt, leistet hier etwas Eigenes. Frage 2: Gelernt werden soll das Modellieren, nicht das Lesen; das Vorlesen darf die KI übernehmen und dauerhaft behalten, den Lösungsweg nicht. Frage 3: Das Kind hält KI-Antworten für richtig und fragt lieber, als selbst anzufangen; also Stufe 1, ein eingebetteter Tutor mit Vorlesefunktion und gesperrtem Endergebnis, dazu die Lehrkraft, die die Rückmeldungen im Gespräch aufgreift. Frage 4: Nach zwei Wochen erklärt das Kind eine neue Aufgabe ohne KI in eigenen Worten; ob der Lösungsweg verstanden ist, zeigt sich daran, nicht an den gelösten Aufgaben in der App. Beim Werkzeug kommt es darauf an, dass sich die Vorlesefunktion unabhängig von den übrigen Hilfen steuern lässt, damit sie als Teilhabemittel bleiben kann, während die Hinweise zum Lösungsweg zurückgenommen werden. In der Übersicht am Ende steht das Kind bei allen Indikatoren auf Stufe 1; die Zuordnung wird nach zwei Wochen an denselben Indikatoren erneut geprüft.

Übersicht: Indikatoren und Einsatzszenarien je Stufe

Die folgende Übersicht ist eine Planungshilfe, kein Diagnoseinstrument: Ihre Stufenzuordnung beruht auf theoretischer Ableitung und ist empirisch nicht validiert; sie ersetzt weder die fachliche Einschätzung der Lehrkraft noch eine Prüfung des Lernens ohne KI. Dasselbe gilt für das Handout. Die Übersicht fasst für Lehrkräfte zusammen, woran sich die Stufen bei den Lernenden erkennen lassen, welche Einsatzszenarien passen, wie viel Begleitung nötig ist und woran Fortschritt sichtbar wird. Die Spalte „Woran Fortschritt erkennbar ist“ nennt ausschließlich beobachtbares Verhalten. Ob gelernt wurde, wird getrennt und ohne KI geprüft (vierte Planungsfrage); stützte sich die Stufenzuordnung darauf, setzte sie voraus, was sie erklären soll. Die Indikatoren dienen der formativen Einschätzung im Unterricht. Als Maß der Kompetenz für die empirische Prüfung des Modells taugen sie nicht, weil mehrere von ihnen, etwa das Nachfragen erst nach einem eigenen

Versuch, genau das Verhalten sind, das die Lenkung erzeugen soll; dort wird die Kompetenz vor der Intervention mit KI-fernen Maßen erhoben (Abschnitt „Für die Forschung“). Die Übersicht steht auch als Handout (PDF, zwei Seiten A4 quer) zum Ausdrucken bereit.

Stufe des Einsatzes	Indikatoren bei den Schülerinnen und Schülern	Passende Einsatzszenarien	Begleitung durch die Lehrkraft	Woran Fortschritt erkennbar ist
1 · Stark gelenkt geringe Kompetenz	Kann fachliche Richtigkeit kaum beurteilen; hält KI-Antworten für zuverlässig; übernimmt Antworten und hält die eigene Lösung meist für besser, als sie ist, bemerkt im Einzelfall aber durchaus, welche Antworten unsicher sind; längere Antworten überfordern. Typisch bei neuem Thema und geringem Vorwissen.	Nur in wenigen, klar umrissenen Situationen. Wenn KI: Tutor in einer Lernanwendung mit hinterlegter Aufgabe; gestufte Hinweise statt Lösungen; Ergebnis gesperrt bis zum eigenen Versuch. Freiarbeit und Üben zu Hause nur mit eingebettetem Werkzeug. Offener Chatbot nur im Plenum oder in Kleingruppen mit Lehrkraft im Gespräch.	Eingebettetes Werkzeug: einführen, beobachten, aufgreifen; das vom Werkzeug übernommene Monitoring zum Thema machen (vorher einschätzen, mit dem Ergebnis vergleichen, den eigenen Weg ohne KI erklären). Offener Chatbot: intensiv, gemeinsam vorbereiten, Anweisung an die KI selbst stellen, im Raum bleiben, eingreifen, ohne KI nachbereiten.	Löst nach einem Hinweis selbst weiter; fragt nach einem eigenen Versuch statt vor ihm; erklärt den eigenen Weg in eigenen Worten; erkennt unpassende Hinweise.

Stufe des Einsatzes	Indikatoren bei den Schülerinnen und Schülern	Passende Einsatzszenarien	Begleitung durch die Lehrkraft	Woran Fortschritt erkennbar ist
2 · Begleitet mittlere Kompetenz	Erkennt grobe Fehler und prüft Teilergebnisse; kennt Grenzen der KI; stellt eigene Fragen, gibt Kontext aber noch nicht gezielt vor; schätzt unmittelbar nach der Bearbeitung realistisch ein, wie gut die eben bearbeitete Aufgabe gelungen ist, und leitet daraus ab, was davon ohne KI wieder gelingen würde.	Offenes Werkzeug mit vorbereiteter Rolle und klarem Auftrag; zwei KI-Lösungswege vergleichen und begründet auswählen; KI-Erklärungen auf Fehler prüfen; KI-Rückmeldung zum eigenen Entwurf einholen und überarbeiten.	Mittel: Auftrag und Rolle bereitstellen, Stichproben, Ergebnisse ohne KI erklären lassen, Nutzung gemeinsam reflektieren.	Findet Fehler in KI-Antworten selbst; gibt Kontext und Ziel vor; begründet, wann KI sinnvoll war und wann nicht.
3 · Selbstgesteuert hohe Kompetenz	Durchschaut auch plausible Fehler; prüft systematisch an Quellen; plant, überwacht und bewertet den Einsatz selbst; schätzt in der Regel zutreffend ein, wo das eigene Verstehen endet, bevor sich das in einer Aufgabe zeigt; eine Restüberschätzung bleibt auch hier.	Frei genutzter Chatbot als Denkpartner: Gegenargumente einholen, Entwürfe kritisieren lassen, Recherche mit Quellenprüfung, eigene Rollen gestalten. Eingebettete Struktur stört nur, wenn sie Bekanntes redundant wiederholt oder nicht abschaltbar ist.	Gering: Transparenzregeln, gelegentliche Reflexion, Lernen weiterhin ohne KI prüfen.	Nutzt KI gezielt und begründet; wechselt bei einem neuen Fachgebiet von sich aus in eine engere Stufe zurück; erkennt selbst, wenn die KI zur Abkürzung geworden ist.

Die Indikatoren beschreiben beobachtbares Verhalten, keine Altersgruppen. Die Entwicklungsbefunde setzen ihnen aber einen Erwartungskorridor: Das vorausschauende Einschätzen des eigenen Lernstands, das Stufe 3 verlangt, verbesserte sich im Grundschulalter über ein Jahr hinweg nicht (Bayard et al., 2021); für die anschließende

Entwicklung liegt nur indirekte Evidenz vor: Weil et al. (2013) berichten einen Anstieg der metakognitiven Diskrimination, nicht der Kalibrierung, und dies in einer Wahrnehmungsaufgabe. Für das vorausschauende Monitoring, das Stufe 3 verlangt, liegt jenseits des Grundschulalters keine einschlägige Evidenz vor; Bayard et al. (2021) zeigen, dass es sich in zwei Kohorten über je ein Jahr hinweg nicht verbesserte, von Klasse 2 zu 3 und von Klasse 4 zu 5, nicht, dass es sich später verbessert. Dass dieser Indikator erst auf Stufe 3 steht, ist eine Setzung, keine Ableitung. Der Unterschied zu Stufe 2 liegt nicht darin, ob überhaupt vorausschauend geurteilt wird, sondern im Bezugspunkt: Auf Stufe 2 wird über eine eben bearbeitete Aufgabe geurteilt, auf Stufe 3 über eine noch nicht bearbeitete; dass Urteile mit dem näheren Bezugspunkt genauer ausfallen, ist für Erwachsene gut belegt und mit Bayard et al. (2021) vereinbar, dort aber nicht geprüft. Zudem ist metakognitive Genauigkeit zwischen Domänen nur teilweise gekoppelt: In einer Übersichtsarbeit mit eingebetteter Metaanalyse über 19 Experimente fanden Rouault et al. (2018) einen starken Zusammenhang zwischen Sinnesmodalitäten ($r = 0{,}55$), zwischen Wahrnehmung und Gedächtnis dagegen keinen belastbaren ($r = 0{,}09$), was sie unter anderem auf begrenzte Teststärke und auf Unterschiede im Aufgabenformat zurückführen; beide Teilanalysen beruhen auf 9 beziehungsweise 10 Effektstärken. Mazancieux et al. (2020) fanden mit einem hierarchisch-bayesianischen Verfahren bei 181 Erwachsenen fünf von sechs Aufgabenpaaren korreliert; deutlich zwischen visueller Wahrnehmung und semantischem Gedächtnis, nicht belastbar dagegen zwischen Wahrnehmung und episodischem Gedächtnis, und sie schließen auf ein Zusammenwirken domänenübergreifender und domänenspezifischer Anteile. Beide Arbeiten betreffen interindividuelle Unterschiede bei Erwachsenen, nicht die Frage, ob sich ein Entwicklungsverlauf von einer Wahrnehmungs- auf eine fachliche Aufgabe überträgt; für fachliches Monitoring im Grundschulalter fand ein Vergleich zweier unabhängiger Querschnittsstichproben bereichsübergreifende Zusammenhänge zwischen Arithmetik und Rechtschreibung erst bei Drittklässlern ($n = 147$), nicht bei Zweitklässlern ($n = 77$); die Deutung als Entwicklungsverlauf stammt von den Autorinnen und Autoren, das Design prüft sie nicht (Bellon et al., 2020). Die Übertragung des an einer Wahrnehmungsaufgabe gewonnenen Verlaufs bleibt damit eine Annahme. Die für Stufe 3 angenommenen Voraussetzungen dürften im Grundschulalter nach der derzeitigen Evidenz deshalb nur selten vollständig vorliegen, auch bei hohem Fachwissen; die Aussage betrifft

Gruppenmittel, nicht das einzelne Kind, und umgekehrt sagt ein höheres Alter für sich genommen nichts über die Stufe. Zusammen mit der Minimum-Regel wirkt der Erwartungskorridor in der Praxis wie eine Altersschranke für Stufe 3, allerdings auf zwei verschiedenen Wegen. Individuell greift die Regel nur, wenn die Selbstregulationsanteile bei diesem Kind gemessen niedrig ausfallen; ein Kind mit hohen Werten in allen vier Teilkompetenzen wird von ihr nicht ausgeschlossen. Für Klassen und Jahrgänge folgt die Schranke daraus, dass niedrige Werte im Grundschulalter der Regelfall sind. Dass beides praktisch zusammenfällt, ist eine bewusste konservative Setzung dieses Beitrags, kein aus den Daten abgeleiteter Befund; wie häufig die Ausnahme ist, bleibt eine empirische Frage, und bestätigt sich in der Prüfung der kompensatorische Summenwert, entfällt die Schranke.

Folgerungen für die Entwicklung von Lernanwendungen

Für die Konzeption von Lernanwendungen dient der Ordnungsrahmen dazu, eine Anwendung zu verorten. Sie wird zwischen eingebetteter KI und offenem Chatbot positioniert, ausgehend von der Zielgruppe und ihrer erwartbaren Kompetenz. Daraus folgen konkrete Entscheidungen: Wie viel Fachkontext ist hinterlegt, welche Ausgaben sind gesperrt, wann und wie wird Hilfe gestuft, wie wird Fading umgesetzt, welche Daten sieht die Lehrkraft. Vier Gestaltungsprinzipien, die sich aus den hier referierten Befunden ableiten lassen (fachdidaktische Fundierung der Rückmeldungen, Dosierung nach dem Prinzip „erst du, dann ich“, eigene Tätigkeit vor der KI-Ausgabe und Verzahnung mit dem umgebenden Lernarrangement), führt ein eigener Beitrag aus. Geprüft sind sie nicht.

Grenzen und offene Fragen

Das SKILL-Modell ist ein Ordnungsrahmen, kein empirisch geprüftes Wirkmodell. Seine drei Stufen vereinfachen bewusst: Kompetenz und Lenkung sind stetige Größen, die Übergänge fließend, und die Teilkompetenzen entwickeln sich nicht im

Gleichschritt. Fünf Fragen bleiben offen. Die Diagnose: Wie lässt sich die lernbezogene KI-Kompetenz einer Person ökonomisch erfassen, sodass die Stufenzuordnung nicht auf Vermutungen beruht? Das Fading: Wann und in welchen Schritten sollte Lenkung zurückgenommen werden, und woran erkennt ein Werkzeug diesen Zeitpunkt? Die genannte Fehlregulation der Hilfesuche (Aleven et al., 2016; Darvishi et al., 2024) zeigt, dass adaptive Regeln dafür erst in Ansätzen erprobt sind.

Die Verrechnung: Dass generative KI stärker wirkt, wenn sie den Unterricht ergänzt, statt die Lehrkraft zu ersetzen, ist als Moderator zwischen Studien metaanalytisch dokumentiert (Strohmaier et al., 2026, vierte Fassung); ein randomisierter Vergleich innerhalb einer Studie liegt nicht vor, und der Kontrast vermengt die Begleitung mit dem Umfang des Unterrichts insgesamt. Ob sich Werkzeuglenkung und Begleitung in den vom Modell angenommenen Kombinationen gegeneinander aufrechnen lassen, also viel Werkzeuglenkung mit wenig Begleitung gegenüber wenig Werkzeuglenkung mit viel Begleitung, ist nicht untersucht.

Die Übertragbarkeit: Die meisten belastbaren Studien stammen aus der Sekundarstufe und der Hochschule. Für die Grundschule liegen erste kontrollierte Untersuchungen vor, etwa zu einer KI-gestützten Sachrechenumgebung (Liu et al., 2025) und zum Vergleich von Chatbot- und lehrkraftgeleitetem Lesen (Zhou et al., 2026), doch messen sie die Leistung bei verfügbarer Unterstützung und unmittelbar nach der Intervention. Studien, die im Grundschulalter verzögert und ohne KI prüfen oder Lenkungsgrade gegeneinanderstellen, fehlen.

Die Richtung: Das Modell nimmt an, dass mit wachsender lernbezogener KI-Kompetenz weniger Lenkung nötig wird. Fernandes et al. (2026) berichten dagegen für Erwachsene, dass die Leistung unter KI-Nutzung steigt, während sich die metakognitive Genauigkeit von ihr entkoppelt, und dass höhere selbstberichtete KI-Kompetenz mit geringerer, nicht höherer metakognitiver Genauigkeit einhergeht.

Trifft das zu, ist der Zusammenhang zwischen Werkzeugroutine und Urteilsfähigkeit nicht monoton, und die kritische KI-Kompetenz muss am Prüfverhalten gemessen werden, nicht an Erfahrung oder Selbsteinschätzung. Der Befund zur KI-Kompetenz ist korrelativ; die Leistungsbefunde stammen aus einer quasi-experimentellen Studie mit externer Vergleichsgruppe und einer randomisierten Studie, in der sich beide Gruppen überschätzten, die Gruppe ohne KI eher stärker, sodass KI die Überschätzung nicht erzeugt, sondern die Leistung ohne entsprechenden Zuwachs der Selbsteinschätzung anhebt; erhoben wurde an Erwachsenen. Der Befund gehört gleichwohl in das Prüfprogramm der Grundannahme.

Ob sich Lenkungsgrade in Lernergebnisse übersetzen, ist zudem offen. Fütterer et al. (2026) verglichen in einem randomisierten Feldversuch mit 371 Lernenden der Klassen 7 bis 9 zwei KI-gestützte Interventionen zur Selbstregulation, eine motivationale und eine strategiebezogene, mit Standard-ChatGPT über sechs Unterrichtsstunden. Für Wissenszuwachs, Anstrengung und Strategienutzung fand sich kein Vorteil gegenüber der Kontrollbedingung; die motivationale Variante entwickelte den wahrgenommenen Nutzen lediglich günstiger als die strategiebezogene, nicht günstiger als Standard-ChatGPT. Geprüft wurden dort vorgeschaltete Lernstrategie-Prompts, nach der hier vorgeschlagenen Ordnung durchweg Varianten der Stufe 2, und nicht der Kontrast zwischen anwendungsseitigen Leitplanken und offener Nutzung. Ein direkter Test der Modellannahme steht damit aus; das Ergebnis mahnt aber, dass didaktische Zusätze nicht von sich aus wirken. Dieselbe Mahnung enthält der Nullbefund von Bassner et al. (2026) im KI-freien Wissenstest (Abschnitt „Begriffsklärung“). Auch die Gerechtigkeitserwartung aus dem Abschnitt „Folgerungen für KI im Unterricht“ bleibt aus diesem Grund ungeprüft.

Für die Forschung: die Grundannahme als prüfbare Hypothese

Der folgende Abschnitt richtet sich an Forschende; für die Unterrichtsplanung ist er

entbehrlich. Vorab ist festzulegen, was als nicht-kontingente Lenkung gilt: Lenkung, die unabhängig vom Bearbeitungsstand ausgelöst wird, etwa eine feste Hinweisreihenfolge, nicht abschaltbare Zwischenschritte oder eine vorgegebene Aufgabengliederung.

Die Grundannahme lautet: Je geringer die lernbezogene KI-Kompetenz, desto größer der Vorteil eines stark gelenkten gegenüber einem offenen Werkzeug in einem verzögerten Test ohne KI; als Entscheidungsschwelle für die konfirmatorische Prüfung wird ein Vorteil von mindestens d = 0,30 für Personen festgelegt, die eine Standardabweichung unterhalb des Kompetenzmittels liegen. Die Schwelle lässt sich nicht aus den vorliegenden Metaanalysen ableiten, deren Schätzungen (Belland et al., 2017: g = 0,46; Lazonder & Harmsen, 2016: d = 0,50) gegebene Unterstützung betreffen, während der vorhergesagte Kontrast wesentlich auf Leitplanken und Kontextualisierung beruht. Sie ist keine Aussage der Grundannahme, sondern der kleinste Effekt, der für die hier skizzierte Studie noch von Interesse wäre. Begründet ist sie so: Kraft (2020) stuft in Feldstudien Effekte ab 0,20 als groß ein, allerdings bezogen auf breite, standardisierte Leistungsmaße gegenüber dem regulären Unterricht. Zwei Umstände verschieben den Maßstab in entgegengesetzte Richtungen. Der hier vorhergesagte Kontrast vergleicht zwei aktive KI-Bedingungen bei gleicher Aufgabe und Zeit, was kleinere Effekte erwarten lässt; zugleich wird mit forschungseigenen Maßen gemessen, bei denen Effektstärken systematisch etwa doppelt so hoch ausfallen wie bei unabhängigen Instrumenten (Cheung & Slavin, 2016). Die Schwelle von 0,30 ist deshalb keine aus Kraft abgeleitete Größe, sondern eine praktische Setzung: Sie markiert den kleinsten Effekt, der Beschaffung, Einarbeitung und Begleitung einer Werkzeugentscheidung im Unterricht rechtfertigen würde. Diese Abwägung ist normativ; sie macht die Prüfung bewusst streng und erhöht das Risiko, einen realen, aber kleinen Effekt als praktisch unerheblich einzustufen. Wird mit einem unabhängigen, standardisierten Maß gemessen, ist die Schwelle in der Präregistrierung abzusenken; entscheidend ist, dass sie vor der Erhebung feststeht.

Gegen die vorliegende Evidenz ist sie ambitioniert. Von den beiden Studien, die den Kontrast im KI-freien Maß bislang prüfen, findet nur Bastani et al. (2025) einen Vorteil der gelenkten gegenüber der offenen Variante, dort hielt die gelenkte Bedingung das Kontrollniveau, während die offene 17 Prozent darunter lag; Bassner et al. (2026) finden keinen. Gegenüber dem Üben ohne KI zeigt in beiden Studien auch die gelenkte Variante keinen Vorteil, und beide messen unmittelbar, nicht verzögert. Die Grundannahme behauptet die Wechselwirkung: Der Vorteil starker Lenkung ist bei geringer Kompetenz wesentlich größer als bei hoher. Ihre starke Fassung behauptet zusätzlich eine Expertise-Umkehr: Bei hoher Kompetenz zeigt nicht-kontingente Lenkung keinen Vorteil oder einen Nachteil. Für kontingente Lenkung macht das Modell keine gesonderte Vorhersage: Weil Kontingenz hier bedeutet, dass Hilfe nur nach Bedarf und nur im nötigen Umfang greift, wäre ein ausbleibender Nachteil eine Folge der Begriffsbestimmung, kein Befund. Prüfbar ist allein, ob die als kontingent implementierte Bedingung im Manipulationscheck tatsächlich bedarfsabhängig auslöste; misslingt das, gilt sie als nicht-kontingent. Als bedarfsabhängig gilt eine Bedingung nur, wenn die Auslöseregel vor der Erhebung als beobachtbare Funktion des Protokolls festgeschrieben ist und im Protokoll nachweislich zu unterschiedlichen Hilfeumfängen bei unterschiedlichen Lernenden geführt hat; der bloße Nachweis, dass eine Regel hinterlegt war, genügt nicht.

Formal ist dies eine Wechselwirkung von Personenmerkmal und Instruktionsform, in der Tradition der Forschung zu Aptitude-Treatment-Interaktionen (Cronbach & Snow, 1977); der Expertise-Umkehr-Effekt ist ihr am besten belegter Vertreter. Diese Tradition mahnt zur Vorsicht: Interaktionseffekte fallen kleiner aus als Haupteffekte, verlangen größere Stichproben und replizieren häufig nicht, wenn das Personenmerkmal unscharf gemessen ist. Die Grundannahme ist deshalb nur so belastbar wie das Maß der lernbezogenen KI-Kompetenz, das sie voraussetzt.

Die folgenden Kriterien setzen zudem eine Präzision voraus, die ein einzelner Feldversuch mit Randomisierung auf Klassenebene kaum erreicht: Der Ausschluss

einer Wechselwirkung von |Δd| = 0,20 verlangt ein Vielfaches der Fallzahl, die für ihren Nachweis nötig wäre. Die Prüfung ist deshalb gestuft anzulegen: Eine erste Studie prüft konfirmatorisch nur den bedingten Vorteil starker Lenkung im unteren Kompetenzbereich; der Ausschluss der Moderation insgesamt bleibt einer Mehrstandortstudie oder einer prospektiven Metaanalyse vorbehalten. Bis dahin gilt die Grundannahme als ungeprüft, nicht als bestätigt.

Als widerlegt gilt die Grundannahme in drei Fällen: wenn ein Äquivalenztest für die Differenz der bedingten Effekte bei einer Standardabweichung unter und über dem Kompetenzmittel gegen die Grenze |Δd| = 0,20 signifikant ausfällt, Lenkung also unabhängig von der Kompetenz wirkt; wenn für den unteren Kompetenzbereich (eine Standardabweichung unter dem Mittel) die obere Grenze des 95-Prozent-Konfidenzintervalls für den Vorteil starker Lenkung unter d = 0,30 liegt, wobei dieses Kriterium nicht die Richtungsaussage widerlegt, sondern ihre praktische Relevanz, denn ein kleinerer gleichgerichteter Effekt bliebe möglich, rechtfertigte aber keine Werkzeugentscheidung im Unterricht; oder wenn die Wechselwirkung das entgegengesetzte Vorzeichen hat, der Vorteil starker Lenkung bei hoher Kompetenz also den bei geringer Kompetenz übersteigt und das Konfidenzintervall dieser Differenz die Null ausschließt. Die starke Fassung gilt darüber hinaus als widerlegt, wenn sich bei hoher Kompetenz ein Vorteil auch für nicht-kontingente Lenkung zeigt, dessen Konfidenzintervall die Null ausschließt; die Grundannahme bleibt davon unberührt, solange der Vorteil bei geringer Kompetenz wesentlich größer ist. Nicht signifikante Befunde ohne diese Präzision gelten als unentschieden, nicht als Widerlegung. Die Zuordnung einer Bedingung zu kontingent oder nicht-kontingent wird vor der Datenerhebung festgeschrieben.

Ein Design dafür präregistriert Hypothesen, Auswertungsplan und Abbruchregeln. Es erhebt die vier Teilkompetenzen vor der Intervention mit KI-fernen Maßen, die von der Lenkung nicht beeinflusst sind: ein Vorwissenstest, die Fehlererkennung an vorgelegten KI-Antworten, ein Selbstregulationsmaß sowie Lese- und

Arbeitsgedächtnisproben. Die Verhaltensindikatoren der Übersicht bleiben der formativen Einschätzung vorbehalten; als Moderator würden sie die Prüfung zirkulär machen, weil sie teilweise Wirkungen der Lenkung sind. Die Teilwerte werden nach einer vorab festgelegten Regel zu einem kontinuierlichen Moderator auf Personenebene verdichtet.

Als erste konfirmatorische Spezifikation wird der gleichgewichtete Summenwert aus standardisierten Teilwerten vorab festgelegt, weil er der messgenaueste Kandidat ist. Die Minimum-Regel aus dem Abschnitt „Teilkompetenzen“ ist keine Aussage der Grundannahme, sondern die Praxisheuristik dieses Beitrags. Weil die Praxisempfehlungen auf ihr beruhen, wird sie als vorab registrierte zweite Hypothese mit eigener Alpha-Korrektur geführt: Sagt das Minimum den Lernzuwachs unter Lenkung besser vorher als der Summenwert, sind die Teilkompetenzen als notwendige Bedingungen zu behandeln; bleibt die Regel unbestätigt, während sich der Summenwert bestätigt, sind Übersicht und Handout auf eine kompensatorische Zuordnung umzustellen. Die Minimum-Regel ist messtheoretisch anspruchsvoller, weil das Minimum aus vier fehlerbehafteten Werten nach unten verzerrt und weniger reliabel ist als jedes Einzelmaß. Das wiegt schwer, weil im Grundschulalter zwar die Monitoring-Urteile selbst intern konsistent ausfallen, die personenbezogenen Genauigkeitsmaße dagegen gering reliabel sind und über sechs Monate erheblich schwanken (Roebers et al., 2021), und weil sich dort latente Profile finden, die sich nicht nur im Niveau, sondern auch darin unterscheiden, wie weit Monitoring und Kontrolle über zwei verschiedene Aufgaben hinweg reichen; in Klasse 4 bleiben schwache Profile über ein Jahr hinweg stabil, in Klasse 2 nicht (van Loon & Roebers, 2024). Die Teilkompetenzen sind deshalb mit mehreren Indikatoren zu erheben, und ihre Reliabilität ist vor der konfirmatorischen Analyse zu berichten; andernfalls ist ein ausbleibender Interaktionseffekt nicht von Messfehlerdämpfung zu unterscheiden.

Die vier Teilkompetenzen als getrennte Moderatoren dienen als Robustheitsprüfung.

Zeigt sich der erwartete Interaktionsterm unter keiner der vorab benannten Regeln, gilt die Grundannahme als nicht gestützt; zeigt er sich nur unter einzelnen, bleibt offen, ob es an der Bezugsgröße oder an der Annahme liegt, und die Frage ist in einer eigenen Studie mit unabhängig validiertem Kompetenzmaß zu entscheiden.

Randomisiert wird auf Klassenebene, um Übertragungseffekte zu vermeiden. Verglichen werden bei gleicher Aufgabe und gleicher Zeit vier Bedingungen desselben Modells, die gestufte Hinweise (ja/nein) und eine Sperre des Endergebnisses (ja/nein) kreuzen, ergänzt um einen fünften Arm: Üben ohne KI. Erst diese Kreuzung trennt, was das Bündel bei Bastani et al. (2025) vermischt, nämlich den Beitrag der gestuften Hilfe vom Beitrag der Leitplanke.

Weil alle vier Zellen dieselbe hinterlegte Aufgabe und Rolle teilen, halten sie den dritten Hebel konstant; die Grundannahme in ihrer Fassung „stark gelenkt gegenüber offen“ verlangt deshalb einen sechsten Arm, in dem dasselbe Basismodell in derselben Oberfläche, aber ohne hinterlegte Aufgabe, ohne Fachkontext und ohne Rollenvorgabe antwortet, bei identischer Aufgabe und Zeit. Ein handelsübliches Produkt als Vergleichsarm wäre ungeeignet, weil es Kontextualisierung mit Modellgeneration, Oberfläche und Antwortstil vermengt und damit den Medienvergleichsfehler wiederholt, den Weidlich et al. (2025) beschreiben; wie realistisch eine so entkleidete Bedingung gegenüber der tatsächlichen Chatbot-Nutzung ist, begrenzt die externe Validität. Dieser Arm ist an die rechtlichen Voraussetzungen gebunden und im Grundschulalter nur über einen von der Lehrkraft geführten Zugang darstellbar; das vollständige Design ist deshalb primär in der Sekundarstufe zu realisieren. Der KI-freie Arm ist unverzichtbar, weil sich in bisherigen Feldversuchen die stark gelenkte Variante der Kontrollgruppe ohne KI nur anglich (Bastani et al., 2025); ohne ihn bliebe die Vorfrage des Modells ungeprüft.

Gemessen wird eine und vier Wochen später ohne KI, ergänzt um eine

Transferaufgabe, eine mündliche Erklärung des eigenen Wegs und die Streuung der Ergebnisse als eigene Zielgröße.

Die Fallzahl folgt aus einer Poweranalyse für diese Wechselwirkung zwischen Klassen- und Personenebene; maßgeblich sind neben der Zahl der Klassen die Zahl der Lernenden je Klasse und die Streuung des Kompetenzmaßes innerhalb der Klassen; für einen bedingten Effekt von d = 0,30 ist mit einer dreistelligen Zahl von Klassen zu rechnen. Lässt sich diese nicht erreichen, ist die Kreuzung auf den zentralen Kontrast (starke Lenkung gegenüber offen, plus KI-freier Arm) zu reduzieren und die Hebelzerlegung in einer eigenen Studie zu prüfen. Ein Manipulationscheck prüft an den Dialogprotokollen, ob die Leitplanken gehalten haben; Prozessdaten wie die Zeit bis zur ersten Hilfeanfrage und eigene Lösungsversuche davor dienen als vermittelnde Größen.

Zwei weitere Annahmen lassen sich prüfen. Die Verrechnungsannahme verlangt ein Design, das Werkzeuglenkung (eingebettet gegenüber offen) und Begleitung (intensiv gegenüber gering) kreuzt und für die beiden gemischten Zellen vergleichbare Ergebnisse vorhersagt. Das Verhältnis der Hebel verlangt, Kontextualisierung und Leitplanken faktoriell zu kreuzen; die Modellannahme lautet hier nicht Additivität, sondern Wechselwirkung: Leitplanken ohne Einbettung auf Anwendungsebene sollten schwächer wirken als dieselben Leitplanken mit Einbettung, weil sie im Dialog aushebelbar bleiben. Lässt sich die Zelle „Leitplanken ohne Kontextualisierung“ nicht herstellen, ist die Zerlegung in drei Hebel für die Praxis auf zwei zusammenzuführen. Ein solches Design wird im Projekt PRIMA-KI vorbereitet, das KI-gestützte Lernbegleitung im Mathematikunterricht der Grundschule erprobt.

Weitere Grenzen: Die Beispielbasis beruht überwiegend auf Mathematik, Programmieren und Schreiben; ob der Zusammenhang von Kompetenz und Lenkung in stärker interpretativen oder gestalterischen Fächern gleich verläuft, ist

offen. Die zitierten Befunde sind an bestimmte Modellgenerationen gebunden; verändern sich Fähigkeiten und Antwortstile generativer Modelle, können sich die Effekte verschieben. Und das Modell klammert Kriterien aus, die in der Praxis mitentscheiden: Datenschutz, Kosten, Verfügbarkeit und rechtliche Vorgaben.

Fazit

Wer entscheiden muss, ob und wie KI in einer Stunde vorkommt, geht die vier Planungsfragen durch und ordnet die Lernenden mit der Übersicht einer Stufe zu. Die Lenkung wird zurückgenommen, wenn die Lernenden zeigen, dass sie ohne sie weiterkommen, und nicht, wenn das Werkzeug es erlaubt. Ob das Modell hält, entscheidet die im Abschnitt zu den Grenzen skizzierte Prüfung.

Häufige Fragen

Was ist das SKILL-Modell?

Ein Ordnungsrahmen für den Einsatz von KI beim Lernen in Schule und Unterricht. Grundregel: Je geringer die Kompetenz einer Person, KI in einer konkreten Sache lernwirksam zu nutzen, desto mehr didaktische Lenkung braucht der Einsatz, aus dem Werkzeug oder durch die Lehrkraft. Mit wachsender Kompetenz wird die Lenkung zurückgenommen. Der Name ist kein Akronym, sondern leitet sich aus dieser Grundregel ab: Kompetenz („skill“) und Lenkung sind miteinander verwoben zu denken.

Sollte KI in der Grundschule eingesetzt werden?

Zurückhaltend und begrenzt. Das Modell empfiehlt keinen KI-Einsatz; es ordnet ihn, wenn er sinnvoll ist. In der Grundschule entstehen die Grundlagen, an denen sich später jede KI-Antwort messen muss; die Selbsteinschätzung des eigenen

Verstehens und das Prüfen fremder Aussagen sind noch in der Entwicklung. Wenn KI, dann eingebettet in eine Lernanwendung mit fester Rolle oder intensiv durch die Lehrkraft begleitet, und nur dort, wo sie etwas leistet, das Material, Gespräch und eigenes Probieren nicht leisten.

Warum kann ein Chatbot beim Lernen schaden?

Weil er den Lösungsprozess überspringen kann. In einem Feldexperiment mit knapp tausend Schülerinnen und Schülern stieg die Leistung beim Üben mit freiem Chatbot-Zugang; in der Prüfung ohne KI lag dieselbe Gruppe dann um 17 Prozent unter dem Leistungsniveau der Kontrollgruppe (Bastani et al., 2025). Leistung mit KI ist nicht dasselbe wie Lernen. Eine Variante mit didaktischen Leitplanken, die Hinweise statt Lösungen gibt, erreichte in der Prüfung wieder das Niveau der Gruppe ohne KI.

Reichen die Lernmodi von ChatGPT oder Gemini?

Sie verschieben das Verhalten in die gewünschte Richtung, garantieren die Grenzen aber nicht: Sprachmodelle geben Lösungen preis, wenn Lernende drängen oder das Thema wechseln (Zhao et al., 2026), und ein Lernmodus lässt sich abschalten. Für Stufe 2 reicht er, für Stufe 1 nicht; dort müssen die Grenzen außerhalb des Dialogs liegen, oder die Lehrkraft führt den Dialog selbst.

Was bedeutet didaktische Lenkung?

Alles, was den Handlungsspielraum der Lernenden im Sinne des Lernziels formt. Bei KI-Werkzeugen wirkt sie über drei Hebel: Scaffolding (gestufte Hilfe, die zurückgenommen wird), Leitplanken (was die KI nicht tut, etwa Lösungen vorwegnehmen) und Kontextualisierung (Aufgabe, Fachkontext und Rolle sind hinterlegt). Dazu kommt die personelle Begleitung durch die Lehrkraft.

Literatur

- Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education, 26*(1), 205–223. https://doi.org/10.1007/s40593-015-0089-1
- Armitage, K. L., & Gilbert, S. J. (2025). The nature and development of cognitive offloading in children. *Child Development Perspectives, 19*(2), 108–115. https://doi.org/10.1111/cdep.12532
- Barcaui, A. (2025). ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. *Social Sciences & Humanities Open, 12*, Artikel 102287. https://doi.org/10.1016/j.ssaho.2025.102287
- Bassner, P., Lenk-Ostendorf, B., Beinstingel, R., Wasner, T., & Krusche, S. (2026). Less stress, better scores, same learning: The dissociation of performance and learning in AI-supported programming education. *Computers and Education: Artificial Intelligence, 10*, Artikel 100537. https://doi.org/10.1016/j.caeai.2025.100537
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), e2422633122. https://doi.org/10.1073/pnas.2422633122
- Bayard, N. S., van Loon, M. H., Steiner, M., & Roebers, C. M. (2021). Developmental improvements and persisting difficulties in children’s metacognitive monitoring and control skills: Cross-sectional and longitudinal perspectives. *Child Development, 92*(3), 1118–1136. https://doi.org/10.1111/cdev.13486
- Belland, B. R., Walker, A. E., Kim, N. J., & Lefler, M. (2017). Synthesizing results from empirical research on computer-based scaffolding in STEM education: A meta-analysis. *Review of Educational Research, 87*(2), 309–344. https://doi.org/10.3102/0034654316670999
- Bellon, E., Fias, W., & De Smedt, B. (2020). Metacognition across domains: Is the association between arithmetic and metacognitive monitoring domain-specific? *PLOS ONE, 15*(3), Artikel e0229932. https://doi.org/10.1371/journal.pone.0229932
- Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development, 81*(6), 1641–1660. https://doi.org/10.1111/j.1467-8624.2010.01499.x
- Bildungsministerkonferenz. (2024). *Handlungsempfehlung für die Bildungsverwaltung zum Umgang mit Künstlicher Intelligenz in schulischen Bildungsprozessen* [Beschluss vom 10.10.2024]. https://www.kmk.org/

- Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Hrsg.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (S. 56–64). Worth Publishers.
- Brown, P., & Gummerum, M. (2025). Trust issues: Adolescents' epistemic vigilance towards online sources. *British Journal of Developmental Psychology, 43*(3), 578–594. https://doi.org/10.1111/bjdp.12559
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Artikel 188. https://doi.org/10.1145/3449287
- Buehler, F. J., Ghetti, S., & Roebers, C. M. (2025). Repeated feedback can benefit seven-year-olds' uncertainty monitoring in a memory task. *Journal of Cognitive Enhancement, 9*(2), 230–243. https://doi.org/10.1007/s41465-025-00322-8
- Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283–292. https://doi.org/10.3102/0013189X16656615
- Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Lawrence Erlbaum.
- Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington.
- Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005
- Danovitch, J. H. (2019). Growing up with Google: How children's understanding and use of internet-based devices relates to cognitive development. *Human Behavior and Emerging Technologies, 1*(2), 81–90. https://doi.org/10.1002/hbe2.142
- Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education, 210*, Artikel 104967. https://doi.org/10.1016/j.compedu.2023.104967
- de Boer, H., Donker, A. S., Kostons, D. D. N. M., & van der Werf, G. P. C. (2018). Long-term effects of metacognitive strategy instruction on student academic performance: A meta-

analysis. *Educational Research Review, 24*, 98–115. https://doi.org/10.1016/j.edurev.2018.03.002

- De Simone, M. E., Tiberti, F., Barron Rodriguez, M., Manolio, F., Mosuro, W., & Dikoru, E. J. (2025). *From chalkboards to chatbots: Evaluating the impact of generative AI on learning outcomes in Nigeria* (Policy Research Working Paper 11125). World Bank. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099548105192529324
- Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, Artikel 105224. https://doi.org/10.1016/j.compedu.2024.105224
- Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning, 10*(3), 347–374. https://doi.org/10.1007/s11409-014-9133-z
- Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students: A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning, 3*(3), 231–264. https://doi.org/10.1007/s11409-008-9029-x
- Doebel, S. (2020). Rethinking executive function and its development. *Perspectives on Psychological Science, 15*(4), 942–956. https://doi.org/10.1177/1745691620904771
- Europäische Union. (2024). Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (KI-Verordnung). *Amtsblatt der Europäischen Union*, L, 2024/1689. https://data.europa.eu/eli/reg/2024/1689/oj
- Europäische Union. (2026). Verordnung (EU) 2026/1744 des Europäischen Parlaments und des Rates zur Änderung der Verordnung (EU) 2024/1689 (Digital Omnibus zur künstlichen Intelligenz). *Amtsblatt der Europäischen Union*, L, 2026/1744. https://eur-lex.europa.eu/eli/reg/2026/1744/oj
- Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56*(2), 489–530. https://doi.org/10.1111/bjet.13544
- Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T., Shen, C., & Welsch, R. (2026). AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior, 175*, Artikel 108779. https://doi.org/10.1016/j.chb.2025.108779

- Fischer, M., Rau, H. A., & Rilke, R. M. (2025). *AI tutoring enhances student learning without crowding out reading effort* (IZA Discussion Paper Nr. 18338). Institute of Labor Economics. https://docs.iza.org/dp18338.pdf
- Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General, 144*(3), 674–687. https://doi.org/10.1037/xge0000070
- Fütterer, T., Bardach, L., Kuhn, J., Keller, S. D., & Gerjets, P. (2026). Enhancing school students' self-regulated learning through generative AI support: A randomized controlled trial. *Educational Psychology Review, 38*(1), Artikel 42. https://doi.org/10.1007/s10648-026-10133-8
- Girouard-Hallam, L. N., & Danovitch, J. H. (2022). Children's trust in and learning from voice assistants. *Developmental Psychology, 58*(4), 646–661. https://doi.org/10.1037/dev0001318
- Google DeepMind. (2026, 9. Juni). *Gemini's guided learning: Results from a randomized controlled trial in Sierra Leone* [Technischer Bericht des Anbieters]. https://deepmind.google/blog/measuring-the-impact-of-learning-with-ai-in-sierra-leone-and-beyond/
- Hamilton, E. R., Rosenberg, J. M., & Akcaoglu, M. (2016). The Substitution Augmentation Modification Redefinition (SAMR) model: A critical review and suggestions for its use. *TechTrends, 60*(5), 433–441. https://doi.org/10.1007/s11528-016-0091-y
- Han, X., Peng, H., & Liu, M. (2025). The impact of GenAI on learning outcomes: A systematic review and meta-analysis of experimental studies. *Educational Research Review, 48*, Artikel 100714. https://doi.org/10.1016/j.edurev.2025.100714
- Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology, 69*, 251–273. https://doi.org/10.1146/annurev-psych-122216-011710
- Herold, A., & Seufert, T. (2025). Who regulates? Effects of scaffolding in system- and self-regulated learning. *Frontiers in Psychology, 16*, Artikel 1543024. https://doi.org/10.3389/fpsyg.2025.1543024 (Korrigendum: https://doi.org/10.3389/fpsyg.2025.1642170)
- Heyes, C. (2016). Born pupils? Natural pedagogy and cultural pedagogy. *Perspectives on Psychological Science, 11*(2), 280–295. https://doi.org/10.1177/1745691615621276
- Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*(2), 99–107. https://doi.org/10.1080/00461520701263368

- Hofer, S. I., & Reinhold, F. (2025). Scaffolding of learning activities: Aptitude-treatment-interaction effects in math? *Learning and Instruction, 99*, Artikel 102177. https://doi.org/10.1016/j.learninstruc.2025.102177
- Holstein, K., McLaren, B. M., & Aleven, V. (2018). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren & B. du Boulay (Hrsg.), *Artificial intelligence in education* (Lecture Notes in Computer Science, Bd. 10947, S. 154–168). Springer. https://doi.org/10.1007/978-3-319-93843-1_12
- Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*(4), 509–539. https://doi.org/10.1007/s10648-007-9054-3
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4
- Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist, 51*(2), 289–299. https://doi.org/10.1080/00461520.2016.1155457
- Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15*(1), Artikel 17458. https://doi.org/10.1038/s41598-025-97652-6
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1
- Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review, 19*(3), 239–264. https://doi.org/10.1007/s10648-007-9049-0
- Kolloff, K., Roebers, C. M., & Buehler, F. J. (2023). 7- and 8-year-olds‘ struggle with monitoring: Inaccuracy persists despite feedback. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 55*(2–3), 91–109. https://doi.org/10.1026/0049-8637/a000276
- Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2506.08872
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher,*

49(4), 241–253. https://doi.org/10.3102/0013189X20912798

- Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42–78. https://doi.org/10.3102/0034654315581420
- Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research, 86*(3), 681–718. https://doi.org/10.3102/0034654315627366
- LearnLM Team, Google. (2024). *LearnLM: Improving Gemini for learning* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2412.16429
- Lee, H., Stinar, F., Zong, R., Valdiviejas, H., Wang, D., & Bosch, N. (2025). Learning behaviors mediate the effect of AI-powered support for metacognitive calibration on learning outcomes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (S. 1–18). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713960
- Lehmann, M., Cornelius, P. B., & Sting, F. J. (2025). *AI meets the classroom: When do large language models harm learning?* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.09047
- Li, X., & Yow, W. Q. (2024). Younger, not older, children trust an inaccurate human informant more than an inaccurate robot informant. *Child Development, 95*(3), 988–1000. https://doi.org/10.1111/cdev.14048
- Liu, J., Sun, D., Sun, J., Wang, J., & Yu, P. L. H. (2025). Designing a generative AI enabled learning environment for mathematics word problem solving in primary schools: Learning performance, attitudes and interaction. *Computers and Education: Artificial Intelligence, 9*, Artikel 100438. https://doi.org/10.1016/j.caeai.2025.100438
- Maurya, K. K., Srivatsa, K. A., Petukhova, K., & Kochmar, E. (2025). Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (Bd. 1, S. 1234–1251). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.57
- Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. J. A. (2020). Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *Journal of Experimental Psychology: General, 149*(9), 1788–1799. https://doi.org/10.1037/xge0000746
- Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. *Developmental Psychology, 49*(3), 404–418. https://doi.org/10.1037/a0029500
- Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one’s understanding: The development of

an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology, 87*(1), 1–32. https://doi.org/10.1016/j.jecp.2003.09.003

- Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x
- Molenaar, I. (2022a). The concept of hybrid human-AI regulation: Exemplifying how to support young learners‘ self-regulated learning. *Computers and Education: Artificial Intelligence, 3*, Artikel 100070. https://doi.org/10.1016/j.caeai.2022.100070
- Molenaar, I. (2022b). Towards hybrid human-AI learning technologies. *European Journal of Education, 57*(4), 632–645. https://doi.org/10.1111/ejed.12527
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Hrsg.), *Psychology of learning and motivation* (Bd. 26, S. 125–173). Academic Press. https://doi.org/10.1016/S0079-7421(08)60053-5
- Newman, R. S. (2002). How self-regulated learners cope with academic difficulty: The role of adaptive help seeking. *Theory Into Practice, 41*(2), 132–138. https://doi.org/10.1207/s15430421tip4102_10
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, Artikel 100041. https://doi.org/10.1016/j.caeai.2021.100041
- OECD. (2026). *OECD digital education outlook 2026: Exploring effective uses of generative AI in education*. OECD Publishing. https://doi.org/10.1787/062a7394-en
- OECD & Europäische Kommission. (2026). *Empowering learners for the age of AI: An AI literacy framework for primary and secondary education*. OECD Publishing. https://doi.org/10.1787/65cd27d4-en
- Otis, N. G., Clarke, R., Delecourt, S., Holtz, D., & Koning, R. (2026). The uneven impact of generative artificial intelligence on entrepreneurial performance: Evidence from a field experiment in Kenya. *Management Science*. Vorab-Onlineveröffentlichung. https://doi.org/10.1287/mnsc.2024.06909
- Pearson, P. D., & Gallagher, M. C. (1983). The instruction of reading comprehension. *Contemporary Educational Psychology, 8*(3), 317–344. https://doi.org/10.1016/0361-476X(83)90019-X
- Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment.

Journal of University Teaching and Learning Practice, 21(6). https://doi.org/10.53761/q3azde36

- Perkins, M., Roe, J., & Furze, L. (2025). Reimagining the Artificial Intelligence Assessment Scale: A refined framework for educational assessment. *Journal of University Teaching and Learning Practice, 22*(7). https://doi.org/10.53761/rrm4y757
- Puntambekar, S., & Hübscher, R. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist, 40*(1), 1–12. https://doi.org/10.1207/s15326985ep4001_1
- Redecker, C. (2017). *European framework for the digital competence of educators: DigCompEdu* (Y. Punie, Hrsg.). Publications Office of the European Union. https://doi.org/10.2760/159770
- Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences, 13*(3), 273–304. https://doi.org/10.1207/s15327809jls1303_2
- Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist, 38*(1), 15–22. https://doi.org/10.1207/S15326985EP3801_3
- Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002
- Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review, 45*, 31–51. https://doi.org/10.1016/j.dr.2017.04.001
- Roebers, C. M., van Loon, M. H., Buehler, F. J., Bayard, N. S., Steiner, M., & Aeschlimann, E. A. (2021). Exploring psychometric properties of children’s metacognitive monitoring. *Acta Psychologica, 220*, Artikel 103399. https://doi.org/10.1016/j.actpsy.2021.103399
- Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience, 1*, Artikel e17. https://doi.org/10.1017/pen.2018.16
- Rovers, S. F. E., Clarebout, G., Savelberg, H. H. C. M., de Bruin, A. B. H., & van Merriënboer, J. J. G. (2019). Granularity matters: Comparing different ways of measuring self-regulated learning. *Metacognition and Learning, 14*(1), 1–19. https://doi.org/10.1007/s11409-019-09188-6
- Rowe, L. (2026). From gradual release of responsibility to gradual release of technology: A case for the staged use of AI in formal education. *Education Sciences, 16*(8), Artikel 1291. https://doi.org/10.3390/educsci16081291

- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1
- Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher, 20*(3), 2–9. https://doi.org/10.3102/0013189X020003002
- Saye, J. W., & Brush, T. (2002). Scaffolding critical reasoning about history and social issues in multimedia-supported learning environments. *Educational Technology Research and Development, 50*(3), 77–96. https://doi.org/10.1007/BF02505026
- Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. https://openreview.net/forum?id=tvhaxkMKAn
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x
- Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior, 160*, Artikel 108386. https://doi.org/10.1016/j.chb.2024.108386
- Ständige Wissenschaftliche Kommission der Kultusministerkonferenz [SWK]. (2024). *Large Language Models und ihre Potenziale im Bildungssystem* [Impulspapier]. https://doi.org/10.25656/01:28303
- Stanković, M., Hirche, E., Kollatzsch, S., & Doetsch, J. N. (2026). *Comment on: Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing tasks* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2601.00856
- Strohmaier, A., Bödefeld, S., Straser, O., & Reinhold, F. (2026). *LLAMA LIMA: A living meta-analysis on the effects of generative AI on learning mathematics* (Version 4) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2601.18685
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292.

https://doi.org/10.1007/s10648-019-09465-5

- Szufnarowska, J., Rohlfing, K. J., Fawcett, C., & Gredebäck, G. (2014). Is ostension any more than attention? *Scientific Reports, 4*, Artikel 5304. https://doi.org/10.1038/srep05304
- Tabak, I. (2004). Synergy: A complement to emerging patterns of distributed scaffolding. *Journal of the Learning Sciences, 13*(3), 305–335. https://doi.org/10.1207/s15327809jls1303_3
- Tetzlaff, L., Simonsmeier, B. A., Peters, T., & Brod, G. (2025). A cornerstone of adaptivity: A meta-analysis of the expertise reversal effect. *Learning and Instruction, 98*, Artikel 102142. https://doi.org/10.1016/j.learninstruc.2025.102142
- UNESCO. (2024). *AI competency framework for students*. https://doi.org/10.54675/JKJB9835
- Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6
- van Loon, M. H., & Roebers, C. M. (2024). Development of metacognitive monitoring and control skills in elementary school: A latent profile approach. *Metacognition and Learning, 19*(3), 1065–1089. https://doi.org/10.1007/s11409-024-09400-2
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner & E. Souberman, Hrsg.). Harvard University Press.
- Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2025). *Tutor CoPilot: A human-AI approach for scaling real-time expertise* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2410.03017
- Weidlich, J., Gašević, D., Drachsler, H., & Kirschner, P. A. (2025). ChatGPT in education: An effect in search of a cause. *Journal of Computer Assisted Learning, 41*(5), Artikel e70105. https://doi.org/10.1111/jcal.70105
- Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., Dolan, R. J., & Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition, 22*(1), 264–271. https://doi.org/10.1016/j.concog.2013.01.004
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

- Yan, L., Greiff, S., Lodge, J. M., & Gašević, D. (2025). Distinguishing performance gains from learning when using generative AI. *Nature Reviews Psychology, 4*(7), 435–436. https://doi.org/10.1038/s44159-025-00467-5
- Zelazo, P. D., & Carlson, S. M. (2012). Hot and cool executive function in childhood and adolescence: Development and plasticity. *Child Development Perspectives, 6*(4), 354–360. https://doi.org/10.1111/j.1750-8606.2012.00246.x
- Zhao, J., Knežević, M., & Käser, T. (2026). Evaluating answer leakage robustness of LLM tutors against adversarial student attacks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics* (S. 30588–30617). Association for Computational Linguistics. https://aclanthology.org/2026.acl-long.1412/
- Zhou, W., Wang, R., Xiong, Y., Zhang, X., Liu, X., Lin, C., Zheng, H., Yin, L., Zheng, D., Xie, B., & Shu, Q. (2026). Chatbot versus preservice teacher-led reading in primary schools: The roles of engagement and learner differences. *Reading and Writing*. Vorab-Onlineveröffentlichung. https://doi.org/10.1007/s11145-026-10775-8
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2