schule verantworten
führungskultur_innovation_autonomie

Thomas Benesch
Private Pädagogische Hochschule Burgenland, Eisenstadt

Das AI-Sandwich als Führungsinstrument für kreative Schulentwicklung

DOI: https://doi.org/10.53349/schuleverantworten.2026.i3.a713

Generative KI-Systeme verändern die Bedingungen schulischer Führung, etwa durch Entscheidungsvorlagen für Bewertungen. Das betrifft Autonomie und Quelle der Anforderungen und wirkt sich direkt auf die motivationale Orientierung und Kreativität der Führungskraft aus. Das AI-Sandwich-Modell (Freeman, 2026) bietet einen Rahmen, um KI kreativ und zugleich verantwortungsvoll einzusetzen. Menschliche Aufsicht reicht nicht aus, um algorithmische Fehler oder Verzerrungen zu korrigieren, wenn kognitive Ankereffekte nicht aktiv adressiert werden. Das AI-Sandwich kann genau hier ansetzen, indem es reflexive Barrieren institutionalisiert. Kreative Führung wird dabei nicht als originelles Handeln der Führungskraft verstanden, sondern als Praxis, die gezielt die Rahmenbedingungen für eigenständige, intrinsisch motivierte Kreativität der Mitarbeitenden schafft. Das AI-Sandwich unterstützt diese Praxis, indem es strategische Autonomie (Human-First), dialogische Ko-Kreativität (AI-Middle) und unhintergehbare menschliche Letztverantwortung (Human-Last) miteinander verbindet. Der Beitrag entfaltet das Modell theoretisch unter Rückgriff auf die Selbstbestimmungstheorie (Ryan & Deci, 2000), die Forschung zur Ambiguitätstoleranz (Budner ,1962, S. 29) und verantwortungsethische Überlegungen (Jonas, 1979). Ziel ist die Bereitstellung eines kategorialen Instruments für Schulleitungen, die KI als Resonanzraum kreativer Führung begreifen.

Künstliche Intelligenz, Kreativität, Schulleitung, Verantwortung, Führungsaufgaben

„Soziale und umgebungsbezogene Faktoren sind zentral für kreative Leistung."
T. M. Amabile, 1996

Einleitung

Generative KI (ChatGPT, Gemini, Claude etc.) dringt zunehmend in schulische Führungsprozesse ein – von der Entscheidungsvorbereitung über die Kommunikation bis zur strategischen Planung. Diese Entwicklung birgt sowohl Chancen als auch Risiken. Die zentrale Gefahr besteht in der algorithmischen Herrschaft (Danaher, 2016, S. 245): KI-Systeme übernehmen un-

merklich Entscheidungen, während menschliche Führungskräfte nur noch diese ausführen. Matthias (2004, S. 177) präzisiert diese Gefahr als Verantwortungslücke. Es entsteht eine wachsende Klasse von Maschinenhandlungen, bei denen die traditionelle Zuschreibung nicht mehr mit unserem Gerechtigkeitsempfinden vereinbar ist, weil niemand mehr ausreichend Kontrolle besitzt – weder die Maschine noch der Mensch. Während algorithmenbasierte Systeme für reibungslose Konnektivität sorgen, steht dem eine subjektgebundene, nur kommunikativ funktionierende Realität gegenüber. Sweller (1988, S. 259) zeigt, dass Expert*innen Probleme nicht durch algorithmische Suche, sondern durch Schemaerkennung lösen – sie erkennen Problemzustände und wissen, welche Schritte angemessen sind. Die kommunikative, reflexive Auseinandersetzung mit einem Problem ist genau das, was bei automatisierter KI-Nutzung verloren geht. O'Neil (2016) zeigt, dass Algorithmen ihre eigene Realität definieren, statt tatsächliche Komplexität der Welt abzubilden – sie ersetzen die Wahrheit durch ihren eigenen Score.

Danaher (2016, S. 254) stellt fest, dass die zunehmende Abhängigkeit von algokratischen Systemen den Raum für aktive menschliche Teilhabe an und das Verständnis von Entscheidungsprozessen einschränkt – was er als Bedrohung für legitime Verfahren ansieht. Die algorithmische Herrschaft droht, diese kommunikative Praxis zu untergraben (Betschart 2025, S. 55). Danaher (2016, S. 252) warnt davor, dass wir durch die Bevorzugung algokratischer Systeme (wegen ihrer instrumentellen Vorteile wie Geschwindigkeit und Genauigkeit) letztlich Systeme schaffen, die zunehmend undurchschaubar werden und die kommunikative, partizipative Basis demokratischer Legitimität untergraben. Demgegenüber steht das Ideal von Führungskräften, die durch geeignete Managementpraktiken Bedingungen für kreatives Arbeiten schaffen – und die bei ihren Mitarbeiter*innen kreativitätsförderliche Persönlichkeitsmerkmale wie Toleranz für Ambiguität voraussetzen (Amabile, 1997, S. 54).

Das AI-Sandwich-Modell (Freeman, 2026) verspricht einen Ausweg aus dem Dilemma zwischen naiver KI-Akzeptanz und pauschaler Ablehnung. Ursprünglich für Lehr-Lernkontexte entwickelt, überträgt der Beitrag diese Logik auf Führungsaufgaben: Sie dient dann nicht als starres Regelwerk, sondern als einprägsame Denkfigur, um KI-gestützte Entscheidungen reflexiv abzusichern. Der Beitrag entfaltet dieses Modell in drei Schritten: Abschnitt 2 klärt den Begriff kreativer Führung, Abschnitt 3 präsentiert die Sandwich-Struktur einschließlich der theoretischen Vertiefung jeder Schicht sowie dem Transfer in den Schulalltag. Der Abschnitt 4 diskutiert die Grenzen des Modells, das Fazit schließt den Beitrag zusammenfassend ab.

Kreative Führung: begriffliche Grundlagen

Kreative Führung zielt im hier verstandenen Sinne nicht auf originelles Handeln der Führungskraft selbst, sondern auf die Ermöglichung eigenständiger, intrinsisch motivierter Kreativität der Mitarbeitenden. Mainemelis, Kark und Epitropaki (2015) fassen diesen Ansatz als ein Führen auf, das andere zur Erreichung eines kreativen Ergebnisses anleitet, und heben die Förderung von Mitarbeiterkreativität als zentrale Dimension hervor. Pointiert formulieren Amabile

und Khaire (2008): Kreativität lässt sich nicht verwalten – nur ihre Rahmenbedingungen sind gestaltbar. In diesem Verständnis unterscheidet sich kreative Führung von transaktionaler Verwaltungsführung durch ihren Fokus auf intellektuelle Stimulation und Inspiration. Amabile (1996, S. 15) betont, dass kreative Führung vor allem dann gelingt, wenn die Führungskraft eine Umgebung schafft, die intrinsische Motivation fördert – also das Arbeiten aus eigenem Interesse heraus. Diese intrinsische Motivation wird jedoch durch die Unvermeidbarkeit von Fehlern bei lernenden Systemen auf eine harte Probe gestellt. Matthias (2004, S. 179) zeigt, dass Fehler bei Systemen mit bestärkendem Lernen keine Produktmängel, sondern notwendige Systemeigenschaften sind – die Führungskraft muss also darauf hinwirken, eine Fehlerkultur zu entwickeln, die diese Fehler nicht als Versagen, sondern als Lernchance begreift. Kultur lässt sich jedoch nicht verordnen oder administrativ etablieren – sie entsteht als geteiltes, implizites Muster von Annahmen und Werten in sozialen Praktiken (Schein, 2010). Die Quelle kreativer Führung liegt in der Ermöglichung von Selbstbestimmung und Neugier.

Amabile (1997, S. 40) definiert Kreativität als Produktion neuartiger und angemessener Ideen – übertragen auf die Gestaltung von Arbeitsumgebungen bedeutet dies, dass Führungskräfte Bedingungen schaffen sollten, unter denen solche Ideen entstehen können Für Schulen gewinnt diese Aufgabe eine spezifische Prägung: Pädagogische Führung zielt nicht auf Verwaltungslogik, sondern auf die unmittelbare Förderung von Lehr- und Lernprozessen ab (Dubs, 2019). Entscheidungen müssen sich daher letztlich an der Frage messen lassen, ob sie Lernprozesse unterstützen. Generative KI verstärkt diesen normativen Anspruch, da sie ethische Urteilsfähigkeit – verstanden als die Fähigkeit zur reflexiven Abwägung von Handlungsfolgen im pädagogischen Kontext – nicht ersetzt, sondern deren aktive Ausübung umso dringlicher macht.

Eine Schlüsselkompetenz kreativer Führung ist Ambiguitätstoleranz – die Fähigkeit, mit mehrdeutigen, unstrukturierten Situationen umzugehen, ohne vorschnelle Schlüsse (Budner, 1962, S. 30). Diese Toleranz ermöglicht den Zugang zu abstrakterem Wissen, was nachweislich zu originelleren Ergebnissen führt als die Fixierung auf konkrete, sofort verfügbare Lösungen (Ward & Kolomyts, 2019, S. 179). Erfolgreiche Schulleitungen sind aufgeschlossen, lernbereit, flexibel, beharrlich, resilient und optimistisch (Leithwood, Harris & Hopkins 2008, S. 36). Diese Eigenschaften sind auch für den souveränen Umgang mit KI unerlässlich.

Amabile (1996, S. 11) verweist darauf, dass besonders kreative Personen gelernt haben, mit externen Erwartungs- und Bewertungsdruck umzugehen, ohne sich davon lähmen zu lassen. Ambiguitätstoleranz ist in ihrem Modell eine wichtige kreativitätsrelevante Persönlichkeitseigenschaft. Wer Mehrdeutigkeit aushält, kann sich besser von extrinsischen Zwängen (zum Beispiel KI-generierten eindeutigen Antworten) distanzieren und eigene originelle Wege gehen. Während Personen mit niedriger Ambiguitätstoleranz dazu neigen, vorschnelle Schwarz-Weiß-Urteile zu fällen, werden mehrdeutige Situationen von toleranteren Personen als wünschenswert, herausfordernd und interessant wahrgenommen – genau diese Haltung eröffnet den Raum für originelles Denken (Furnham & Marks, 2013, S. 717).

Budner (1962, S. 30) spezifiziert drei Typen mehrdeutiger Situationen: Neuheit (völlig neue Situation ohne vertraute Hinweise), Komplexität (eine Vielzahl gleichzeitig zu berücksichtigender Hinweise) und Unlösbarkeit (widersprüchliche Hinweise, die unterschiedliche Strukturen nahelegen). Alle drei treten in der schulischen Führungspraxis mit KI auf. Neue KI-Tools, komplexe Datenlagen oder widersprüchliche Empfehlungen der KI erzeugen systematisch Ambiguität: sie liefern Wahrscheinlichkeiten, halluzinieren, widersprechen sich. Führungskräfte benötigen daher Ambiguitätstoleranz, um KI souverän zu nutzen. Allerdings setzt dies voraus, dass Führungskräfte die Mehrdeutigkeit von KI-Outputs überhaupt erst als solche erkennen – was die bestärkende, richtungsweisende Antwortstruktur von KI-Systemen systematisch erschwert (Furnham & Marks, 2013). Unsere anspruchsvollste Form des Wissens ist voller Unsicherheiten, Widersprüche und Mehrdeutigkeiten – nicht reine, kalte Logik (Birhane 2021, S. 4). Diese Anforderung ist eingebettet in übergreifende gesellschaftliche Megatrends wie Individualisierung, Globalisierung, Diversität und Digitalisierung, wobei Digitalisierung oft Rückzug in Bestätigungsschleifen fördert (Hufer, 2023, S. 28).

Das AI-Sandwich-Modell: Grundstruktur

Freeman (2026) formuliert für den souveränen Umgang mit generativer KI im Bildungskontext – einschließlich der Führungsebene – drei Kernprinzipien: (1) Think first, then AI – eigene Denkleistung vor KI-Konsultation; (2) AI as co-thinker, not replacement – KI als Ko-Denkerin, nicht als Ersatz; (3) No AI output without human final pass – jede KI-Ausgabe durchläuft menschliche Prüfung. Der Begriff ‚Ko-Denkerin' bezeichnet bei Freeman (2026) die dialogisch-impulsgebende Funktion der KI, nicht eine Gleichsetzung mit menschlichem Denken im philosophischen Sinne.

Das Modell weist theoretische Verwandtschaften auf: zum Scaffolding-Ansatz (Wood, Bruner & Ross, 1976), wonach Unterstützung schrittweise reduziert wird; zum Human-in-the-Loop-Prinzip der KI-Entwicklung; und zur reflexiven Führung, die permanente Selbst- und Fremdreflexion fordert.

Human-First: strategische Autonomie

Die erste Schicht verlangt von Schulleitungen, vor KI-Konsultation eine eigene Position zu entwickeln. Diese Maxime entspricht der Erkenntnis, dass Bildung nur als Selbstbildung denkbar ist – jede Form von Fremdbildung würde die Eigentümlichkeit menschlicher Lebensweise verletzen (Zierer, 2024, 31). Nutzer*innen empfinden einen ständigen Fragenstrom eines lernenden Systems als unausgeglichen und nervig; sie schalten ihr mentales Modell ab und sind nicht bereit, als bloße „Orakel" zu fungieren (Amershi et al., 2014, 109). Dies verhindert den Automatismus des Path of Least Resistance, wonach Menschen beim Generieren neuer Ideen zuerst auf die zugänglichsten Basisbeispiele zurückgreifen und deren Eigenschaften projizieren – was die kreative Bandbreite einengt (Ward & Kolomyts, S. 2019, 178). Dies verhindert cognitive offloading – das unreflektierte Auslagern von Denkarbeit an die Maschine.

Theoretisch fundiert ist dieses Prinzip durch die Selbstbestimmungstheorie (Ryan & Deci, 2000). Die Selbstbestimmungstheorie postuliert, dass Autonomie ein angeborenes, universelles Bedürfnis ist. Wird dieses Bedürfnis durch eine vorschnelle KI-Konsultation unterlaufen, erlebt die Führungskraft ihr Handeln fremdbestimmt – was nach der Selbstbestimmungstheorie die Qualität der anschließenden Entscheidung mindert (Ryan & Deci, 2000, S. 68). Autonomie ist ein grundlegendes psychologisches Bedürfnis. Wird es durch KI-gesteuerte Entscheidungsprozesse untergraben, sinkt die intrinsische Motivation. Human-First schützt die Autonomie der Führungskraft, indem es die Rangordnung klärt: Menschliche pädagogische Urteilskraft hat Vorrang vor algorithmischer Berechnung. Die Priorisierung menschlicher Urteilskraft bedeutet, dass die Führungskraft das Problemziel und seine Struktur verstehen muss, bevor sie eine KI-gestützte Produktion von Lösungen bewertet. Dies entspricht der Erkenntnis, dass das Realisieren einer richtigen Lösung zeitlich und logisch vor der eigenständigen Herstellung dieser Lösung liegen muss – nur so bleibt der Mensch der Souverän des Prozesses (Wood, Bruner & Ross, 1976, S. 90). Dies bedeutet nicht Technikfeindlichkeit, sondern eine strategische Priorisierung, die auch die Grenzen von KI anerkennt – das Fehlen von implizitem Wissen (Polanyi, 1966).

AI-Middle: Dialog als ko-kreativer Prozess

Die mittlere Schicht institutionalisiert einen dialogischen Umgang mit KI. Die Selbstbestimmungstheorie betont, dass Wahlmöglichkeiten und Selbststeuerung die wahrgenommene Autonomie erhöhen. Indem die KI nicht als Befehlsempfänger, sondern als Ko-Denkerin agiert, wird die extrinsische Vorgabe in einen intrinsisch motivierenden Erkundungsprozess umgewandelt (Ryan & Deci, 2000, S. 70–71). Der dialogische Umgang ist die einzig mögliche Antwort auf die Black-Box-Problematik. Da das Wissen in neuralen Netzen nicht direkt einsehbar ist, können wir ihr Verhalten nur indirekt durch Testmuster und Beobachtung evaluieren – genau das tut der ko-kreative Dialog, indem er die KI immer wieder mit neuen Rückfragen und Widersprüchen konfrontiert (Matthias, 2004, S. 178–179). Die Führungskraft tritt nicht als Befehlende auf, sondern als Dialogpartner*in durch Nachfragen, Widerspruch oder dem Einfordern von Alternativen (Freeman, 2026). Metakognitives Wissen befähigt dazu, kognitive Aufgaben, Ziel und Strategien auszuwählen, zu bewerten, zu revidieren oder zu verwerfen. Genau das tut die Führungskraft im Dialog mit der KI, wenn sie nicht passiv übernimmt, sondern aktiv prüft, ob die vorgeschlagenen Strategien zu den eigenen pädagogischen Zielen passen (Flavel, 1979, S. 908). Die Verwendung eines nicht-spezifischen Ziels reduziert die kognitive Belastung erheblich, da beim Problemlösen nicht ständig Unterschiede zwischen aktuellem Zustand und Ziel abgeglichen werden muss. Der dialogische Umgang mit KI – bei dem nicht einfach eine Lösung abgerufen, sondern iterativ hinterfragt wird – kann ähnlich entlastend wirken, weil er die kognitive Last auf mehrere Schritte verteilt und reflexive Pausen erzwingt (Sweller, 1988, S. 262–263). Schulleitungen müssen eine Kultur des kontinuierlichen Lernens vorleben, indem sie selbst Neugier zeigen und sich mit KI-Tools auseinandersetzen. Sie bieten gezielte Fortbildungen zu ethischer KI-Nutzung, Datenkompetenz und instruktionalen Anwendungen an und fördern kollaborative Lernumgebungen, in denen Ex-

perimentieren und Reflexion möglich sind. Dies erfordert sowohl eigene Denkleistung (Human-First) als auch den ko-kreativen Austausch mit KI (AI-Middle) (DeMatthews et al., 2026, S. 10). Im ACDC Modell (Adjust, Check, Develop, Challenge) fordert Zierer (2024, S. 44), dass Medienkompetenz einen mühsamen Lernprozess mit vielen Gesprächen benötigt – Selbststeuerung lässt sich nicht per Knopfdruck einschalten. Der dialogische, reflexive Umgang mit KI ist genau die pädagogische Praxis, die hier beschrieben wird. Dies setzt metakognitive Kompetenz voraus (Flavell, 1979): die Fähigkeit, das eigene Denken zu beobachten und zu steuern. Flavell (1979, S. 906) bezeichnet damit jenes abgespeicherte Weltwissen, das sich auf Menschen als kognitive Wesen und deren diverse Aufgaben, Ziele, Handlungen und Erfahrungen bezieht. Genau diese Wissensart ist gemeint, wenn Führungskräfte ihr eigenes Denken im Umgang mit KI beobachten und lenken sollen – es handelt sich nicht um allgemeine Intelligenz, sondern um spezifisches Wissen über die eigenen kognitiven Prozesse.

Die Tiefe des dialogischen Prozesses hängt vom Entscheidungstyp ab: Bei strategischen Entscheidungen (z. B. der mittelfristigen Personal- und Ressourcenplanung oder der Festlegung von Qualitätszielen für die Schulentwicklung) sind mehrere Iterationen mit KI nötig, bei operativen (z. B. Terminplanung) genügt eine einzige Rückfrage. Diese Differenzierung korrespondiert mit der Unterscheidung zwischen informellen und formellen Diagnosen. Während formelle, standardisierte Diagnosen eher für eine Digitalisierung in Frage kommen, bleibt der Wert informeller Diagnosen für die schnelle und sichere Handlungsfähigkeit im Unterricht erhalten – was eine differenzierte Betrachtung des KI-Einsatzes nahelegt (Betschart, 2025, 57).

Kreativität entsteht durch das Zusammenbringen ungewohnter Perspektiven. KI kann hier als kognitive Erweiterung fungieren, indem sie Kombinationen generiert, die menschlichen Teams nicht einfallen. Die korrespondiert mit der Erkenntnis, dass gegensätzliche oder ungewöhnliche Paarungen zu Interpretationen zwingen, die in den Einzelkonzepten nicht angelegt sind, und damit die gedankliche Vielfalt erhöhen (Ward & Kolomyts, 2019, S. 187). Amabile (1996, S. 17) beschreibt, dass kreative Durchbrüche oft aus dem Zusammenspiel ungewohnter Perspektiven entstehen. Eine KI kann als Ko-Denkerin neue Assoziationen liefern. Entscheidend ist jedoch, dass diese Kombinationen nicht unkritisch übernommen werden, sondern in einen dialogischen Prozess eintreten, der die menschliche Urteilskraft wahrt. Allerdings zeigt die Forschung, dass Menschen zur unkritischen Übernahme von KI-Empfehlungen neigen, sobald die KI als ‚intelligent' wahrgenommen wird. Zusammenarbeit zwischen Mensch und Maschine führt daher nicht zwangsläufig zu besseren Ergebnissen (Vaccaro & Waldo, 2019, S. 37). Der dialogische Modus durchbricht diese Automatismen. Metakognitives Wissen kann zu einer breiten Vielfalt an Erfahrungen über Selbst, Aufgaben, Ziele und Strategien führen. Die KI als kognitive Erweiterung generiert ungewohnte Perspektiven, die wiederum solche neuen metakognitiven Erfahrungen auslösen – sie zwingen die Führungskraft, ihre bisherigen Schemata zu hinterfragen und neu zu justieren (Flavell, 1979, S. 908).

Zugleich trainiert die AI-Middle-Schicht Ambiguitätstoleranz. Diese gilt als Voraussetzung für konstruktive demokratische Beteiligung, da sie hilft, einfachen Lösungen zu widerstehen und Raum für kritisches Denken zu öffnen (Hufer 2023, 28). KI-Antworten sind selten eindeutig; Führungskräfte lernen, mit Unschärfe zu arbeiten und Entscheidungen auf der Basis unvoll-

ständiger Informationen zu treffen. Budner (1962, S. 30) zeigt, dass Personen auf mehrdeutige Situationen mit phänomenologischer oder operativer Unterwerfung (Angst, Unbehagen, Vermeidungsverhalten) oder mit Verleugnung (Umdeutung der Realität) reagieren – beides Indikatoren für Intoleranz. Der Begriff phänomenologische Unterwerfung von Budner (1962) bezeichnet die innere, gefühlsmäßige Reaktion auf eine als bedrohlich empfundene mehrdeutige Situation. Der dialogische Umgang mit der KI in der AI-Middle-Schicht trainiert genau die entgegengesetzte Haltung: Statt zu vermeiden oder zu vereindeutigen, wird die Unschärfe aktiv ausgehalten und produktiv genutzt.

Human-Last: die Rückholung der Verantwortung

Die dritte Schicht besagt: Keine KI-generierte Ausgabe darf ungeprüft in die schulische Praxis übernommen werden. Sweller (1988, S. 261) betont, dass beim konventionellen Problemlösen die Aufmerksamkeit auf die Reduktion von Unterschieden zwischen aktuellem Zustand und Ziel gerichtet ist – frühere Zustände und Operatoren werden ignoriert. Genau diese selektive Aufmerksamkeit verhindert Schemaerwerb. Ungeprüfte KI-Outputs führen zu einem ähnlichen Effekt. Die Führungskraft konzentriert sich nur auf das Ergebnis, nicht auf den Prozess, und verpasst so die Gelegenheit, eigene Schemata zu entwickeln.

Diese menschliche Letztverantwortung ist Ausdruck des grundlegenden Bildungsprinzips, dass jeder Mensch sein Personsein als lebenslange Aufgabe selbst zu leben hat – sie lässt sich nicht an eine Maschine delegieren (Zierer 2024, S. 31). Das Prinzip der Letztverantwortung findet seine theoretische Begründung in der Unüberbrückbarkeit des Hiatus zwischen Wissen und Handeln. Eine pädagogische Diagnose im Sinne der Gesamtbeurteilung eines Lernenden bleibt auf eine menschliche Instanz angewiesen, die mit ihrer Urteilskraft diese Lücke schließt – eine Fähigkeit, die sich nicht operationalisieren lässt (Betschart 2025, S. 58).

Nutzer*innen sind nicht mit „Black Box“ Lernsystemen zufrieden; sie wollen nuanciertes Feedback geben, um das System zu steuern, und sie sind bereit, Details über das System zu lernen (Amershi et al., 2014, S. 111). Dies ist ein verantwortungsethischer Imperativ (Jonas, 1979): Lernende Maschinen können definitionsgemäß keine Verantwortung im moralischen Sinne übernehmen, da ihnen die notwendigen Bedingungen – Kontrolle, Wissen um die Handlungsumstände und freie Wahl zwischen Alternativen – fehlen (Matthias, 2004, S. 175). Gleichzeitig verlieren auch Hersteller und Betreiber die Kontrolle, sodass eine Verantwortungslücke entsteht (ebd., S. 183). Die letzte Entscheidung muss beim Menschen liegen. Dieser Imperativ verhindert die Externalisierung des internalen Wahrnehmungsorts der Kontrolle. Nach der Selbstbestimmungstheorie (Ryan & Deci, 2000, S. 69) führt eine externe Wahrnehmung der Verursachung zu reduzierter Anstrengungsbereitschaft und schlechterer Bewältigung von Rückschlägen. Die menschliche Letztentscheidung bewahrt die Handlungsautonomie der Führungskraft, anstelle algorithmische Outputs zu vollstrecken. Vaccaro & Waldo (2019, S. 35) zeigen in ihrer Studie, dass die formale Letztentscheidung des Menschen nicht vor kognitiven Verzerrungen schützt. Die menschliche Letztverantwortung wird unterlaufen, wenn sie nicht reflexiv institutionalisiert ist. Danaher (2016, S. 259) meint, dass die

bloße Forderung nach menschlicher Überprüfung das Problem der Undurchschaubarkeit nicht löst. Zum einen können die Technologien selbst undurchschaubar sein, zum anderen würde die Bedrohung durch Algokratie lediglich durch die Bedrohung durch Epistokratie ersetzt werden – also durch die Herrschaft einer menschlichen Wissenselite.

Human-Last schützt vor Algorithmokratie (Danaher, 2016, S. 252; O'Neil, 2016) – einer Herrschaftsform, in der Algorithmen faktisch regieren, ohne demokratische Legitimation. Dieser Schutz ist auch kognitiv begründet, denn der Konformitätseffekt zeigt, dass selbst bei expliziter Aufforderung zur Abweichung vorgegebene Eigenschaften unkritisch übernommen werden – die menschliche Schlussprüfung ist daher das einzige Mittel, diesen Automatismus zu durchbrechen (Ward & Kolomyts, 2019, S. 182). Die Mathematisierung sozialer Probleme bringt einen Anschein von Objektivität und stellt Operationen als wertfrei, neutral und amoralisch dar – das ist eine Illusion (Birhane, 2021, S. 2). O'Neil (2016) betont, dass mathematische Modelle niemals wertfrei sind, sondern menschliche Vorurteile und ideologische Entscheidungen in eine scheinbar neutrale Form gießen. Im Bildungssystem, das auf Grundrechten wie Chancengleichheit beruht, ist dies eine ernstzunehmende Gefahr. Zugleich eröffnet der gezielte Einsatz von KI jedoch auch Chancen für mehr soziale Gerechtigkeit – etwa durch Algorithmen zur Optimierung der sozialen Durchmischung von Schulklassen, wie sie in der Schweizer Stadt Uster seit 2023 erfolgreich erprobt werden (SRF, 2023; SRF, 2026). Entscheidend ist daher nicht der KI-Einsatz an sich, sondern dessen transparente, ethisch reflektierte und an Gerechtigkeitskriterien ausgerichtete Gestaltung. Die institutionelle Verankerung der Letztverantwortung muss geklärt sein (z. B. Schulleitung allein oder im Team). Zugleich setzt Human-Last eine Fehlerkultur voraus, in der Widerspruch erlaubt ist (Edmondson, 2018). Eine reine Fokussierung auf Ambiguitätstoleranz würde allerdings die Frage nach Macht- und Herrschaftsverhältnissen ausblenden – auch hinter KI-gestützten Entscheidungen stehen Interessen, die transparent gemacht werden müssen (Hufer, 2023, 29).

Ein konkretes Beispiel verdeutlicht die Anwendung: Eine Klasse zeigt über mehrere Wochen ein belastendes soziales Klima – Mobbingvorfälle, Rückzug von Schüler*innen, zunehmende Unruhe. Die Schulleitung lädt die betroffenen Lehrpersonen zu einer gemeinsamen Reflexionsrunde ein. In der *Human-First*-Schicht formulieren alle Beteiligten zunächst ihre eigene Wahrnehmung der Situation und ihre ersten Handlungsimpulse, ohne KI-Konsultation – das sichert die strategische Autonomie der pädagogischen Urteilskraft. In der *AI-Middle*-Schicht tritt die Gruppe dann in einen ko-kreativen Dialog mit der KI: Sie lassen sich die Situation aus verschiedenen Perspektiven beschreiben (Lehrperson, betroffene Schüler*innen, Mitschüler*innen, Elternteil), generieren alternative Hypothesen zur Entstehung der Dynamik und sammeln mögliche Interventionsansätze. Die KI liefert dabei nicht die ‚richtige' Lösung, sondern erweitert den Lösungsraum – durch ungewohnte Perspektiven, systemische Zusammenhänge oder widersprüchliche Interpretationen. In der *Human-Last*-Schicht prüfen die Lehrpersonen gemeinsam mit der Schulleitung die KI-Outputs kritisch, verwerfen unpassende Vorschläge, passen andere an und entscheiden verantwortet, welche Maßnahmen im konkreten Fall umgesetzt werden sollen. Das Ergebnis ist kein fertiger Plan der KI, sondern ein reflektierter, gemeinsam getragener Handlungsrahmen – und zugleich ein Erfahrungsraum,

der zeigt, wie KI im Sinne des Sandwich-Modells souverän als Führungsinstrument genutzt werden kann.

Transfer in die Praxis

Die folgenden zehn Punkte übersetzen die theoretischen Prinzipien in konkrete Handlungsanleitungen für den Schulleitungsalltag. Sie sind als reflexive Werkzeuge zu verstehen, nicht als starres Regelwerk:

Human-First (Strategische Autonomie): eigene Position vor KI-Konsultation

1. Eigene Hypothese notieren – Schreiben Sie vor jedem KI-Prompt Ihre eigene Antwort stichpunktartig auf. Das bewahrt davor, unreflektiert in die vorgedachte Richtung der KI zu gehen.

2. Zeitlimit setzen – Nehmen Sie sich maximal fünf Minuten Zeit, um selbst zu überlegen. Das verhindert sowohl Grübeln als auch das vorschnelle Auslagern von Denkarbeit an die Maschine.

AI-Middle (Dialog als ko-kreativer Prozess): dialogisch mit der KI arbeiten

3. Drei-Alternativen-Regel – Fordern Sie die KI immer auf: „Nenne drei mögliche Lösungen, auch konträre." Das erweitert Ihren Blickwinkel und beugt vorschnellen Festlegungen vor.

4. Widerspruch einbauen – Nach der ersten KI-Antwort antworten Sie: „Begründen Sie, warum Ihr zweitbester Vorschlag besser wäre als Ihr bester." Das zwingt zur Reflexion und durchbricht gedankliche Pfadabhängigkeiten.

5. Halluzinations-Check – Lassen Sie sich von der KI ihre Quellen nennen – und überprüfen Sie mindestens eine davon.

6. Metakognitions-Prompt – Fragen Sie: „Welche Annahmen treffen Sie in Ihrer Antwort? Welche Informationen würden Ihre Antwort verändern?" Das macht implizite Voraussetzungen sichtbar und schult kritisches Denken.

Human-Last (Rückholung der Verantwortung)

7. Vier-Augen-Prinzip bei strategischen Entscheidungen – Lassen Sie jeden KI-Output von einer zweiten Person (oder im Team) gegenlesen. Das minimiert individuelle Blindstellen und stärkt kollektive Verantwortung.

8. Diskriminierungs-Check – Prüfen Sie systematisch: „Benachteiligt dieser Vorschlag eine Gruppe? (z. B. nach Herkunft, Geschlecht, Förderbedarf)". Das sichert die bildungspolitische Grundverpflichtung auf Chancengleichheit.

9. Fehlerdokumentation – Dokumentieren Sie jeden korrigierten KI-Fehler in einer internen Log-Datei. Das macht Fehler sichtbar, schafft eine lernförderliche Fehlerkultur und baut kollektive Erfahrung auf.

10. Organisationale Routine – Verankern Sie die Sandwich-Struktur in einer schriftlichen Dienstanweisung (z. B. als Teil des schulischen Qualitätshandbuchs). Das schafft Verbindlichkeit und verhindert den Rückfall in unreflektierte Routinen.

Diese zehn Punkte sind keine Checkliste, sondern ein Instrument zur Reflexion des eigenen KI-Handelns. Sie helfen, die Balance zu halten zwischen produktiver KI-Nutzung und dem Erhalt der eigenen professionellen Urteilskraft. Entscheidend ist nicht die mechanische Abarbeitung, sondern die wiederholte, bewusste Anwendung – bis sie zur selbstverständlichen Führungsroutine wird.

Die Abbildung 1 visualisiert das AI-Sandwich-Modell in seiner Dreischichtigkeit als Strukturierungsrahmen für kreative Schulentwicklung. Sie fasst die zentralen Prinzipien, theoretische Bezüge und praktische Leitlinien zusammen. Die Darstellung macht deutlich, dass es nicht um Technikoptimierung, sondern um die Sicherung pädagogischer Urteilskraft im Zeitalter generativer KI geht. Das Modell versteht KI nicht als Ersatz, sondern als Resonanzraum für professionelles Führungshandeln.

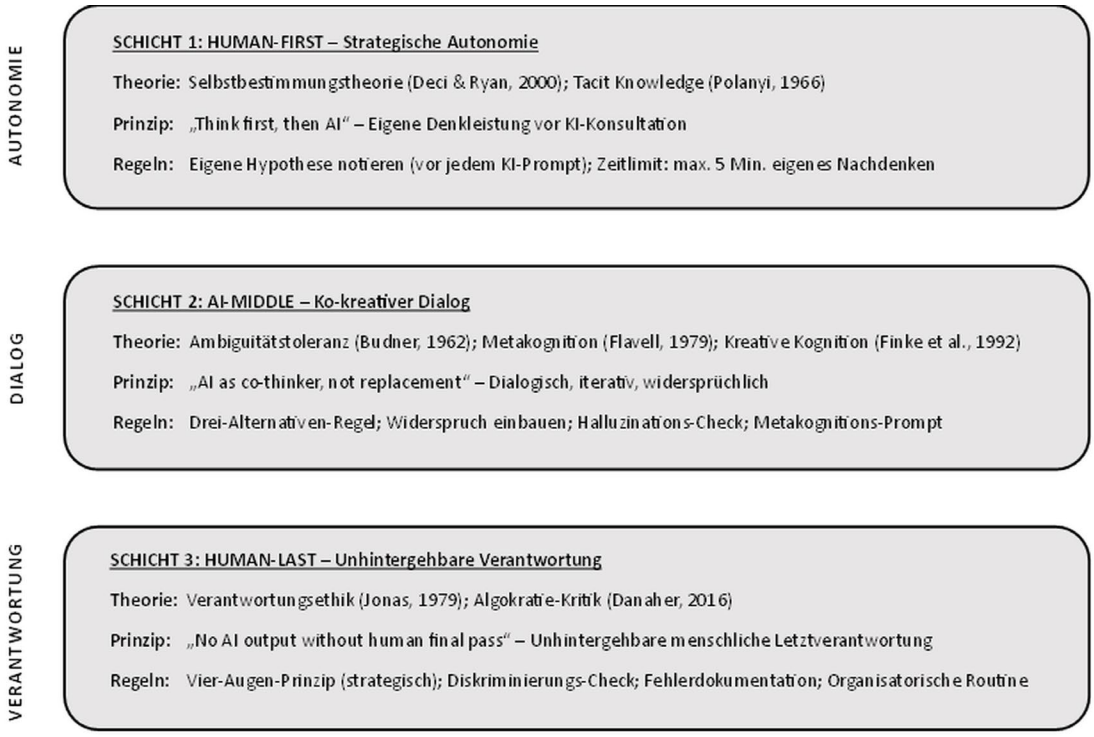

Abbildung 1: Das AI-Sandwich-Modell – Dreischichtiger Rahmen für kreative Führung mit generativer KI | eigene Darstellung

Grenzen des Modells

Sechs Einwände sind zu diskutieren. Erstens der kognitive Mehraufwand: Die Sandwich-Struktur erfordert zusätzliche reflexive Schleifen, die nach dem Konzept der kognitiven Belastung (Sweller, 1988) das Arbeitsgedächtnis stärker beanspruchen als eine einstufige KI-Konsultation. Dies kann sich in einem subjektiv höheren Zeitaufwand niederschlagen. Freeman (2026) kontert, dass die Folgekosten ungeprüfter KI-Entscheidungen höher sind. Zweitens die Frage der Zuschreibung: Wer genau ist der ‚Human'? Dies muss je nach Entscheidungstyp (strategisch, operativ, disziplinarisch) spezifiziert werden. Drittens droht Instrumentalisierung: Eine pro forma durchgeführte Endkontrolle ohne wirklichen Widerspruch entwertet das Modell. Vaccaro & Waldo (2019, S. 34) belegen in ihrer Studie diese Gefahr: Teilnehmer*innen nutzten den Algorithmus nur dann, wenn er mit ihrem eigenen Urteil übereinstimmte – eine scheinbare menschliche Prüfung, die in Wahrheit die eigene Voreingenommenheit bestätigt. Andere passten ihre Antwort lediglich minimal an den Algorithmus an, ohne eine eigenständige Prüfung vorzunehmen. Viertens braucht das Sandwich-Modell eine grundlegende KI-Kompetenz bei Führungskräften, die nicht überall gegeben ist. Fünftens setzt Human-Last voraus, dass die Führungskraft den KI-Output fachlich beurteilen kann. Bei hochkomplexen Analysen (z. B. Daten zu Lernverläufen) ist dies oft nicht der Fall – dann muss die ‚menschliche Letztprüfung' durch ein Team oder externe Expertise ersetzt werden. Sechstens ist zu fragen, ob alle Widersprüche ertragen werden sollen. Konflikte erfordern Engagement, nicht Tolerierung (Hufer, 2023, S. 29). Ambiguitätstoleranz gegenüber KI-Output darf nicht in Gleichgültigkeit gegenüber diskriminierenden oder demokratiefeindlichen Vorschlägen umschlagen.

Fazit

Das AI-Sandwich bietet eine theoretisch fundierte Heuristik für kreative Schulleitung im KI-Zeitalter. Die Heuristik ist theoretisch fundiert, weil sie die von Matthias (2004, S. 183) diagnostizierte doppelte Kontrollproblematik aufnimmt: Weder die Maschine noch der Mensch können die Lücke allein schließen. Das Sandwich institutionalisiert reflexive Barrieren, die den Technologieeinsatz erlauben, ohne die moralische Konsistenz der Gesellschaft aufzugeben. Human-First sichert strategische Autonomie (Ryan & Deci, 2000), AI-Middle erschließt kreative Potenziale durch dialogische Nutzung, Human-Last institutionalisiert unhintergehbare menschliche Verantwortung (Jonas, 1979; Danaher, 2016). Schulleitungen müssen ethische Aufsicht über KI gewährleisten – von der Entwicklung bis zur Umsetzung und fortlaufenden Bewertung. Dies erfordert, dass sie Algorithmen und Datenquellen regelmäßig überprüfen, um zu verhindern, dass KI bestehende Ungleichheiten verstärk oder neue Schäden für Schüler*innen oder Mitarbeiter*innen verursacht (DeMatthews et al., 2026, S. 6). Wer die letzte Verantwortung an ein System delegiert, verliert die motivationale Grundlage für echte Kreativität (Amabile 1996, S. 15). Diese These wird durch die organisationstheoretische Perspektive gestützt, die zeigt, dass eine Veränderung der organisationalen Arbeitsumgebung

durch digitale Technologien den Spielraum für Varianzen und Abweichungen minimiert und damit kreative Spielräume von Professionellen einschränkt (Betschart 2025, S. 55).

Das Modell ist kein Algorithmus, sondern eine Reflexionsstruktur. Ob es sich in der Praxis bewährt, ist eine empirische Frage – seine theoretische Konsistenz ist der erste Schritt.

Literaturverzeichnis

Amabile, T. M. (1996). *Creativity in context: Update to the social psychology of creativity*. London: Routledge.

Amabile, T. M., & Khaire, M. (2008). Creativity and the role of the leader. *Harvard Business Review, 86*(10), 100–109.

Amabile, T. M. (1997). Motivating creativity in organizations: On doing what you love and loving what you do. *California Management Review, 40*(1), 39–58.

Amershi, S., Cakmak, M., Knox, W. B. & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine, 35*(4), 105–120.

Betschart, B. (2025). Professionalität und digitalisierte pädagogische Diagnosen in schulischen Organisationen: Herausforderungen und Innovationspotenzial für die Profession der Lehrpersonen. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung,* (67). https://doi.org/10.21240/mpaed/67/2025.11.22.X

Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns, 2*(2), 100205.

Budner, (1962). Intolerance of ambiguity as a personality variable. *Journal of Personality, 30*(1), 29–50.

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology, 29*(3), 245–268. https://doi.org/10.1007/s13347-015-0211-1

DeMatthews, D., Reyes, P., Hart, T. D. & James, L., III. (2026). Leadership for artificial intelligence use in schools: A six-domain framework for ethical, equitable, and effective integration. *Educational Management Administration & Leadership*. Advance online publication. https://doi.org/10.1177/17411432261418940

Dubs, R. (2019). *Die Führung einer Schule. Leadership und Management* (3. Aufl.). Franz Steiner.

Edmondson, A. C. (2018). *The fearless organization: Creating psychological safety in the workplace for learning, innovation, and growth*. Hoboken, NY: John Wiley & Sons.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911.

Freeman, L. (2026, January 21). *The AI Sandwich: Human-First, Human-Last in an AI-Rich Educational Environment.* OLC Insights. https://onlinelearningconsortium.org/olc-insights/2026/01/the-ai-sandwich/ Stand vom 22. Juli 2026.

Furnham, A. & Marks, J. (2013). Tolerance of ambiguity: A review of the recent literature. *Psychology, 4*(9), 717–728. https://doi.org/10.4236/psych.2013.49102

Jonas, H. (1979). *Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation. Insel*. Frankfurt a.M.: Suhrkamp.

Hufer, K.-P. (2023). Ambiguitätstoleranz: Ein Kurswechsel der politischen Bildung. *weiter bilden, 30*(4), 27–29.

Leithwood, K., Harris, A. & Hopkins, D. (2008). Seven strong claims about successful school leadership. *School Leadership & Management, 28*(1), 27–42. https://doi.org/10.1080/13632430701800060

Mainemelis, C., Kark, R., & Epitropaki, O. (2015). Creative leadership: A multi-context conceptualization. *Academy of Management Annals, 9*(1), 393–482. https://doi.org/10.5465/19416520.2015.1024502

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown.

Polanyi, M. (1966). *The tacit dimension*. New York: Doubleday.

Ryan, R. M. & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Schein, E. H. (2010). *Organizational culture and leadership* (4th ed.). Jossey-Bass.

SRF. (2023, August 21). Schweizweite Premiere in Uster – Zum ersten Mal teilt ein Computer Schulklassen ein. *Schweizer Radio und Fernsehen*. https://www.srf.ch/news/schweiz/schweizweite-premiere-in-uster-zum-ersten-mal-teilt-ein-computer-schulklassen-ein

SRF. (2026, August 26). Wie ein Algorithmus für besser durchmischte Klassen sorgt. *Schweizer Radio und Fernsehen*. https://www.srf.ch/news/schweiz/innovatives-zuteilungssystem-mit-steuerdaten-zu-durchmischten-schulklassen

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Vaccaro, M. & Waldo, J. (2019). The effects of mixing machine learning and human judgment: Collaboration between humans and machines does not necessarily lead to better outcomes. *ACM Queue, 17*(4), 19–40.

Ward, T. B., & Kolomyts, Y. (2019). Creative cognition. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity* (pp. 175–199). Cambridge: Cambridge University Press

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

Zierer, K. (2024). *ChatGPT als Heilsbringer? Über Möglichkeiten und Grenzen von KI im Bildungsbereich*. Münster: Waxmann.

Autor

Thomas Benesch, Tit. Univ. Prof. HS-Prof. Dr. habil DDr.
Seit 2012 Dozent und Hochschulprofessor an der Privaten Pädagogischen Hochschule Burgenland sowie Titularprofessor für Wirtschafts- und Organisationswissenschaften an der Westungarischen Universität. Zuvor als Universitätsassistenz und Fachhochschulprofessor in den Bereichen Biometrie und Informationsmanagement tätig. Autor von über 200 wissenschaftlichen Publikationen zu Bildungswissenschaft, angewandter Statistik und Fachdidaktik. Nebenberuflich als Unternehmensberater, Coach und Gutachter aktiv. Mitglied in Fachgesellschaften (ÖFEB, GDM).
Kontakt: thomas.benesch@ph-burgenland.at