1.3 Verantwortungsbewusster Einsatz von KI

Kernbotschaften

- **Verantwortung ist eine Gestaltungsaufgabe, keine Kontrollaufgabe.** Die Erkennung von KI funktioniert nicht zuverlässig.
- **Und KI-Erkennung ist unfair, wenn sie versagt.** Detektoren markieren menschliche Arbeiten fälschlich als KI-generiert und übersehen KI-generierte Arbeiten. Bildungsbehörden in Frankreich und den Niederlanden raten von ihrem Einsatz ab.
- **Selbstsichere Erfindung und ererbte Verzerrungen ergeben sich daraus, wie diese Systeme gebaut und bewertet werden.** Die Konsequenz sollte sein: sie zu lehren, statt auf eine Reparatur zu warten.
- **Das Wirkprinzip lautet: erst denken, dann KI.** Menschen dazu zu bringen, sich zunächst auf eine eigene Antwort festzulegen, schlägt das Erklären der Maschine, und in die Software eingebaute Grenzen beseitigen den Schaden, den uneingeschränkter Zugang anrichtet.
- **Prompts allein halten nicht.** Ein Chatbot, der als Tutor agieren soll, neigt dazu, wieder in die Antwortvergabe zurückzufallen, was verlässliche Grenzen zu einer Beschaffungsfrage macht.
- **Verstärkung ist der Normalfall; Ausgleich muss aus Zugang, Gestaltung und Lernbegleitung gemeinsam aufgebaut werden.** Das Knappste, aber Wichtigste der drei ist eine geschulte Lehrkraft.
- **Inklusion ist der klarste Gewinn**, und dort sehen Lehrkräfte selbst das stärkste Argument.

> Wenn Sie sich nur einen Satz von dieser Seite merken: **Die Dämme brechen. Bringen Sie das Schwimmen bei.**

Das Lesetagebuch, zweiter Blick

Zurück zu unserem Neuntklässler und seinem verdächtig druckreifen Lesetagebuch. Der Instinkt ist fast universell, und es ist der Instinkt einer guten Lehrkraft: *herausfinden, ob er KI benutzt hat.* Einen besseren Damm bauen.

Dieses Kapitel handelt davon, warum dieser Instinkt für sich genommen nicht weiterführt. Wir kommen dazu, was stattdessen funktioniert.

Beginnen wir mit dem unbequemen Teil. **Lehrkräfte können KI-Text nicht zuverlässig von Schülertext unterscheiden.** In zwei Studien eines deutschen Teams erhielten 89 Lehramtsanwärter und 200 erfahrene Lehrkräfte eine Mischung aus Schüleraufsätzen und Chatbot-Aufsätzen und sollten diese sortieren. Keine der beiden Gruppen konnte das zuverlässig, aber beide Gruppen waren zuversichtlich, dass sie es könnten. Erfahrung half nur wenig: Auch die erfahrenen Lehrkräfte übersahen die meisten KI-Texte. An einer britischen Universität reichten Forschende von Chatbots erzeugte Antworten über die üblichen Kanäle bei echten, nicht beaufsichtigten Online-Prüfungen ein, ohne die Korrigierenden zu informieren.

„Sie dürfen sie nutzen, aber Sie müssen das Ergebnis in eigenen Worten erklären können."

Der Detektor, der die Falschen im Stich lässt

Forschende testeten sieben der am weitesten verbreiteten KI-Detektoren an TOEFL-Aufsätzen von Personen, die Englisch als Fremdsprache lernen; die Aufsätze stammten aus einem chinesischen Online-Forum. Daneben legten sie Aufsätze amerikanischer Achtklässler mit Englisch als Muttersprache. Die Arbeiten der Muttersprachler wurden fast immer korrekt eingestuft. Von den fremdsprachigen Aufsätzen wurden mehr als sechs von zehn fälschlich als maschinell geschrieben markiert.

Der Mechanismus ist kein Geheimnis. Viele dieser Werkzeuge erkennen nicht die Arbeit von Large Language Models. Die Werkzeuge messen *Vorhersehbarkeit*. Text, der gängige Wörter in erwarteter Reihenfolge verwendet, wird als künstlich eingestuft. Wer in einer Zweitsprache schreibt, mit kleinerem Wortschatz und starreren Satzstrukturen, kann genau dieses Signal erzeugen. Detektoren, die für einen bestimmten Test trainiert wurden, sowie Studien in anderen Sprachen zeigen allerdings geringere Unterschiede. Bislang gibt es auf dem Markt keine Detektoren, die zuverlässig genug sind, um jemandem eine vorsätzliche Täuschung vorzuwerfen.

Die Fehler des Werkzeugs treffen systematisch die Schülerinnen und Schüler, die ohnehin schon am meisten zu tragen haben. Institutionen haben nachgerechnet. Die Vanderbilt University wies darauf hin, dass selbst wenn ein solches Werkzeug nur in 1 % der Fälle irren würde, das bei 75.000 eingereichten Arbeiten immer noch rund **750 fälschlich markierte Arbeiten pro Jahr** bedeuten würde. Folglich schaltete sie es ab. Ein von Forschenden für das deutsche Bundesbildungsministerium verfasster Leitfaden ist unmissverständlich: Die Technologie sei *„nicht zuverlässig genug, um einen rechtssicheren Nachweis zu liefern"*. Ein unabhängiger europäischer Test von vierzehn Detektoren kam zum selben Urteil: *„weder genau noch*

zuverlässig“.

Ein Großteil Europas hat diese Debatte hinter sich gelassen. Das französische Bildungsministerium und die französische Datenschutzbehörde raten Schulen davon ab, Detektoren einzusetzen; die niederländische IKT-Agentur Kennisnet hält sie für höchst unzuverlässig. Im Hochschulbereich verbieten einzelne Hochschulen, etwa die ZHAW in Zürich, ihren Einsatz für die Bewertung. In England erlauben die Prüfungsgremien Erkennungssoftware nur als ein Element eines Urteils, das bei der Lehrkraft liegt, die *„ihre Schülerinnen und Schüler am besten kennt“*. Auch ein Datenschutzaspekt lohnt sich zu merken: **Die Arbeit einer Schülerin oder eines Schülers in ein Online-Werkzeug einzufügen, bedeutet, personenbezogene Daten zu verarbeiten.** Dafür sind eine Rechtsgrundlage und ein Vertrag mit dem Anbieter erforderlich. → 1.4 Konformer Einsatz von KI

Es scheint, dass die Betrugspanik dünner ist als der Lärm darum. Forschende befragten Schülerinnen und Schüler an drei US-amerikanischen High Schools vor der Einführung von ChatGPT und erneut danach. Das Ausmaß des Betrugs blieb weitgehend gleich. Eine Folgebefragung von mehr als 4.000 Schülerinnen und Schülern an sechs weiteren US-amerikanischen High Schools ein Jahr später kam zum selben Ergebnis.

Die eigentliche Frage war nie: *„Wie erwische ich ihn?*" Sie lautet: **Was misst eine Hausaufgabe eigentlich noch, sobald sich ihr Produkt auf Abruf erzeugen lässt?**

„Zuerst habe ich das als Betrug gesehen. Heute sehe ich es als Signal, dass sich die Aufgabe selbst ändern muss.“

Zwei Dinge, die nicht behoben werden, also müssen sie gelehrt werden

Selbstsichere Erfindung ist eingebaut. Diese Systeme erzeugen plausible Fortsetzungen eines Textes, und plausibel ist nicht dasselbe wie wahr. Übersichtsarbeiten zur Forschung sind sich einig, dass bislang keine Technik das Problem beseitigt hat, auch wenn es sich verringern und teilweise erkennen lässt. Der Grund: Die Standardtests zur Bewertung von Large Language Models belohnen selbstsicheres Raten gegenüber dem Eingeständnis von Unsicherheit, und solange so gemessen wird, bleibt eine gewisse Rate an flüssigen, wohlgeformten Falschaussagen bestehen.

Auch die Schieflagen im Trainingsmaterial sind eingebaut. Zudem schrumpfen sie nicht, wenn sich die Modelle verbessern. Eine europäische Studie zu Geschlechterstereotypen über verschiedene Sprachen hinweg fand, dass größere Modelle – selbst nach zusätzlichem Training, das sie sicherer und fairer machen soll – manchmal *stärker* statt schwächer stereotypisieren. Wenn sich eine Eigenschaft nicht wegentwickeln lässt, wird sie zum Lehrplan. Und sie lässt sich lehren: In einem finnischen Programm stieg der Anteil der Schüler:innen, die erklären konnten, woher die Verzerrung eines Bildgenerators stammt, von **7 % auf 44 %**. Die Konsequenz für den Unterricht muss sein, Fehleranalyse statt Fehlervermeidung zu trainieren. Mit anderen Worten: Die Fehler der

Maschine werden zur Lektion.

„Lassen Sie die Lernenden die Lösung des Chatbots prüfen, als wären sie die Lehrkraft. Gerade durch die Suche nach den Fehlern lernen sie schließlich die Methode."

Was funktioniert: Gestaltung statt Kontrolle

Die Evidenz weist auf ein einfaches, leicht zu merkendes Prinzip hin: **erst denken, dann KI.**

- **Den Menschen zuerst festlegen lassen.** In einem Experiment mit rund 200 Erwachsenen folgten Personen, die ihre eigene Entscheidung treffen mussten, *bevor* sie den Vorschlag der Maschine sahen, deren falschen Vorschlägen seltener als Personen, denen eine kurze Erklärung der maschinellen Argumentation gezeigt wurde. Die KI zu erklären löste das Problem nicht. Geholfen hat das Erzwingen des eigenen Urteils. Das eigene Urteil zu erzwingen heißt, Menschen zum Nachdenken zu zwingen. Da Nachdenken anstrengend ist, gefielen diese Methoden den Teilnehmenden am wenigsten, und sie halfen am meisten jenen, die ohnehin gerne anstrengend nachdenken. Das ist das, was Sie alle als Matthäus-Effekt kennen: Wer schon hat, dem wird gegeben.
- **Die Grenzen einbauen.** Der Chatbot aus Kapitel 1.2, der Antworten zurückhielt und stattdessen Rückfragen stellte, ist dieselbe Idee als Software: Die Schüler:innen, die ihn nutzten, verloren in der Prüfung nichts, während die mit der uneingeschränkten Version deutlich verloren.
- **Den Prozess sichtbar machen, statt dem Produkt hinterherzujagen.** Entwürfe, Arbeitsnotizen, ein kurzes Gespräch über die Arbeit. Diese zeigen Ihnen, was ein Schüler oder eine Schülerin kann, ohne dass jemand beweisen muss, was er oder sie getan hat.
- **Dänemark erprobt das Prinzip:** An teilnehmenden Schulen dürfen Schüler:innen KI bei der Vorbereitung auf die mündliche Englischprüfung nutzen, und die schriftliche Prüfung enthält einen handschriftlichen, hilfsmittelfreien Teil.

Zwei Einschränkungen sollte man kennen, bevor man sich darauf verlässt.

Die erste: Der beliebte Rat, Lehrkräfte sollten einen Chatbot einfach anweisen, als sokratischer Tutor zu agieren, übersteht ein längeres Gespräch möglicherweise nicht. Erste Hinweise aus einer noch nicht begutachteten Studie mit simulierten Schülerinnen und Schülern (Oktober 2026) deuten darauf hin, dass solche Tutoren unter anhaltendem Druck wieder dazu übergehen, Antworten zu liefern. Verlässliche Grenzen müssen in das System eingebaut sein, was sie zu einer Beschaffungsentscheidung macht statt zu einer Prompting-Fähigkeit.

Die zweite: Eine KI ihre Argumentation erklären zu lassen, klingt wie die naheliegende Absicherung, aber Erklärungen verhindern übermäßiges Vertrauen nicht zuverlässig. In einer Studie ließen sie die Akzeptanz falscher Ratschläge unverändert, und zu technische oder zu vereinfachte Erklärungen können fehlgeleitetes Vertrauen sogar verstärken. Erklärung ist nur eine

Voraussetzung für Lernen.

Haltung: zwei Fehlermodi, nicht einer

Wir sorgen uns, dass Lehrkräfte der KI zu sehr vertrauen. Wir sollten uns jedoch ebenso um das Gegenteil sorgen. Selbst erfahrene Menschen können durch übermäßiges Vertrauen hereingelegt werden. In einem kleinen Pilotprojekt an US-amerikanischen Schulen beurteilten vier von zehn befragten Lehrkräften das Feedback eines KI-Bewertungswerkzeugs als nicht hilfreich, überwiegend vage oder sogar falsch. Die Forschung zum Automation Bias zeigt, dass Menschen dazu neigen, dem Vorschlag einer Maschine zu folgen, selbst wenn sie Grund haben, daran zu zweifeln. Einen Fehler zu erkennen und ihn zu korrigieren sind zwei verschiedene Schritte.

Der gegenteilige Pol ist Verheimlichung. Lehrkräfte verbergen ihre eigene KI-Nutzung vor Kolleg:innen und Schüler:innen und verbieten diesen, was sie selbst tun: 77 % nutzen KI privat, während 57 % sagen, Schüler:innen sollten sie nie zur Ideenfindung nutzen. Zugleich fand eine Studie unter akademisch starken Schüler:innen der Klassen 6 bis 8, dass zwischen einem Drittel und zwei Fünfteln bereits sagten, die KI wisse mehr als ihre Lehrkräfte. **Ein Berufsstand, der seine Praktiken verbirgt, kann verantwortungsvolle Praxis nicht vorleben**, und Transparenz ist nicht nur eine Höflichkeit. Sie ist der Mechanismus, durch den Lernen am Vorbild funktioniert.

Ebnet KI das Spielfeld oder kippt sie es?

Bei arbeitsbezogenen Aufgaben gleicht KI Unterschiede aus. In einem Experiment schnitten Teilnehmende mit höherer Bildung deutlich besser ab als solche mit niedrigerer Bildung. Mit einem KI-Assistenten verschwanden etwa drei Viertel dieser Lücke. Als der Assistent wieder entzogen wurde, kehrte ein beträchtlicher Teil der Lücke zurück. Was die KI schloss, war die Lücke im Output, nicht die Lücke in der Fähigkeit.

Bei Lernaufgaben ist es umgekehrt. Eine Studie fand im Schnitt gar keinen Nutzen, aber einen wachsenden Abstand zwischen Schüler:innen, die bereits viel wussten, und solchen, die es nicht taten. Eine andere untersuchte Schüler:innen, die in einer Fremdsprache schrieben: Die schwächeren Schreiber:innen erhielten *mehr* Korrekturen von der KI und setzten *weniger* davon erfolgreich um, und mehrere empfanden die Flut an Feedback als entmutigend. **Feedback ohne die Fähigkeit, es zu sortieren, hilft nicht; es begräbt.**

Die Lücke besteht auch schon, bevor überhaupt jemand ein Werkzeug anfasst. In Deutschland **sehen 80 % der Schüler:innen aus den privilegiertesten Elternhäusern KI als Chance, gegenüber 55 % aus den am wenigsten privilegierten**. In Großbritannien verläuft die Trennlinie eher durchs Lehrerzimmer als durch den Geräteschrank: **45 % der Lehrkräfte an Privatschulen haben eine formale KI-Fortbildung erhalten, gegenüber 21 % an staatlichen Schulen.** Innerhalb des staatlichen Sektors liegen wohlhabendere Schulen wiederum vor ärmeren. Eine Befragung von mehr als 25.000 Kindern in 17 europäischen Ländern aus dem Jahr 2026 findet dieselbe soziale Kluft darin, wie viel und wie vielfältig sie KI nutzen. Ein Satz fasst

das ganze Problem besser zusammen als jede Zahl: *„Die Reichen haben Zugang zu Technologie und zu Menschen, die ihnen bei der Nutzung helfen, während die Armen nur Zugang zur Technologie haben.*" Das ist das stärkste verfügbare Argument dafür, dass Lehrkräfte entscheidend sind, um diese Lücke zu schließen.

Der Engpass ist also, welche Lernenden eine geschulte Lehrkraft haben. **KI gleicht aus, wenn drei Dinge zusammenkommen: sicherer Zugang, ein für das Lernen (statt für das Liefern von Antworten) gestaltetes Werkzeug, und eine Lehrkraft, die die Nutzung anleitet.** Fehlt eines davon, verstärkt sie stattdessen. Da alle drei tendenziell an denselben Schulen vorhanden bzw. an denselben Schulen abwesend sind, **ist Verstärkung der Normalfall, und Ausgleich ist etwas, das man aktiv aufbauen muss.**

Revision #14
Created 2026-09-01 09:16:52 UTC by Alexander
Updated 2026-10-20 15:02:43 UTC by Alexander