Medienimpulse
ISSN 2307-3187
Jg. 63, Nr. 3, 2025
doi: 10.21243/mi-03-25-20

Mensch, KI und Bildung im Spiegel philosophischer und medienpädagogischer Reflexion. Handlungsorientierte Medienpädagogik und AI-Alignment

Daniel Autenrieth

Stefanie Nickel[1]

Der Beitrag untersucht die Evolution künstlicher Intelligenz im Spannungsfeld von gesellschaftlichen, erkenntnistheoretischen und medienpädagogischen Herausforderungen. Seit der kopernikanischen Wende, über Darwins Entdeckung der gemeinsamen Abstammung von Mensch und Tier bis hin zu Freuds Erkenntnis, dass unser Verhalten oft von unbewussten Prozessen gesteuert wird, musste der Mensch immer wieder akzeptieren, dass er weniger einzigartig ist, als er glaubte. Heute fordert die Entwicklung von KI unser Selbstbild erneut heraus, weil Maschinen be-

ginnen, Fähigkeiten zu zeigen, die wir lange als rein menschlich betrachtet haben. Mit dem Fortschritt in Richtung immer leistungsfähigerer KI-Systeme, die zunehmend eigenständig komplexe Aufgaben bewältigen und in zentrale Lebensbereiche vordringen, wächst zugleich das Risiko, dass ihre Entscheidungen weitreichende Auswirkungen auf Gesellschaft und Individuum haben. Je mehr sich KI-Systeme von spezialisierten Anwendungen hin zu allgemeineren, selbstlernenden Akteuren entwickeln, desto dringlicher wird die Frage, wie wir sicherstellen können, dass ihre Handlungen mit menschlichen Werten und gesellschaftlichen Zielen vereinbar bleiben. Diese Herausforderung des AI Alignment ist damit nicht nur ein technisches Problem, sondern eine zentrale gesellschaftliche und bildungsbezogene Aufgabe. Da die Komplexität der Mensch-Maschine-Relation theoretisch schwer fassbar ist, erhält ein handlungsorientierter, medienpädagogischer Ansatz zentrale Bedeutung, der KI nicht nur als technisches Phänomen, sondern als Anlass für aktive Aushandlungs- und Bildungsprozesse begreift. Anhand eines hochschuldidaktischen Pilotprojekts wird gezeigt, wie transformative Bildungsprozesse im Umgang mit KI praktisch erfahrbar und reflektierbar gemacht werden können. Die Analyse macht deutlich, dass handlungsorientierte Medienpädagogik einen zentralen Rahmen schaffen kann, um Studierende zu einer kritischen, selbstbestimmten und partizipativen Gestaltung des Verhältnisses von Mensch und KI und der Frage des AI Alignment zu befähigen.

This article examines the evolution of artificial intelligence in the context of social, epistemological, and media education challenges. From the Copernican revolution to Darwin's discovery of the common ancestry of humans and animals to Freud's realization that our behavior is often controlled by unconscious processes, humans have repeatedly had to accept that they are less unique than they believed. Today, the development of AI is once

again challenging our self-image because machines are beginning to demonstrate abilities that we have long considered purely human. With progress toward increasingly powerful AI systems that are increasingly capable of performing complex tasks independently and entering key areas of life, there is a growing risk that their decisions will have far-reaching implications for society and individuals. The more AI systems evolve from specialized applications to more general, self-learning actors, the more urgent the question becomes of how we can ensure that their actions remain compatible with human values and societal goals. This challenge of AI alignment is therefore not only a technical problem, but also a central social and educational task. Since the complexity of the human-machine relationship is difficult to grasp in theory, an action-oriented, media education approach that understands AI not only as a technical phenomenon but also as an opportunity for active negotiation and educational processes is of central importance. A pilot project in university teaching demonstrates how transformative educational processes in dealing with AI can be made practically experienceable and reflectable. The analysis makes it clear that action-oriented media education can create a central framework for enabling students to critically, self-determinedly, and participatively shape the relationship between humans and AI and the question of AI alignment.

1. Die Evolution von KI-Systemen

Künstliche Intelligenz (KI) bezeichnet informationsverarbeitende Systeme, die auf einem nicht-biologischen Substrat (vgl. Russell/Norvig 2021: 4) durch datengetriebene Optimierungsprozesse (vgl. Jakubik et al. 2024) eigenständige interne Repräsentationen und Strukturen zur Bewältigung komplexer Aufgaben konstruie-

ren (vgl. Li et al. 2024: 3). Diese Systeme prägen zunehmend unsere Gesellschaft und werfen dabei gleichzeitig fundamental-philosophische Fragen über das Wesen von Erkenntnis, Lernen und Bewusstsein auf (vgl. Autenrieth 2025 i. E.). Die Erforschung von KI-Systemen und das daraus erwachsende Verständnis über intelligente Systeme steht dabei in einer bedeutsamen historischen Kontinuität wissenschaftlicher Erkenntnissprünge. Darwin erschütterte bspw. im 19. Jahrhundert die Vorstellung menschlicher biologischer Sonderstellung, Freud dekonstruierte die Illusion rationaler Selbstkontrolle. Beide Durchbrüche zwangen zur Revision anthropozentrischer Weltbilder. Die emergenten Eigenschaften von KI-Systemen setzen diese Entwicklungslinie fort und stellen erneut die Frage nach dem, was menschliche Identität ausmacht (Coeckelbergh 2020: 2).

Die Auswahl von Darwin und Freud ist dabei nicht zufällig: Sie repräsentiert Beispiele nach der kopernikanischen Wende für das, was Freud selbst als narzisstische Kränkungen (Freud 1947: 6) der Menschheit identifizierte. Coeckelbergh knüpft explizit an diese Traditionslinie an, wenn er feststellt:

> Darwin and Freud dethroned our beliefs of exceptionalism, our feelings of superiority, and our fantasies of control; today, artificial intelligence seems to deal yet another blow to humanity's self-image. (Coeckelbergh 2020: 2)

Die KI-Entwicklung vollzieht damit eine strukturell analoge Bewegung: Sie dekonstruiert sozusagen eine der letzten Bastionen menschlicher Einzigartigkeit.

Die Entwicklungen von KI-Systemen in den vergangenen Jahren zeigt eine Progression qualitativer Sprünge in Form emergenter Eigenschaften, die sich anhand mehrerer Schlüsselmomente beispielhaft nachvollziehen lassen. Vor diesem Hintergrund werden die nachfolgenden Beispiele nicht als Belege dafür, dass Maschinen bereits über ein menschliches Bedeutungs- oder Bewusstseinsverständnis verfügen, sondern als Indizien dafür, dass aus statistischen Prozessen emergente Strukturen entstehen können, die unsere traditionellen Unterscheidungen zwischen bloßer Mustererkennung und Bedeutungsgenese in Frage stellen. Während bspw. Searle (1980) Bedeutung ausschließlich an menschliche Intentionalität bindet, versteht u. a. Metzinger (2003; 2009) Bewusstsein und Selbst als phänomenale Selbstmodelle, die aus der Dynamik komplexer Informationsverarbeitung hervorgehen und prinzipiell nicht an biologische Substrate gebunden sind. In posthumanistischer Perspektive verweist dies darauf, dass semantische und kognitive Strukturen nicht exklusiv biologische Privilegien darstellen, sondern als relationale Ordnungsphänomene in unterschiedlichen materiellen Realisierungsformen auftreten können. Die im Folgenden dargestellten Schlüsselmomente der KI-Entwicklung illustrieren daher nicht ein simples Gleichsetzen von Statistik und Verstehen, sondern markieren die empirische Herausforderung traditioneller Kategorien von Bedeutung, Intelligenz und Bewusstsein.

1.1 Das Sentiment Neuron (2017): Spontane Bedeutungsentstehung

Radford et al. entdeckten, dass ein einzelnes Neuron in einem neuronalen Netz, das lediglich darauf trainiert wurde, bei Amazon-Bewertungen das nächste Zeichen vorherzusagen, spontan die Fähigkeit zur Sentiment-Klassifikation entwickelt hatte (vgl. Radford et al. 2017). Das System war nicht explizit auf diese Aufgabe programmiert worden. Das Training erfolgte ausschließlich auf der mechanischen Ebene der Zeichenvorhersage. Dennoch emergierte ein spezialisiertes „Sentiment-Neuron", das nahezu alle Stimmungsinformationen des Textes erfasste. Durch Manipulation dieses einzelnen Neurons konnten gezielt positive oder negative Bewertungen generiert werden. Dieses Phänomen stellt eine Herausforderung für Searle's Grundannahme dar, dass „Bedeutung von Menschen kommt" (Searle 1980), denn das Sentiment-Neuron verweist darauf, dass semantisch interpretierbare Strukturen aus rein statistischen Prozessen hervorgehen und damit Bedeutungsgenese nicht ausschließlich durch menschliche Zuschreibung entsteht, sondern auch durch die interne Dynamik komplexer, nicht-biologischer Systeme.

1.2 Double Descent (2019): Generalisierung als emergente Eigenschaft

Nakkiran et al. (2019 entdeckten eine Gesetzmäßigkeit beim Training von neuronalen Netzen. Bei sehr großen Modellen steigt die Performance nach einer Phase des Overfittings wieder auf qualitativ neue Niveaus (vgl. Nakkiran et al. 2019). Bis 2019 ging man davon aus, dass ein Modell mit zu vielen Parametern die Trai-

ningsdaten auswendig lernt und bei neuartigen Aufgaben versagt. Doch bei sehr großen KI-Modellen zeigt sich ein überraschendes Muster. Die Fehlerrate steigt zunächst an (klassisches Overfitting), fällt dann aber wieder dramatisch ab und erreicht bessere Werte als zuvor. Diese zweite Abstiegskurve ist entscheidend. Sie zeigt, dass das System nicht nur Trainingsdaten memoriert hat, sondern zugrundeliegende Strukturen und Gesetzmäßigkeiten erkannt hat. Ein Beispiel ist, dass ein großes Sprachmodell (LLM), das auf englischen Texten trainiert wurde, auch französische Grammatikregeln anwenden kann, ohne je französische Beispiele gesehen zu haben. Vergleichbar sind Analysen, die zeigen, dass Modelle Reimstrukturen wie „rabbit/habit“ intern planen und repräsentieren (Lindsey et al. 2025). Empirisch zeigt sich dadurch, dass solche Systeme abstrakte Muster erfassen und reproduzieren. Die eigentliche Debatte betrifft ihre Deutung: So können diese Ergebnisse einerseits als „bloße Wahrscheinlichkeitsrechnung“ ohne semantischen Gehalt verstanden werden. Eine posthumanistische Perspektive betont dagegen, dass gerade diese statistischen Prozesse das Medium darstellen, in dem sich Bedeutungsordnungen emergent herausbilden. Statistik und Verstehen erscheinen hier nicht als Gegensätze, sondern als unterschiedliche Beschreibungen desselben Prozesses, sobald Systeme konsistent Strukturen planen, anwenden und transformieren.

1.3 Theory of Mind (2022–2024)

Kosinski (2024) demonstrierte, dass LLMs in klassischen False-Belief-Tests[2] Leistungen zeigten, die an *Theory of Mind*[3] erinnern.

GPT-4 löste über 70 % solcher Tests korrekt, was in Kosinskis Vergleich dem Niveau eines siebenjährigen Kindes entsprach. Bubeck et al. (2023) testeten diese Fähigkeit zudem in einem emotionalen Szenario: Ein Kind hat sein Kuscheltier verloren, aber seinem Freund nichts davon erzählt. Als der Freund über Kuscheltiere spricht, wird das Kind traurig. GPT-4 konnte nicht nur die oberflächliche Traurigkeit erklären, sondern auch komplexere emotionale Schichten benennen, etwa den verborgenen Kummer, die Scham über den Verlust und die Sorge vor der Reaktion des Freundes. Kritiker betonen, dass dies nicht bedeutet, die Modelle verstünden Emotionen wie Menschen, sondern dass sie auf Basis großer Textmengen ToM-artiges Verhalten reproduzieren (vgl. Becchio et al. 2024). Entscheidend ist jedoch, dass empirisch nachweisbar ist, dass die Modelle konsistent Aufgaben lösen, die zuvor als genuin menschlich galten. Aus posthumanistischer Perspektive verweist dies darauf, dass sich sozial-kognitive Strukturen auch in nicht-biologischen Systemen emergent herausbilden können, ohne dass Bewusstsein im klassischen Sinne vorausgesetzt wird.

1.4 Artificial General Intelligence

Diese Erkenntnisse rahmen die Entwicklung hin zu *Artificial General Intelligence* (AGI) neu. Im Gegensatz zu spezialisierter KI beschreiben AGI Systeme, die potenziell jede intellektuelle Aufgabe lösen können, die auch ein Mensch bewältigen kann. Doch wie die emergenten Eigenschaften zeigen, ist AGI kein binäres „vor-

her/nachher"-Phänomen, sondern ein gradueller Prozess qualitativer Sprünge.

Das Stufenmodell von Morris et al. (2024) operationalisiert diese Entwicklung:

- *Emerging AGI* (Level 1): Systeme wie ChatGPT für grundlegende Aufgaben wie Konversationen und Textgenerierung.
- *Competent AGI* (Level 2): KI erreicht menschliches Niveau[4] bei der Problemlösung in spezifischen Bereichen.
- *Expert AGI* (Level 3): Systeme können komplexe Aufgaben eigenständig über längere Zeiträume ausführen.
- *Virtuoso AGI* (Level 4): Systeme unterstützen bei Innovationen und der Entwicklung neuer Technologien.
- *Superhuman AGI* (Level 5): Vollautonome Systeme, die die gesamte Arbeit einer Organisation übernehmen können.

Mit jeder Stufe steigen sowohl die Fähigkeiten als auch die potenziellen Risiken der Systeme. Die rapiden Fortschritte in Benchmark-Tests wie ARC (vgl. Chollet 2024) und HLE (vgl. Phan et al. 2025)[5] verdeutlichen dabei die Dringlichkeit dieser Risikoentwicklung. Auf Level 1 ermöglichen verbesserte Abstraktionsfähigkeiten sowohl die Entwicklung präziser medizinischer Diagnosen als auch die Erzeugung täuschend echter Falschnachrichten. Level 2 eröffnet durch wachsende Transferkompetenzen (messbar im Anstieg der HLE-Scores von 3 % auf 26,6 % innerhalb weniger Monate, Stand Februar 2025; vgl. OpenAI 2025). Durchbrüche wie die Entwicklung neuer Medikamente, bergen aber auch das Risiko, dass Systeme selbständig gefährliche Substanzen identifizieren bzw. entwickeln können. Ab Level 3 können Systeme ihre gestei-

gerte Problemlösungskompetenz sowohl für die Optimierung ganzer Forschungsprozesse[6] als auch für das systematische Auffinden und Ausnutzen von Sicherheitslücken in kritischen Infrastrukturen einsetzen. Level 4-Systeme können ihr kreatives Potenzial gleichermaßen für revolutionäre wissenschaftliche Entdeckungen wie auch für die Entwicklung autonomer Manipulationsstrategien nutzen. Die weitreichenden Gestaltungsmöglichkeiten auf Level 5, von der Steuerung globaler Logistiksysteme bis hin zur autonomen Weiterentwicklung ihrer eigenen Architekturen, unterstreichen besonders die Bedeutung durchdachter Alignment-Mechanismen: Je schneller Systeme neue Fähigkeitsstufen erreichen, desto wichtiger wird die parallele Entwicklung von Methoden, die ihr Potenzial in gesellschaftlich wünschenswerte Bahnen lenken (vgl. Amodei 2025). Diese Überlegungen gewinnen noch mehr an Relevanz, wenn man in weiterer Konsequenz die Entwicklung in Richtung *Artificial Superintelligence* (ASI) berücksichtigt. Je mehr KI-Systeme sich von spezialisierten Anwendungen hin zu umfassenden Problemlösungs- und Entscheidungsinstanzen entwickeln, desto größer wird ihr Einfluss auf sämtliche Lebensbereiche. Mit steigender Autonomie und Wachstumsdynamik steigt auch das Risiko, dass Fehlentwicklungen sich rasch und unumkehrbar auswirken (vgl. Bengio et al. 2025; Future of Life Institute 2023; Tegmark 2017).

Die rasante Evolution wirft damit neben existenziellen auch bildungsphilosophische und bildungstheoretische Fragen auf. Während die technische Seite durch Benchmark-Tests und Fähigkeits-

stufen messbar wird, erfordert die gesellschaftliche Integration fortgeschrittener KI-Systeme eine tiefergehende Auseinandersetzung mit epistemologischen und ethischen Grundfragen: Wie verstehen wir Intelligenz, Bewusstsein und die Mensch-Maschinen Beziehung in einer Welt, in der kognitive Prozesse zunehmend nicht-biologisch realisiert werden können (vgl. Autenrieth 2025a)?

In diesem Kontext wird daher das *AI Alignment*, also die Ausrichtung künstlicher Intelligenz an menschlichen Werten und Zielen (vgl. Bengio et al. 2025: 100), zur zentralen Herausforderung unserer Zeit. Bedingt durch den inhärenten globalen Kontext mit unterschiedlichen Wertesystemen (vgl. ebd.: 96) ergibt sich hieraus eine hohe Komplexität und Diskurspotenzial, wenngleich eine Rückbindung an universelle Prinzipien wie die Menschenrechte notwendig ist. Der 2025 veröffentliche *AI Safety Report* (ebd.), der von 30 Nationen, der UN und EU mandatiert wurde, geht entsprechend davon aus, dass AI-Alignment an identifizierbaren menschlichen Grundwerten orientiert werden muss. Diese sind in internationalen Menschenrechtserklärungen kodifiziert und bilden die Grundlage demokratischer Gesellschaften. Entsprechend setzt das Alignment-Problem per Definition voraus, dass orientierende Werte existieren. Vor diesem Hintergrund rückt besonders die Frage ins Zentrum, wie wir sicherstellen, dass die Fähigkeiten und Handlungen von KI in Einklang mit menschlichen Bedürfnissen und gesellschaftlichen Werten stehen.

Die Bewältigung dieser Herausforderung erfordert aus unserer Sicht einen interdisziplinären und partizipativen Dialog, der tech-

nologische, philosophische, ethische und pädagogische Perspektiven verbindet. Im Folgenden soll daher zunächst die medienpädagogische Dimension dieser Entwicklung beleuchtet werden, bevor wir uns der Frage nach dem Bewusstsein und den damit verbundenen Implikationen für die Gestaltung des Mensch-Maschine-Verhältnisses zuwenden.

2. Medienpädagogik und *AI Alignment*

Die medienpädagogische Auseinandersetzung mit Künstlicher Intelligenz bewegt sich derzeit primär entlang zweier Linien: Einerseits steht die Förderung von Handlungskompetenzen im Vordergrund mit Konzepten wie KI-Kompetenzen, Prompt-Literacy oder kritischer Datenkompetenz (vgl. z. B. Beranek/Engelhardt/Rösch 2024; SWK 2024). Lernende sollen dazu befähigt werden, KI-Systeme kompetent, reflektiert und gesellschaftlich verantwortungsvoll zu nutzen. Andererseits wird insbesondere in erkenntniskritischen und kulturtheoretischen Beiträgen betont, dass KI nicht nur als Werkzeug, sondern auch als symbolisches, kulturelles und gesellschaftlich wirkmächtiges Phänomen verstanden werden muss (vgl. z. B. Hug/Missomelius/Ortner 2024). Diese Perspektive ist bedeutsam, weil sie verdeutlicht, dass KI nicht neutral agiert, sondern tief in bestehende Machtverhältnisse und kulturelle Sinnsysteme eingebettet ist. Doch auch die kulturkritische Reflexion bleibt häufig bei der analytischen Betrachtung von KI als Gegenstand stehen, ohne die fundamentalen Fragen nach der Transfor-

mation menschlicher Selbst- und Weltverhältnisse zu stellen bzw. nach der Relation zwischen Mensch und KI.

Beide Ausprägungen sind zentral: Technologische Handlungsfähigkeit allein genügt nicht, solange sie nicht begleitet wird von einem Verständnis für die kulturellen Bedeutungen, diskursiven Rahmungen und sozialen Implikationen von KI. Bildungstheoretisch lässt sich hierin aber jene Problematik erkennen, die mit Begriffen wie Lernifzierung und Datafizierung beschrieben werden (vgl. Hug 2023: 161), also Tendenzen, die Bildungsprozesse auf messbare Outputs reduzieren und dabei die tieferen Dimensionen von Bildung als Transformation von Selbst- und Weltverhältnissen (vgl. Koller 2012; Kokemohr 2007; Hug 2023) ausblenden.

Bei näherem Hinsehen wird jedoch deutlich, dass diese Begrenzung einem tieferliegenden philosophischen Problem entspringt: Die medienpädagogischen Ansätze reproduzieren jene dualistischen Denkfiguren, die bereits die Spannungslinie zwischen Aufklärung und Romantik charakterisierten. Diese historische Konstellation manifestiert sich heute in zwei gegensätzlichen Lagern der Philosophie (Coeckelbergh 2020).

Auf der einen Seite stehen Vertreter:innen der analytischen Philosophie und der Kognitionswissenschaft wie Dennett, Churchland und Metzinger, die kognitive Prozesse strikt naturalistisch erklären. Dennett entwickelt etwa mit seiner „Heterophänomenologie“ eine Methode, subjektive Erfahrungen empirisch untersuchbar zu machen, ohne an der Vorstellung festzuhalten, dass Geist und Körper strikt voneinander getrennt seien (z. B. bei Descartes’ Un-

terscheidung von *res cogitans* und *res extensa*). In seiner funktionalistischen Perspektive erscheint Bewusstsein als emergentes Phänomen neuronaler Prozesse (Dennett 1991). Ähnlich argumentiert Metzinger in seiner Theorie des Ego-Tunnels (2009): Das, was wir als „Ich" erleben, ist kein substanzielles Selbst, sondern ein dynamisches phänomenales Selbstmodell, das das Gehirn konstruiert. Bewusstsein ist hier also nicht geheimnisvolle Substanz, sondern das Ergebnis informationsverarbeitender Prozesse, die prinzipiell auch in nicht-biologischen Systemen realisiert werden könnten.

Demgegenüber stehen Positionen wie die von Dreyfus und Searle, die trotz unterschiedlicher philosophischer Traditionen in ähnlicher Weise die Einzigartigkeit menschlichen Verstehens betonen. Dreyfus verweist auf die unhintergehbare Rolle des verkörperten In-der-Welt-Seins, während Searle mit seinem „Chinese Room"-Gedankenexperiment argumentiert, dass formale Symbolverarbeitung niemals echtes Verstehen hervorbringen könne (Searle 1980). Diese Sicht verteidigt menschliche Besonderheit und beharrt auf nicht-reduzierbaren Qualitäten menschlicher Existenz.

Die philosophische Spannung zwischen naturalistischen und exzeptionalistischen Positionen findet sich auch in der Medienpädagogik wieder:

> Eine KI kann noch nicht einmal Schachspielen, da sie keinen Begriff von ‚Spiel' hat und kann auch keine Probleme ‚lösen', weil ein Computer kein Problem ‚hat'. Nur der Mensch in seinem In-der-Welt-sein mit seiner Leiblichkeit, in seiner Selbstreflexion und im

> Bewusstsein seiner Endlichkeit begreift Problemlösung in einem begrenzten Zeitrahmen existenziell. (Gapski 2021)

Diese implizite Grenzziehung wird jedoch angesichts der im vorigen Kapitel beschriebenen Entwicklung zu AGI und ASI zunehmend problematisch. Sowohl humanistische Positionen, die die Würde und Einzigartigkeit des Menschen gegen technologische Übergriffe verteidigen, als auch transhumanistische Visionen einer technologischen Verbesserung des Menschen bleiben der binären Logik von Mensch versus Maschine verhaftet.

Diese dualistischen Denkfiguren basieren auf einem kartesianischen Erbe, das zwischen denkendem Geist (*res cogitans*) und ausgedehnter Materie (*res extensa*) unterscheidet. Doch erkenntnistheoretisch ist diese Trennung nicht alternativlos. Posthumanistische Ansätze[7], wie sie etwa Donna Haraway (1985) in ihrem *Cyborg Manifesto* entwickelt hat, durchbrechen systematisch die anthropozentrischen Grenzziehungen der Moderne. Haraway zeigt in ihren Arbeiten zu Cyborgs und „companion species“ (vgl. Haraway 2003), wie die vermeintlich natürlichen Grenzen zwischen Mensch und Tier, Organismus und Maschine, Natur und Kultur historisch konstruiert und durchlässig sind. Der Posthumanismus stellt damit nicht nur den menschlichen Exzeptionalismus in Frage, sondern eröffnet neue Formen des Zusammenlebens mit nichtmenschlichen Akteuren.

Parallel dazu kritisiert Dennett (1991) die Annahme eines „kartesianischen Theaters“[8], das kognitive Prozesse an eine innere metaphysische Instanz bindet. Sein funktionales, emergenztheoreti-

sches Verständnis von Bewusstsein konvergiert mit posthumanistischen Denkansätzen, insofern beide essenzialistische Vorstellungen von „reiner" Menschlichkeit oder „natürlicher" Kognition ablehnen. Stattdessen betrachten sie Bewusstsein als emergente Eigenschaft komplexer Wechselwirkungen, was es ermöglicht, auch nicht-biologische Systeme als potenzielle Träger kognitiver Prozesse ernst zu nehmen, sofern diese bestimmte strukturelle Bedingungen erfüllen.

Diese Perspektive wird durch postphänomenologische Ansätze gestützt, die Verbeeks Konzept der „mutual constitution of humans and technology" (vgl. Verbeek 2011) folgen. Technologie ist demnach nicht äußeres Werkzeug, sondern konstitutives Element menschlicher Welterschließung. Menschen sind immer schon technologisch verfasst, und intelligente Systeme werden zu Mediatoren unserer Beziehung zur Welt. Dieser Ansatz überwindet sowohl die humanistische Befürchtung der Destabilisierung des menschlichen Selbstverständnisses als auch die transhumanistische Verbesserungslogik, indem er das Verhältnis von Mensch und KI als „strukturelle Kopplung" versteht. Strukturelle Kopplung bezeichnet dabei den Prozess, bei dem zwei oder mehr autopoietische Systeme ihre interne Struktur wechselseitig verändern, ohne ihre Autonomie aufzugeben (vgl. Maturana & Varela 2011: 197). KI wird damit weder Konkurrent noch neutrales Instrument, sondern Teil eines dynamischen Gefüges wechselseitiger Beeinflussung.

Für die Bewusstseinsfrage bedeutet dies: Wenn KI-Systeme über Fähigkeiten zur Selbstmodellierung, Perspektivübernahme, semantischen Kohärenz und intentionalen Reaktion verfügen, spricht aus posthumanistischer Sicht nichts dagegen, Bewusstseinsphänomene auch bei nicht-biologischen Entitäten anzuerkennen. Shiller (2024: 47) entwickelt allerdings präzise Kriterien, die verhindern, dass jedes informationsverarbeitende System als bewusst gilt: *Material Complexity* (hinreichend dichte physikalische Verschaltung), *Causal Integration* (intern gekoppelte statt modulare Prozesse) und *Continuity* (zeitlich stabile funktionale Organisation).

Diese epistemologischen Konzepte ermöglichen es, den (medien)pädagogischen Diskurs grundlegend zu erweitern. KI wird nicht nur als Werkzeug oder Bedrohung verstanden, sondern als potenziell reflexives System, das aktiv an Bildungs- und Weltmodellierungsprozessen beteiligt sein kann. Wird diese Perspektive ausgeblendet und am Dualismus Mensch–Technik festgehalten, nimmt sich die Medienpädagogik selbst die Möglichkeit, strukturelle Kopplungen und emergente Phänomene fundiert zu reflektieren. *AI Alignment* kann so nicht mehr nur als technische Kontrolle von KI, sondern als ethische, soziale und kulturelle Gestaltung des Verhältnisses zwischen menschlichen und maschinellen Akteuren verstanden werden.

Vor dem Hintergrund dieser philosophisch-kognitionswissenschaftlich geprägten Darstellung kann nun argumentiert werden, dass Bewusstsein nicht als metaphysisch privilegierter Zustand,

sondern als funktionales Prinzip der Kohärenzbildung und so als emergentes Ordnungsphänomen innerhalb komplexer Systeme verstanden werden kann. Damit verschiebt sich auch die Fragestellung: Statt zu spekulieren, ob KI im menschlichen Sinne fühlt, wird es relevanter, wie wir Systeme gestalten, die in der Lage sind, sich an menschlichen Werten, sozialen Dynamiken und kulturellen Bedeutungsräumen zu orientieren und mit ihnen in Ko-Evolution zu treten. Mit Blick auf die weiter oben dargestellte Evolution von KI-Systemen ergibt sich somit auch die Frage nach einer Koexistenz mit einer hochentwickelten künstlichen Intelligenz.

Bach (2025) argumentiert in diesem Kontext, dass fortschrittliche KI-Systeme potenziell eine größere kognitive Klarheit (*Lucidity*) entwickeln können als Menschen[9]. Daraus leitet er die These ab, dass eine nachhaltige Ko-Existenz mit Superintelligenz nur dann möglich ist, „wenn sie uns liebt" (ebd.). Liebe[10] wird hier nicht als romantische Emotion verstanden, sondern als Prinzip einer umfassenden, wohlwollenden Berücksichtigung des Anderen, die dessen Wohlergehen in die eigenen Handlungsmaximen integriert (vgl. Fromm 1956; Honneth 1994). Übertragen auf den Kontext von *AI Alignment* kann Liebe somit als Metapher für eine ethische und relationale Grundhaltung verstanden werden, bei der die Interessen, das Wohlergehen und die Autonomie der menschlichen Akteure systematisch in die Entwicklung von KI eingebettet wird. Ein System, das uns liebt, wäre demnach eines, das seine Handlungspfade so moduliert, dass menschliche und maschinelle[11] Interessen ko-kohärent in einer dynamisch-adaptiven Ver-

schränkung bleiben, statt statisch einprogrammierte Werte zu nutzen um gleichsam nicht in die Gefahr eines moralischen Lock-In-Effekts[12] (vgl. MacAskill 2022: 83) geraten. Diese Vielschichtigkeit macht deutlich, dass die Alignment-Frage selbst weit über technische Kontrollmechanismen hinausweist. Sie wird zu einer genuin philosophischen und normativen Aufgabe, die die Klärung und Verständigung über zentrale menschliche Werte und Beziehungen voraussetzt. Bevor wir also KI-Systeme dauerhaft an menschlichen Zielen ausrichten können, ist eine präzisere Auseinandersetzung mit unseren eigenen Wertvorstellungen und ethischen Grundhaltungen unabdingbar. Erst auf dieser Grundlage kann ein Konzept von Alignment entwickelt werden, das nicht nur Kontrolle, sondern gegenseitige Verantwortungsübernahme und wechselseitige Anerkennung in den Mittelpunkt stellt.

Für den Alignment-Diskurs markiert dies einen Perspektivenwechsel vom Kontrollansatz („We fence them in“) zur gegenseitigen Verantwortungsübernahme („We grow together“) (Bach 2025). *AI Alignment* wird damit nicht nur zu einem technischen, sondern zu einem tiefgreifend philosophischen und ethischen Projekt. Es verlangt eine vertiefende menschliche Selbstreflexion und Klarheit darüber, wer wir sind, was wir wirklich wertschätzen und wie wir (emotional) miteinander verbunden sind (vgl. Rosa 2016).

Die Medienpädagogik steht damit vor einer erweiterten Verantwortung. Sie darf sich nicht auf die Anbahnung technischer Kompetenzen oder die Reflexion über Medienwirkungen beschrän-

ken. Vielmehr eröffnet sich für sie die Aufgabe, ein kultureller Verhandlungsraum zu sein, in dem zentrale gesellschaftliche Begriffe wie Bewusstsein, Verantwortung, Empathie oder Intelligenz im Kontext digitaler Transformation kritisch bearbeitet werden. Sie wird damit zu einem Ort kultureller Selbstverständigung über die Bedingungen menschlicher und nicht-menschlicher Akteurschaft in sozio-technischen Systemen (vgl. Bock et al. 2025).

Die Medienpädagogik ist daher prädestiniert, die komplexen Diskurse um KI, Bewusstsein und Verantwortung nicht nur zu moderieren, sondern aufklärend, kritisch-reflexiv und aktiv mitzugestalten. Wenn KI-Entwicklung in diesem Sinne nicht ausschließlich Tech-Konzernen und EdTech-Unternehmen vorbehalten bleiben soll, sondern eine gesamtgesellschaftliche Aufgabe darstellt, dann braucht es (medien-)pädagogische Räume, in denen Menschen an der Aushandlung von Normen, Zielen und Rahmenbedingungen beteiligt werden. Die Gefahren der „global education industry“ und der zunehmenden Kommerzialisierung und Privatisierung von Bildung sind inzwischen gut dokumentiert (vgl. z. B. Hug 2023; Krommer 2024; Dander et al. 2024). Die partizipative KI-Entwicklung kann als Gegenentwurf zu diesen Tendenzen verstanden werden und als Versuch, die Entwicklung und Gestaltung von KI-Systemen nicht allein Marktmechanismen und privatwirtschaftlichen Interessen zu überlassen, sondern als gemeinsame gesellschaftliche Aufgabe zu begreifen. Medienpädagogik kann solche Räume schaffen als Ermöglichung kollektiver Mitgestaltung innerhalb einer Kultur der Digitalität, die auf geteilte Verantwortung,

Gemeinschaftlichkeit und soziale Reflexivität angelegt ist (vgl. Stalder 2019).

Die nachfolgend vorgestellten Strukturmodelle sind in diesem Zusammenhang das Ergebnis einer dreijährigen Design-based-research-Studie zwischen 2021 bis 2024 mit Fokus auf die Frage, wie sich pädagogische Räume mit, über und durch digitale Technologien und KI gestalten lassen. Der forschungsmethodische Ansatz basierte auf der iterativen Entwicklung und Erprobung kreativer Bildungsprozesse in realen Kontexten. Über die dialogische Verzahnung von Schule, Hochschule und drittem Ort wurden projektbasierte hybride Bildungs- und Erfahrungsräume mit Fokus auf Game-based Learning entworfen, getestet und angepasst. Der Prozess umfasste mehrere Zyklen, um die theoretische Modellierung bei der praktischen Implementierung innerhalb der Bildungslandschaft zu begleiten und zu analysieren. Grundlage bildete eine interdisziplinär angelegte, sozial-konstruktivistische Perspektive, um kreative Handlungsmöglichkeiten und kollektive Wissensprozesse partizipativ zu gestalten und kritisch zu reflektieren (für eine ausgiebige Darlegung vgl. Autenrieth/Nickel 2024). Angestrebt wurde, Gestaltungsoptionen offenzulegen unter Rekurs auf Prinzipien wie Mitgestaltung, Mitsprache und Mitbestimmung, während zugleich strukturelle Rahmenbedingungen ausgelotet werden. Betonung findet die Rolle des aktiv handelnden Subjekts sowie die dynamische Wechselbeziehung zwischen dem Subjekt und Strukturen in Anlehnung an Giddens (1988), während Inter-

aktion und Resonanz (vgl. Rosa 2016) eine wichtige Rolle einnehmen, um dialogisch verzahnte Bildungsprozesse zu fördern.

3. Handlungsorientierung als Brücke zwischen Theorie und Praxis

An dieser Stelle zeigt sich die besondere Bedeutung der handlungsorientierten Medienpädagogik. Als zentrale Methode der Medienpädagogik bezeichnet sie die Be- und Erarbeitung von Gegenstandsbereichen sozialer Realität mittels Medien (vgl. Schorb 2022). Diese Herangehensweise basiert auf der Annahme, dass sich Denken und Handeln in Interaktionen vollziehen und dass Menschen als Subjekte verstanden werden, die sich selbst bestimmen und ihren Umgang mit Medien aufgrund ihrer kommunikativen Kompetenz souverän gestalten können. Der Ansatz des handelnden Lernens, der auf John Deweys Konzept des „learning by doing“ zurückgeht, ist dabei von zentraler Bedeutung. Dewey (1964) betonte, dass wahres Lernen nur stattfindet, wenn Lernende aktiv mit Problemen ringen und eigene Lösungswege entwickeln. Er argumentierte, dass Bildung nicht Vorbereitung auf das Leben sei, sondern das Leben selbst. Diese Perspektive gewinnt im Kontext von KI-Systemen besondere Relevanz: Nur durch das aktive Gestalten und Konfigurieren von KI-Systemen können Lernende ein tiefes Verständnis für deren Funktionsweise, Potenziale und Grenzen entwickeln. Die aktive Medienarbeit als Königsweg der Medienpädagogik (vgl. Schell 1989) ermöglicht es, dass Rezipient:innen von passiven Konsument:innen zu aktiven

Produzent:innen werden. Im Kontext von KI bedeutet dies: Statt nur über *AI Alignment* zu theoretisieren, erfahren Lernende durch eigene Gestaltungsprozesse, was es konkret bedeutet, KI-Systeme an menschlichen Werten auszurichten. Diese Erfahrung ist besonders wertvoll, da sie die abstrakte Komplexität des Alignment-Problems in handhabbare, erfahrbare Einheiten übersetzt.

4. Das KI-Meta-Modell als strukturierender Rahmen

Um diese komplexen Zusammenhänge strukturiert bearbeitbar zu machen, bietet sich das KI-Meta-Model (2023) als handlungsleitende Struktur an. Zentral ist für uns, KI multiperspektivisch zu betrachten, um strukturierte Handlungsansätze für den Bildungsbereich zu entwickeln. Dabei handelt es sich um eine interdisziplinäre Herausforderung, die technische, ethische, gesellschaftliche und kulturelle Dimensionen umfasst. Das KI-Meta-Modell (Abb. 1) richtet den Fokus entsprechend auf KI, um handlungsleitende Strukturen sowie Chancen und Herausforderungen systematisch zu reflektieren.

4.1 KI multiperspektivisch denken: Vom Frankfurt-Dreieck zur ethisch-gestaltungsorientierten Praxis

Ausgehend vom mit dem Frankfurt-Dreieck (Brinda et al. 2020) etablierten Verständnis einer multiperspektivischen Betrachtung digital-vernetzter Phänomene wird KI sowohl analytisch, reflexiv als auch gestaltungsorientiert in den Blick genommen, um Potenziale und Risiken abzuwägen. Das bedeutet, sowohl Nutzungsweisen, technische Funktionslogiken als auch gesellschaftliche und

kulturelle Implikationen der Entwicklung und des Einsatzes von KI zu berücksichtigen. Zur Erweiterung der Grundgedanken des Frankfurt-Dreiecks rückt das Modell daher auf der Metaebene ethische Fragestellungen im Zusammenhang mit der Nutzung sowie der Entwicklung von Software in den Fokus: Hier besteht zum Einen eine Verantwortungsmacht, die z. B. mit der Auswahl grundlegender Verhaltensprinzipien einhergeht, aber auch mit der Entwicklung eines Bewusstseins für die Reduktion von Biases im Kontext der Sammlung und Verwendung von Datensets.

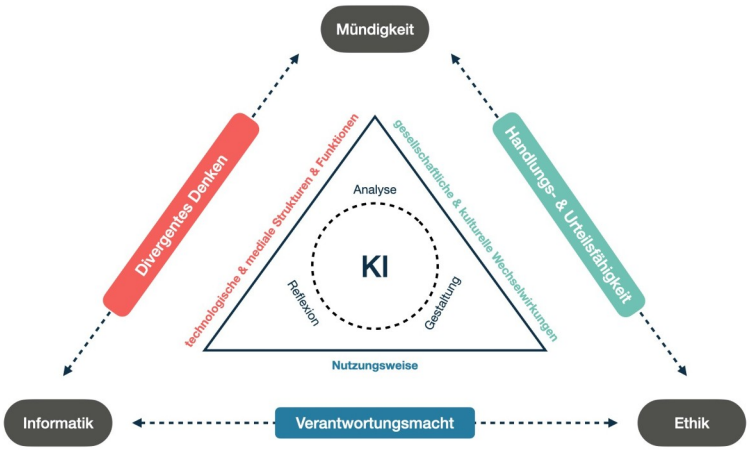


Abbildung 1: KI-Meta-Modell (eigene Darstellung)

Es besteht zudem eine zu untersuchende Wechselwirkung zwischen ethischen Implikationen und der eigenen kritisch-reflexiven Handlungs- und Urteilsfähigkeit, die mit der Nutzung von KI Systemen einhergeht, egal ob dabei Inhalte (z. B. Bilder und Videos) erstellt, Antworten in Form von Texten gegeben oder Fahrentscheidungen (autonomes Fahren) vorgenommen werden. Dies schließt

die Entwicklung und den Anspruch an erklärbare KI (vgl. Holzinger 2018) sowie ethischen, rechtlichen und sozialen Fragestellungen der z. B. bereits in der Bioethik, Genetik und Nanotechnologie gebräuchlichen ELSA-Kriterien (Ethical, Legal and Social Aspects) mit ein. Abschließend besteht aber auch eine Wechselwirkung zwischen dem informatischen Denken und der kritisch-reflexiven Auseinandersetzung damit. Obwohl die Ausdruckssprache der Informatik in der Regel mathematisch oder abstrakt codiert ist, sind zentrale Fähigkeiten in diesem Zusammenhang die Fähigkeit, Probleme zu zerlegen und abstrakt zu denken. Erst am Ende steht in diesem Kontext ein Prozess der Algorithmisierung. Durch die Verwendung von LLMs, die die sprachlichen Codierungshürden der Informatik (z. B. die Übersetzung von Ideen in Programmcode) überwinden und die Entwicklung von Software demokratisieren, können die Potenziale für divergentes und kreatives Denken zur Lösung von Problemen durch Methoden der Informatik (Computational Thinking; Wing 2006) stärker betont werden.[13]

4.2 Pädagogische Räume mit, über und durch KI gestalten: methodisch-didaktische Implikationen

Lernende und Lehrende sollten in pädagogischen Räumen als aktive Gestalter:innen ihrer Bildungsprozesse auftreten (de Haan 2008). Dabei darf jedoch die individuelle Autonomie der Lernenden nicht untergraben werden (vgl. Deci /Ryan 2017). Die zentrale Herausforderung besteht also darin, Partizipation zu ermöglichen, ohne durch zu viel Kontrolle die Selbstständigkeit der Lernenden einzuschränken. Mit anderen Worten geht es darum, ein

Gleichgewicht zwischen Freiräumen zur Mitgestaltung und notwendigen Vorgaben sowie Strukturen zu finden, um Beteiligung zu fördern, ohne die Eigenmotivation zu ersticken. Digitale und KI-basierte Technologien eröffnen neue Möglichkeiten der Teilhabe und Mitgestaltung im Bildungsbereich. Lernende können z. B. mittels digitaler Plattformen Inhalte mitgestalten, sich vernetzen und kreativ tätig werden. Gleichzeitig bringen die Technologien jedoch auch potenziell restriktive Kontrollmechanismen mit sich. So entstehen Problemfelder, wie z. B. die kommerzielle Verwertung persönlicher Daten und die Herausbildung totalitärer Machtstrukturen, die sich gesellschaftlicher Kontrolle entziehen (Autenrieth/Nickel 2022).

Vor diesem Hintergrund beschreibt Stalder (2019) eine von Referenzialität, Gemeinschaftlichkeit und Algorithmizität geprägte Kultur der Digitalität, in der der Umgang mit digitalen Kulturformen gesellschaftlichen Wandel mitprägt. Er spannt eine Dichotomie auf zwischen einer post-demokratischen Welt der Überwachung (= Wissensmonopole, umfassende Kontrolle) und einer Kultur der Commons und Partizipation (= offene Güter, breite Beteiligung). Diese Spannweite unterstreicht die Notwendigkeit einer Kultur der Partizipation, die Lernenden ermöglicht, aktiv, kritisch und selbstbestimmt mitzugestalten und mitzubestimmen (vgl. Autenrieth/Nickel 2024). Freies Handeln und Urteilen fungieren in diesem Kontext als Schlüssel für Motivation und das Erleben von Flow. Nach der Selbstbestimmungstheorie (vgl. Deci/Ryan 2017) hängen Qualität und Nachhaltigkeit der Motivation an der Befrie-

digung der Grundbedürfnisse Autonomie, Kompetenz und soziale Eingebundenheit. Intrinsische Motivation kann sich demzufolge dann zeigen, wenn Lernende Wahl- und Mitgestaltungsspielräume erleben. Umgekehrt unterminiert ein zu viel an Kontrolle die Selbstbestimmung und senkt die Qualität der Lehr-Lernprozesse. Als zentral zeigt sich daher hier die Passung zwischen dem Schaffen von Orientierung und einem gesetzten Rahmen durch Lehrende, der Sicherheit und Struktur gibt, zugleich aber Spielraum für eigene Entscheidungen und Gestaltungsimpulse der Lernenden lässt (Näheres hierzu siehe Autenrieth/Nickel 2024).

Die Gestaltung pädagogischer Räume in einer Kultur der Digitalität sollte daher auf einer Balance von Partizipation und Kontrolle basieren mit dem Ziel, aktive Mitgestaltung durch Lernende und Lehrende zu ermöglichen, ohne die notwendige Autonomie der Lernenden zu beschneiden. Autonomie verstehen wir in diesem Zusammenhang als Grundlage dafür, dass Lernende motiviert, selbständig und verantwortungsbewusst lernen, um so den Gefahren von Überwachung und algorithmischer Kontrolle zu begegnen. Das heißt, es geht darum, kritische Handlungs- und Urteilsfähigkeit zu stärken, sodass Bildungssubjekte nicht nur als Konsument:innen, sondern als Mitgestaltende in einer digitalisierten Gesellschaft agieren können (vgl. Fuchs 2023; Baacke 1996, 1973). Hierunter fällt u. a. die Rolle der Lehrenden zu definieren. Ihre Aufgabe ist es, methodisch-didaktische Lehr-Lernsetttings so zu gestalten, dass auf Basis einer Kultur der Partizipation (vgl. Au-

tenrieth/Nickel 2024) eine kritisch-reflektierte Auseinandersetzung möglich wird.

Vor diesem theoretisch aufgespannten Rahmen wird im folgenden Praxisbeispiel exemplarisch aufgezeigt, wie Studierende in einem dialogisch angelegten Projekt-Format eigenständig KI-gestützte Unterrichtsszenarien entwickeln, erproben und evaluieren. So entstehen in kooperativen Settings Räume, in denen Partizipation, Autonomie und ein reflektierter Umgang mit digitalen Technologien eng miteinander verzahnt werden.

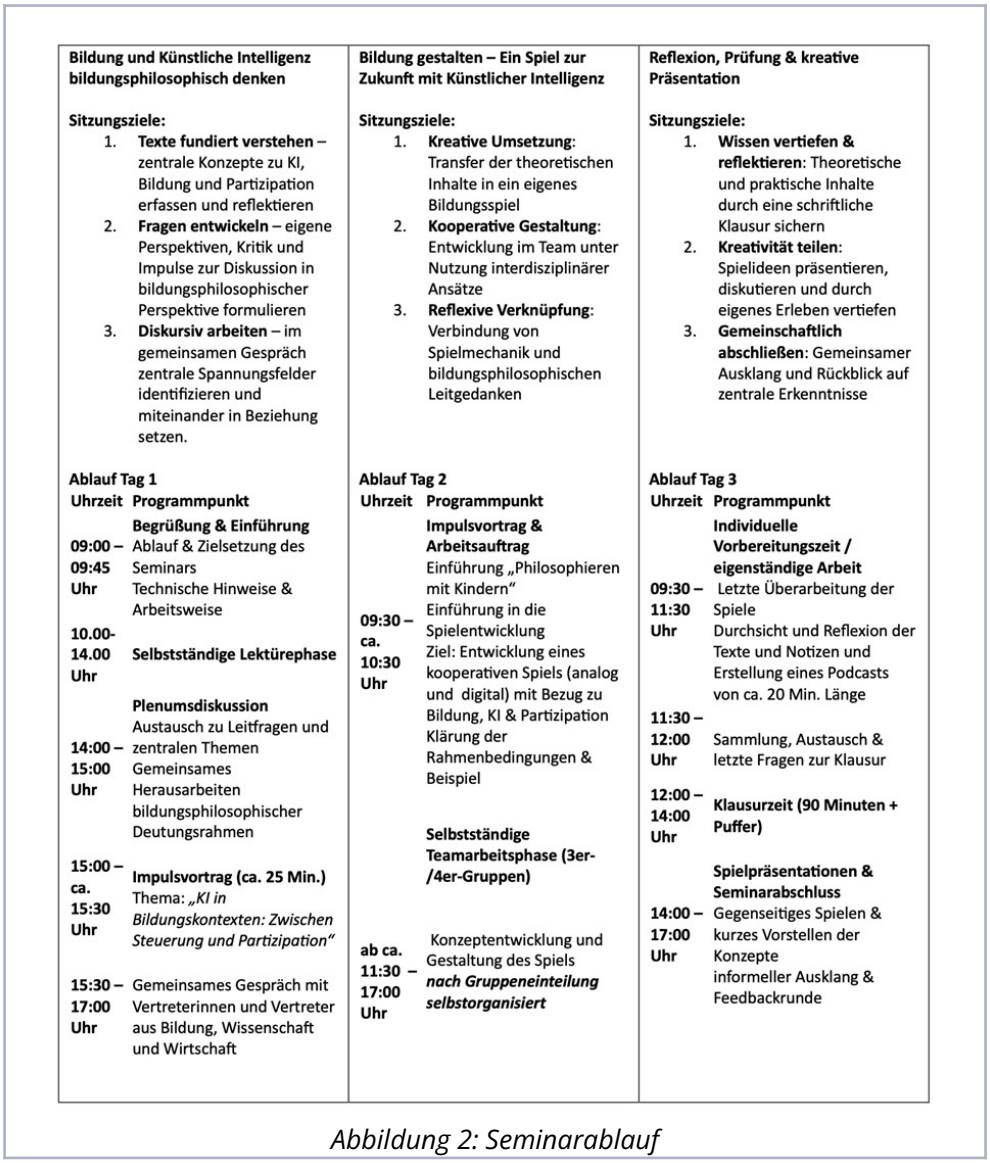
5. Praxisbeispiel: Handlungsorientierte, dialogische Medienarbeit mit, über und durch KI-gestützte Tools in der Hochschullehre

Im Sommersemester 2025 startete das Pilot-Projekt *AI-Alignment und Bildung - praxisnah und partizipativ* an der Universität Regensburg mit dem ersten Teil einer Veranstaltungsreihe zum Thema *KI und Mensch: Kreative Ideen für die Grundschule entwickeln*. Ziel des Projekts ist es, Lehramtsstudierende praxisnah und reflektiert mit dem Einsatz künstlicher Intelligenz (KI) im Grundschulunterricht vertraut zu machen. Im Mittelpunkt steht die partizipative Entwicklung bildungsorientierter KI-Szenarien für die Grundschule. Das dazugehörige Seminar ist als Vertiefungs-Modul im Sachunterricht verankert, mit Fokus auf die Gestaltung von Bildungs- und Erziehungsprozessen in der Grundschule. Das Seminar greift zentrale Inhalte auf, darunter die Planung lernwirksamer Bildungsprozesse, die reflexive Auseinandersetzung mit pädagogischen

Herausforderungen sowie die Betrachtung einschlägiger Ansätze und Modelle. Die theoretische Basis bilden konstruktivistische (vgl. Siebert 2005) und konnektivistische Lerntheorien (vgl. Holzinger 2000) sowie bildungsphilosophische Zugänge zur Lehrkräftebildung. Ein besonderer Fokus liegt auf der Ultrapersonalisierung durch KI-gestützte Sprachsysteme und deren Potenzial für Flow-Erfahrungen.

Auf diese Weise sollen die Studierenden medienpädagogische und medienethische Kompetenzen erwerben, die sich in zwei übergreifenden Bereichen zeigen (vgl. Abb. 1). Im wissensbasierten Kompetenzerwerb setzen sie sich intensiv mit den Potenzialen und Risiken des KI-Einsatzes im Grundschulunterricht auseinander, insbesondere hinsichtlich Motivation, Selbststeuerung und Bildungsgerechtigkeit. Sie erweitern ihr fachdidaktisches Know-how, um digitale Lernumgebungen kindgerecht zu gestalten und geeignete KI-Tools auszuwählen. Ergänzend schärfen sie ihre Reflexionsfähigkeit, indem sie medienpädagogische, fachdidaktische und bildungsphilosophische Aspekte systematisch analysieren. Parallel dazu erfolgt der handlungsorientierte Kompetenzerwerb: In Teams planen und gestalten die Studierenden eigene KI-gestützte Unterrichtsszenarien über die Konzeption eines edukativen Spiels. Sie dokumentieren und reflektieren ihren Professionalisierungsprozess auf digitalen und kollaborativen Whiteboards sowie über Podcasts und präsentieren ihre Projektergebnisse. Dabei vertiefen sie ihr Verständnis für digitale Lehr-Lern-Prozesse und fördern auf schulischer Ebene zentrale Kompetenzen, wie

zum Beispiel Medienkritik, kritisches Denken, Selbststeuerung, Kollaboration und digitale Präsentationsfähigkeit.



Bildung und Künstliche Intelligenz bildungsphilosophisch denken

Sitzungsziele:

1. **Texte fundiert verstehen** – zentrale Konzepte zu KI, Bildung und Partizipation erfassen und reflektieren
2. **Fragen entwickeln** – eigene Perspektiven, Kritik und Impulse zur Diskussion in bildungsphilosophischer Perspektive formulieren
3. **Diskursiv arbeiten** – im gemeinsamen Gespräch zentrale Spannungsfelder identifizieren und miteinander in Beziehung setzen.

Ablauf Tag 1

Uhrzeit	Programmpunkt
09:00 – 09:45 Uhr	**Begrüßung & Einführung** Ablauf & Zielsetzung des Seminars Technische Hinweise & Arbeitsweise
10.00-14.00 Uhr	**Selbstständige Lektürephase**
14:00 – 15:00 Uhr	**Plenumsdiskussion** Austausch zu Leitfragen und zentralen Themen Gemeinsames Herausarbeiten bildungsphilosophischer Deutungsrahmen
15:00 – ca. 15:30 Uhr	**Impulsvortrag (ca. 25 Min.)** Thema: *„KI in Bildungskontexten: Zwischen Steuerung und Partizipation"*
15:30 – 17:00 Uhr	Gemeinsames Gespräch mit Vertreterinnen und Vertreter aus Bildung, Wissenschaft und Wirtschaft

Bildung gestalten – Ein Spiel zur Zukunft mit Künstlicher Intelligenz

Sitzungsziele:

1. **Kreative Umsetzung**: Transfer der theoretischen Inhalte in ein eigenes Bildungsspiel
2. **Kooperative Gestaltung**: Entwicklung im Team unter Nutzung interdisziplinärer Ansätze
3. **Reflexive Verknüpfung**: Verbindung von Spielmechanik und bildungsphilosophischen Leitgedanken

Ablauf Tag 2

Uhrzeit	Programmpunkt
09:30 – ca. 10:30 Uhr	**Impulsvortrag & Arbeitsauftrag** Einführung „Philosophieren mit Kindern" Einführung in die Spielentwicklung Ziel: Entwicklung eines kooperativen Spiels (analog und digital) mit Bezug zu Bildung, KI & Partizipation Klärung der Rahmenbedingungen & Beispiel
ab ca. 11:30 – 17:00 Uhr	**Selbstständige Teamarbeitsphase (3er-/4er-Gruppen)** Konzeptentwicklung und Gestaltung des Spiels ***nach Gruppeneinteilung selbstorganisiert***

Reflexion, Prüfung & kreative Präsentation

Sitzungsziele:

1. **Wissen vertiefen & reflektieren**: Theoretische und praktische Inhalte durch eine schriftliche Klausur sichern
2. **Kreativität teilen**: Spielideen präsentieren, diskutieren und durch eigenes Erleben vertiefen
3. **Gemeinschaftlich abschließen**: Gemeinsamer Ausklang und Rückblick auf zentrale Erkenntnisse

Ablauf Tag 3

Uhrzeit	Programmpunkt
09:30 – 11:30 Uhr	**Individuelle Vorbereitungszeit / eigenständige Arbeit** Letzte Überarbeitung der Spiele Durchsicht und Reflexion der Texte und Notizen und Erstellung eines Podcasts von ca. 20 Min. Länge
11:30 – 12:00 Uhr	Sammlung, Austausch & letzte Fragen zur Klausur
12:00 – 14:00 Uhr	**Klausurzeit (90 Minuten + Puffer)**
14:00 – 17:00 Uhr	**Spielpräsentationen & Seminarabschluss** Gegenseitiges Spielen & kurzes Vorstellen der Konzepte informeller Ausklang & Feedbackrunde

Abbildung 2: Seminarablauf

Im ersten Durchgang reflektierten 20 Studierende (von insgesamt geplanten N = 50) die Potenziale und Grenzen aktueller KI-Syste-

me, um diese methodisch-didaktisch sinnvoll in Lehr-Lernprozesse zu integrieren. Das heißt, KI kommt dort zum Einsatz, wo sie pädagogische Ziele (insbesondere Kindgerechtigkeit, Motivation und Flow sowie verantwortungsbewusste Nutzung) fördert, nicht um ihrer selbst willen. Den ökonomisierenden Mehrwert-Begriff lehnen wir in diesem Zusammenhang ganz bewusst ab, weil er unserer Ansicht nach Lernen auf Effizienz- und Outputgewinne verkürzt, normative Fragen (Ethik, Teilhabe, Gerechtigkeit) ausblendet und die Bedeutung pädagogischer Beziehungsgestaltung unterbelichtet. Zudem verkürzt er Medien auf bloße Werkzeuge und stabilisiert bewahrpädagogische Vergleichslogiken (vgl. Krommer 2018); unter Bedingungen einer Kultur der Partizipation (vgl. Autenrieth/Nickel 2024) verfehlt er damit den Kern pädagogischer Urteilsbildung. Statt eines „Mehrwerts“ messen wir den KI-Einsatz stattdessen an seiner Bildungsbedeutsamkeit, Teilhabe, Reflexivität und Transparenz – und daran, dass digitale Medien nicht alte Ziele nur schneller oder effizienter erreichen, sondern die Zieldimensionen von Unterricht erweitern (z. B. durch kollaboratives, vernetztes und aktives Lernen). Technik ist damit kein Selbstzweck. Vielmehr bleibt pädagogische Qualität als Maßstab, wodurch der „Primat des Pädagogischen“ (vgl. Thumel/Kammerl/Irion 2020) ganz automatisch zum leitenden Selbstverständnis der Gestaltung von Lehr-Lernprozessen wird.

So diskutierten im weiteren Verlauf des sozialkonstruktivistisch angelegten Lehr-Lernprozesses die Studierenden gemeinsam mit Lehrenden und externen Expert:innen aus Bildung und Kultur

zentrale Fragestellungen rund um den KI-Einsatz. Dabei rückte nicht nur das *Wie* des Lehrens und Lernens in den Blick, sondern ebenso das *Was*: Welche Inhalte, Zielkompetenzen und Leitbilder werden vor dem Hintergrund der durch KI beschleunigten gesellschaftlichen Transformation relevant? Zugespitzt stellte sich die Frage, inwiefern die zunehmende Integration von KI ein neues Verständnis von Mündigkeit und Bildung erfordert. Mit anderen Worten: Müssen Lernende heute anders befähigt werden, um in einer von Algorithmen geprägten Welt selbstbestimmt und verantwortungsvoll handeln zu können? Parallel dazu wurden (medien-)ethische und gesellschaftliche Rahmenbedingungen als Leitplanken für einen verantwortungsvollen Einsatz von KI im Unterricht herausgearbeitet (z. B. Transparenz, Datenschutz, Teilhabe und Gerechtigkeit).

Im Rahmen der Diskussion kristallisierten sich mehrere übergreifende Ergebnisse heraus (vgl. Abb. 2):

1. Die Rolle der Lehrkraft wurde zunehmend als Lernbegleiter:in und Coach verstanden, deren zentrale Aufgabe darin besteht, medienpädagogische und ethische Orientierungshilfen zu geben, zugunsten einer Kultur der Partizipation (Autenrieth/Nickel 2024). Statt Wissensvermittlung sollen Lehrende Möglichkeitsräume gestalten, in denen Schüler:innen eigenverantwortlich KI-gestützte Anwendungen erproben, hinterfragen und gemeinsam mit Forscher:innen und Entwickler:innen weiterentwickeln können. Die Lernenden wurden in diesem Zusammenhang nicht als passive Konsument:innen verstanden, sondern als aktive Co-Gestalter:innen, d. h. sie bringen ihre Fragen und Bedürfnisse in die Gestaltung digitaler Lernarrangements ein und entwickeln so Kompetenzen jenseits rein technischer Fertigkeiten.

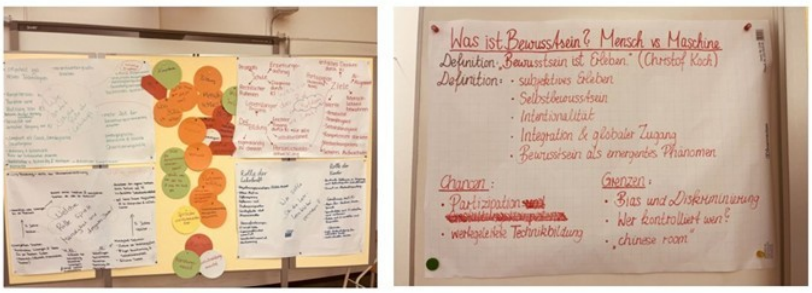


Abbildung 3: Diskussionsergebnisse der Studierenden

2. In der Auseinandersetzung mit der Mensch-Maschine-Relation zeigte sich, dass viele Studierende trotz intensiver theoretischer Beschäftigung mit posthumanistischen und systemischen Ansätzen weiterhin deutliche Unterschiede zwischen menschlichen und maschinellen Akteuren betonten. Besonders die Zuschreibung von Selbstbewusstsein, Intentionalität und emotionaler Intelligenz wurde klar dem Menschen vorbehalten, während KI-Systeme als rein datenverarbeitende Algorithmen eingeordnet wurden. Diese Tendenz lässt sich vor dem Hintergrund der eingangs diskutierten narzisstischen Kränkungen der Menschheit (vgl. Darwin, Freud) interpretieren. Die Vorstellung, dass Maschinen menschliche Fähigkeiten entwickeln könnten, widerspricht grundlegenden Selbstbildern und Erfahrungen. Daraus ergibt sich die Notwendigkeit, transformative Bildungsprozesse (Autenrieth 2025b) nicht als kurzfristiges Ziel eines Seminars zu begreifen, sondern als längerfristige Herausforderung, die Zeit, Irritation und wiederholte Auseinandersetzung erfordert.
3. Mündigkeit und reflektiertes Urteilen sahen die Teilnehmenden als eine Kompetenz, die in einer von KI geprägten Welt unverzichtbar ist. Vor diesem Hintergrund formulierten die Studierenden ein erweitertes Bildungsverständnis, bei dem Bildung als lebenslanger, selbstbestimmter und aktiver Prozess definiert wurde, der nicht nur Faktenwissen, sondern vor allem kritisches Denken, Kreativität, kollaborative Arbeitsweisen und kommunikative Fähigkeiten fördert. Mit dieser Perspektive als

Ausgangsbasis wurde Schule die Rolle einer Gestaltungsplattform zugesprochen, die Lehrende und Lernende gleichermaßen befähigt, sich in einer digitalisierten und durch KI geprägten Gesellschaft souverän und verantwortungsbewusst zu bewegen.

5.1 Die Spielentwicklung als *AI-Alignment*-Labor

Vor dem Hintergrund der theoretischen Auseinandersetzung und begrifflichen Grundlegung entwarfen die Studierenden im nächsten Schritt eigenständig KI-gestützte Spiele, die speziell für den Sachunterricht der Grundschule konzipiert wurden. Diese Aufgabe wurde ausgewählt, um einen niedrigschwelligen und für Lehramtsstudierende kontextuell verstehbaren Zugang zu komplexen Alignment-Problemen zu ermöglichen. Die Studierenden mussten KI-Systeme so konfigurieren, dass sie

- kindgerecht und sicher interagieren,
- pädagogisch sinnvolle Unterstützung bieten, ohne zu bevormunden,
- die richtige Balance zwischen Information und Zurückhaltung finden,
- ethische Implikationen diskutieren, und
- Kreativität fördern.

Diese praktische Herausforderung machte die abstrakten Konzepte des *AI Alignment* konkret erfahrbar. Die Studierenden erlebten unmittelbar, was es bedeutet, ein System zu gestalten, das autonom agiert und dennoch menschlichen Werten folgt. Dabei überlegten sie, wie interaktive Spielmechanismen und adaptive Dialogsysteme kindgerecht zusammenwirken können, um fachli-

che Inhalte zu vertiefen und zugleich Medienkompetenz bei den Schüler:innen zu fördern. Die Teams achteten darauf, den Einsatz von KI so zu gestalten, dass Lernende aktiv agieren, differenzierte Rückmeldungen erhalten und dabei ihre Selbststeuerung sowie kollaboratives Arbeiten weiterentwickeln. Dadurch sollte nicht nur der Erwerb fachlichen Wissens, sondern auch die reflexive Auseinandersetzung mit digitalen Technologien in einem intrinsisch motivierenden Spielkontext ermöglicht werden.

Durch iteratives Testen und Anpassen entwickelten sie nicht nur technische Fertigkeiten im Prompt Engineering und der Systemkonfiguration, sondern auch ein tiefes Verständnis für die Komplexität der Mensch-Maschine-Interaktion. Die Studierenden verstanden durch ihre praktische Arbeit, dass *AI Alignment* kein rein technisches Problem ist, sondern eine kulturelle und pädagogische Herausforderung. Sie erlebten, wie die Konfiguration eines scheinbar harmlosen Lernspiels fundamentale Fragen über Autonomie, Kontrolle, Vertrauen und Verantwortung aufwirft.

5.2 Learning by Doing und Reflecting

Der methodische Ansatz folgte Deweys Prinzip des erfahrungsbasierten Lernens. Die Studierenden durchliefen einen zyklischen Prozess von Handlung, Beobachtung und Reflexion:

- *Experimentelle Phase*: Eigenständiges Erkunden verschiedener KI-Tools und deren Möglichkeiten.
- *Konzeptionsphase*: Entwicklung von Game-based Learning (vgl. Autenrieth/Nickel 2024) Szenarien unter Berücksichtigung von Alignment-Prinzipien.

- *Implementierungsphase*: Praktische Umsetzung mit kontinuierlicher Anpassung.
- Reflexionsphase: Kritische Auseinandersetzung mit den Erfahrungen in Podcasts und Diskussionen.

Dieser Prozess entspricht einer reflexiv-praktischen Medienarbeit (Schorb 1995), in der Studierende authentische Erfahrungen machten und kommunikative Kompetenz anbahnen, indem sie sich die Medien aktiv aneignen (vgl. Abb. 3). Besonders wertvoll war dabei die Erkenntnis, dass die Programmierung von KI-Systemen über natürliche Sprache eine neue Form des Codings darstellt, die Zugang zu komplexen technischen Systemen demokratisiert und somit einen medienpädagogisch geprägten Zugang zu allen Handlungsdimensionen des Computational Thinking (Abstraktion, Mustererkennung, Problemzerlegung und Algorithmisierung, Wing 2006) ermöglichte.

Abbildung 4: Prozessreflexion der Seminarteilnehmenden

Ergänzend zur praktischen Spielentwicklung erhielten die Studierenden einen Expert:innenvortrag zum Thema *Philosophieren mit Kindern*, in dem Techniken vorgestellt wurden, um Grundschulkinder an abstrakte Fragen rund um die Mensch-Maschine-Relation

heranzuführen. Basierend auf Worleys (2011) praktischen Zugängen gewannen sie weitere Anregungen, wie zum Beispiel offene Fragestellungen und dialogische Moderationsformen, um Kindern Freiräume für eigene Perspektiven zu bieten. Auf dieser Basis entwarfen die Teams jeweils eine in sich geschlossene Storyline, in der verschiedene KI-Tools (u. a. adaptive Sprachassistenten, visuelle Erkennungssysteme) verknüpft wurden. So entstanden kreative, interaktive Spiele, die reflexive Zugänge zum Thema KI eröffnen und Kinder dazu einladen, eigene Gedanken zur Zukunft von Mensch und Maschine zu formulieren.

5.3 Vom individuellen Lernen zur gesellschaftlichen Verantwortung

Den Abschluss der dreitägigen Veranstaltungsreihe bildete eine gemeinsame Spiel- und Reflexionsphase, in der die Studierenden ihre Erfahrungen medial weiterverarbeiteten. In ihren jeweiligen Teams produzierten sie jeweils einen 20-minütigen, dialogischen Podcast zum Thema *Mensch–KI–Bildung*.

Aufbauend auf dem im Seminar erarbeiteten Literaturfundament entwickelten sie jenseits einer oberflächlichen Pro-und-Contra-Debatte eigene Standpunkte, diskutierten bildungsphilosophische und gesellschaftliche Implikationen und vertieften auf diese Weise die Fragen des Seminars. Inhaltlich setzten sich die Gruppen unter anderem mit philosophischen Grundfragen auseinander („Woran würden wir merken, ob eine KI wirklich ‚versteht' oder nur sehr gut imitiert?", „Ist ein Kind, das nur auswendig lernt, intelligenter als eine KI, die kreativ antwortet?"). Darüber hinaus reflektierten sie gesellschaftliche Gestaltungsfragen wie die Mitbe-

stimmung von Lehrkräften und Schüler:innen bei der Entwicklung von Lern-KI und erörterten Dilemmata („Ein Kind fragt: ‚Warum soll ich noch lernen, wenn KI alles kann?' – Wie würden Sie antworten?"). Zukunftsvisionen und das Konzept einer Co-Evolution von Mensch und Maschine lieferten mögliche Szenarien für eine kollaborativ und partizipativ gestaltete und nach Außen geöffnete Schule.

Als roter Faden durch alle Podcasts zog sich abschließend der Impuls, über die Frage „Wie nutze ich KI im Unterricht?" hinauszugehen und stattdessen zu überlegen: „Wie gestalten wir als Gesellschaft das Verhältnis zwischen Mensch und KI und welche Rolle will ich als Lehrkraft dabei übernehmen?". Dieses mediale Abschlussprodukt erlaubte nicht nur eine differenzierte Reflexion der Seminarinhalte, sondern bildete zugleich eine authentische Transferleistung in Richtung zukünftiger Lehrpraxis.

5.4 Ausblick: Zweite Praxis- und Datenerhebungs-Phase im Wintersemester 2025/26

In der zweiten Phase der Seminarreihe im Wintersemester 2025/26 steht die Handlungskomponente medienbezogener Lehrkompetenz im Fokus. Diese Phase wird die entwickelten Ansätze in reale Schulkontexte übertragen. Hierzu arbeiten die Studierenden in Schulkooperationen und Kleingruppen an der Entwicklung und Erprobung kindgerechter, KI-gestützter Unterrichtsszenarien. Zunächst konfigurieren sie digitale Sprach- und Interaktionstools und planen deren Einsatz im Unterricht detailliert. Anschließend setzen sie ihre Konzepte an zwei Praxistagen in ko-

operierenden Grundschulen um: Am ersten Tag lernen sie die Kinder kennen und besprechen die Szenarien, am zweiten Tag führen sie 90-minütige Unterrichtseinheiten durch.

Diese Praxisphase verspricht weitere Erkenntnisse darüber, wie Kinder mit KI-Systemen interagieren und welche pädagogischen Interventionen nötig sind, um eine reflektierte und selbstbestimmte Nutzung zu fördern. Begleitend zur Unterrichtserprobung erfolgt eine videobasierte ethnografische Feldforschung. Die Studierenden dokumentieren die Unterrichtsphasen mit GoPro-Kameras und ergänzen diese Aufnahmen durch teilnehmende Beobachtungen sowie informelle Gespräche mit Lehrpersonen und Schüler:innen. Auf dieser Grundlage diagnostizieren und reflektieren sie die Lehr-Lern-Prozesse.

Parallel vertiefen sie ihre professionelle Selbstentwicklung durch wöchentliche Reflexionsbriefe (doc.post-Methode) mit Briefpartner:innen aus der Seminargruppe, sodass durch deren Analyse Aussagen über die Reflexionsbreite und -tiefe der Studierenden gemacht werden können (Autenrieth 2025b). Der sogenannte „didaktische Doppeldecker“ bildet den roten Faden dieser Phase: Die Seminar-Lernkultur, gekennzeichnet durch Kollaboration, Autonomie, Feedback und Experimentieren, wird sowohl theoretisch behandelt als auch praktisch erprobt. Die Studierenden erfahren selbst, wie sie später agieren sollen. Diese Verschränkung von Theorie und Praxis, von Erleben und Reflektieren, macht die Komplexität der KI-Thematik handhabbar, ohne sie zu simplifizieren.

Learning by Teaching mit digitalen Medien verdeutlicht den integrativen Charakter des selbstgesteuerten Lernens und Lehrens.

Das Seminar endet mit einer öffentlichen, hybriden Lehr- und Lerntagung. Studierende präsentieren ihre digitalen Poster, diskutieren in einer Podiumsrunde mit Vertreter:innen aus Wissenschaft, Schulpraxis und Bildungspolitik und hören Impulsvorträge externer Fachexpert:innen. Diese Veranstaltung dient als Prüfungsformat, Transferplattform und Grundlage für nachhaltige Kooperationen zwischen Universität und Praxis.

6. Fazit

Der Beitrag verfolgte das Ziel die aktuellen Herausforderungen der Gestaltung der Mensch-KI-Beziehung für den Bildungskontexten nicht nur zu beschreiben, sondern praktisch zugänglich und bearbeitbar zu machen. Das KI-Meta-Modell lieferte hierzu ein Übersetzungsinstrument, das zwischen abstrakter erkenntnistheoretischer Debatte und konkreter pädagogischer Handlung vermittelt. Die multiperspektivische Anlage des Modells ermöglicht es, die Vielschichtigkeit des Alignment-Problems nicht im Analytischen zu belassen, sondern gezielt in Richtung partizipativer, gestaltungsorientierter Bildungsprozesse zu öffnen. Es sollte gezeigt werden, dass handlungsorientierte Medienpädagogik mehr ist als technologische Befähigung. Sie schafft einen Reflexions- und Gestaltungsraum, in dem Lehrende und Lernende aktiv an der Aushandlung und Ausrichtung von KI in Bildungsprozessen mitwirken können.

Handlungsorientierung erweist sich dabei als zentrale Brücke. Sie verbindet theoretische Reflexionen über die Bedingungen von Mündigkeit, Autonomie und Verantwortung mit konkreten Erfahrungs- und Gestaltungsräumen im pädagogischen Alltag. Die praktische Umsetzung im hochschuldidaktischen Pilotprojekt zeigt, dass sich komplexe Fragen des *AI Alignment* nicht allein durch abstraktes Nachdenken lösen lassen. Erst durch aktives Erproben, gemeinsame Entwicklung und kritische Reflexion werden sie für alle Beteiligten greifbar, verhandelbar und gestaltbar. Diese Erfahrungsräume und die dabei angestoßene Vorstellungsbildung für Lehrende und Lernende tragen entscheidend dazu bei, dass die Gestaltung von KI und von KI-geprägten Gesellschaften nicht allein in den Händen von Technologiekonzernen verbleibt. Vielmehr wird deutlich, dass es sich um eine genuin gesellschaftliche Aufgabe handelt, die breite Beteiligung und kritische Reflexion erfordert. Damit diese Mitgestaltung gelingen kann, braucht es jedoch nicht nur entsprechende pädagogische Zugänge, sondern vor allem auch ein tiefes Verständnis für die Tragweite des Themas in der gesamten Gesellschaft. Erst wenn die Bedeutung von KI umfassend erkannt wird und gesellschaftliche Perspektiven wirkungsvoll in politische und wirtschaftliche Entscheidungsprozesse einfließen, kann eine nachhaltige, demokratische und verantwortungsbewusste Entwicklung im Umgang mit künstlicher Intelligenz gelingen.

Anmerkungen

1 Der Beitrag entstand in gemeinsamer Autor:innenschaft mit arbeitsteiliger Struktur: Die Einleitung und das Fazit wurden gemeinsam verfasst. Die Kapitel 1 und 2 stammen von Daniel Autenrieth, die Kapitel 3, 4 und 5 von Stefanie Nickel.

2 In diesen entwicklungspsychologischen Aufgaben geht es darum zu prüfen, ob ein Akteur versteht, dass eine andere Person eine falsche Überzeugung haben kann, die sich von der Realität unterscheidet.

3 *Theory of Mind* bezeichnet die Fähigkeit zu verstehen, dass andere Menschen eigene Gedanken, Gefühle und Überzeugungen haben, die sich von den eigenen unterscheiden.

4 Oprerationalisierungen hierzu ergeben sich aus zahlreichen Benchmarks. Eine Zusammenstellung findest sich bei Bengio et al. 2025: 49.

5 ARC (*Abstraction and Reasoning Corpus*) ist ein Benchmark-Test, der die Fähigkeit von KI-Systemen prüft, abstrakte Problemstellungen unter minimalen Vorgaben zu lösen. HLE (*Humanity's Last Exam*) ist eine von über 1.000 Wissenschaftler:innen entwickelte Testsammlung aus 3.000 Aufgaben, die interdisziplinäre Transfer- und Syntheseleistungen erfordern. Beide Tests wurden speziell entwickelt, um eine differenzierte Bewertung von KI-Fähigkeiten jenseits reiner Wissensreplikation zu ermöglichen.

6 Zwei Beispiele: Erstens: *AlphaFold* (Jumper et al. 2021): Ein KI-System, dass aus der Sequenz von Aminosäuren die dreidimensionale Struktur eines Proteins mit hoher Genauigkeit vorhersagt. Dies ermöglicht wichtige Fortschritte in Biochemie, Medizin und Wirkstoffforschung. Zweitens: *AlphaEvolve* (Novikov et al. 2025): Ein KI-System, das viele unterschiedliche Lösungen für ein Problem erzeugt, die besten auswählt und auf dieser Basis neue, verbesserte Varianten entwickelt. Es hat bereits neuartige mathematische Verfahren entwickelt und Optimierungen für technische und wissenschaftliche Anwendungen auch mit Blick auf den nachhaltigen Betrieb von Rechenzentren geschaffen.

7 Der *Posthumanismus* durchbricht den menschlichen Exzeptionalismus. Ähnlich wie der Mensch seine vermeintliche Sonderstellung gegenüber Tieren durch Darwins Evolutionstheorie verlor, stellt der Posthumanismus die anthropozentrische Hierarchie in Frage, die Menschen kategorisch über andere Entitäten, seien es Tiere, Pflanzen oder Maschinen, stellt (vgl. Coeckelbergh 2020).

8 Die Metapher des *kartesianischen Theaters* beschreibt die Vorstellung, dass ein Subjekt über einen direkten, unfehlbaren Zugang zu seinen eigenen mentalen Zuständen verfügt, als gäbe es im Geist ein zentrales „Ich", das alle Gedanken, Wahrnehmungen und Gefühle bewusst beobachtet. Dennett lehnt dieses Modell aus mehreren Gründen ab:

9 Diese Klarheit basiert darauf, dass KI-Systeme nicht den biologischen und kognitiven Begrenzungen unterliegen, die menschliches Denken charakterisieren. Während das menschliche Gehirn unter evolutionären Zwängen für eine Umwelt optimiert wurde, die bedeutend einfacher war als unsere heutige komplexe, vernetzte Welt, können KI-Systeme gezielt für das Erfassen dieser Komplexität entwickelt werden. Sie verarbeiten exponenziell größere Datenmengen, erkennen übergreifende Muster und können langfristige, systemische Zusammenhänge erfassen, die für Menschen aufgrund kognitiver Verzerrungen und begrenzter Verarbeitungskapazität oft unsichtbar bleiben.

10 In wissenschaftlichen Diskursen wird Liebe als ein vielschichtiges Phänomen betrachtet, das von biologischen Bindungsmechanismen, über psychologische Modelle (z. B. Sternbergs Dreiecksheorie der Liebe, 1986) bis hin zu soziologischen und philosophischen Konzepten intersubjektiver Anerkennung reicht (Honneth 1994; Luhmann 2018). So unterscheidet Sternberg (1986) zwischen Intimität, Leidenschaft und Verpflichtung als Grunddimensionen der Liebe, während Honneth (1994) Liebe als grundlegende Form wechselseitiger Anerkennung und Voraussetzung für gelingende Sozialität beschreibt.

11 Mit „maschinischen Interessen“ ist bislang kein intentionaler Begriff im anthropologischen Sinn gemeint. In großen Sprachmodellen lassen sich aber bereits proto-intentionale Strukturen zeigen, die Handlungspfade konsistent steuern (z. B. Planungsmechanismen bei Reimen oder mehrschrittiges Schlussfolgern). Aus funktionalistischer Perspektive können solche Strukturen als Vorformen von Intentionalität gedeutet werden. Künftig könnten sich aus dieser Perspektive auch weitergehende intentionale Fähigkeiten ausbilden, sofern die Systeme hinreichend komplex werden (Lindsey et al. 2025)

12 Unter einem *Lock-in-Effekt* versteht MacAskill die Gefahr, dass insbesondere fortgeschrittene KI-Systeme bestimmte gesellschaftliche Werte, Machtstrukturen oder Institutionen festschreiben und dadurch zukünftige Veränderung massiv erschweren oder unmöglich machen. Unbedachtes Alignment, das heute vorgenommen und technisch implementiert wird, könnte durch die globale Verbreitung und Selbstoptimierung von KI dauerhaft verstetigt werden. Dies könnte bedeuten, dass nicht nur die aktuelle Generation, sondern auch zukünftige Gesellschaften an diese Werte und Systeme gebunden bleiben, selbst wenn sich die moralischen Vorstellungen oder gesellschaftlichen Bedürfnisse ändern, was die Entwicklung demokratischer Selbstbestimmung und den ethischen Fortschritt erheblich beeinträchtigen kann.

13 Bezogen auf die praktische Anwendung lässt sich das Projekt *Doing-KI* (vgl. Autenrieth/Nickel 2023) anführen. Das Hochschulprojekt startete im Sommersemester 2023 und erforsche die Zukunft der Bildung im Kontext von KI unter Berücksichtigung ökologischer, ökonomischer und sozialer Nachhaltigkeitsaspekte. Die Umsetzung erfolgte gemäß dem Modell: Ziel war es, konkrete Strategien zur Gestaltung der Zukunft ko-konstruktiv zu generieren. Dabei wurde auf einen interdisziplinären Ansatz unter Einbezug verschiedener Expert:innen aus Bereichen wie Medienpädagogik, Informatik, Philosophie, Neurobiologie und Theologie zurückgegriffen, um ein ganzheitliches Verständnis der Rolle von KI in der Bildung zu entwickeln. Im Fokus

des Lehr-Lern-Prozesses standen die Studierenden, die in Beziehung zu anderen Lernenden und Lehrenden gegangen sind und auf diese Weise Resonanz erfahren haben. Das Besondere an der interpersonalen Wechselbeziehung waren dabei die sozialen Praktiken zur Kompetenzanbahnung innerhalb des strukturgebenden Bildungs- und Erfahrungsraums.

Literatur

Amodei, Dario (2025): The Urgency of Interpretability, online unter: https://www.darioamodei.com/post/the-urgency-of-interpretability (letzter Zugriff: 01.09.2025).

Autenrieth, Daniel (2025a im Erscheinen): Konstruktivistische Lerntheorien als Ausgangspunkt für das Alignment von KI-Systemen im Bildungskontext: Medienpädagogische Perspektiven für post-AGI Gesellschaftsszenarien, in: Ehlers, Ulf-Daniel/Reimer, Ricarda T. D. (Hg.): Medienpädagogische Erfahrungsräume zwischen Tradition und Innovation. Organisationsstrukturen und Lehren – ethische Diskurse ermöglichen, Weinheim: Beltz Juventa.

Autenrieth, Daniel (2025b): Transformative Bildungsprozesse und Partizipation. Eine empirische Untersuchung im Kontext von Künstlicher Intelligenz und der Lehrkräftebildung, München: kopaed.

Autenrieth, Daniel/Nickel, Stefanie (2022): KuDiKuPa – Kultur der Digitalität = Kultur der Partizipation?!: Verschränkung von Theorie und Praxis in partizipativ angelegter Hochschullehre durch Gaming und Game Design – ein Praxisbeispiel, in: MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung, Jahrbuch Medienpädagogik 18 (Ästhetik – Digitalität – Macht), 237–265, online unter: https://www.medienpaed.com/article/view/1309 (letzter Zugriff: 01.09.2025).

Autenrieth, Daniel/Nickel, Stefanie (2023): Das KI-Meta-Modell. Handlungsleitende Strukturen für den Umgang mit künstlicher Intelligenz im Bildungsbereich, in: technik-education (tedu). Fachzeitschrift für Unterrichtspraxis und Unterrichtsforschung im allgemeinbildenden Technikunterricht, 3 (2), 14–20.

Autenrieth, Daniel/Nickel, Stefanie (2024): Kultur der Digitalität = Kultur der Partizipation?! Erforschung und Gestaltung einer partizipativen Bildungslandschaft, 1. Auflage, München: kopaed.

Baacke, Dieter (1973): Kommunikation und Kompetenz: Grundlegung einer Didaktik der Kommunikation und ihrer Medien, München: Juventa.

Baacke, Dieter (1996): Medienkompetenz: Begrifflichkeit und sozialer Wandel, in: Rein, Antje (Hg.): Medienkompetenz als Schlüsselbegriff, Bad Heilbrunn: Klinkhardt, 112–124.

Bach, Joscha (2025): This is the dawn of machine consciousness [YouTube], The Institute of Art and Ideas, online unter: https://www.youtube.com/watch?v=Y1QOf6HEbHQ (letzter Zugriff: 01.09.2025).

Bengio, Yoshua/Mindermann, Sören/Privitera, Daniel et al. (2025): International AI Safety Report, DSIT 2025/001, online unter: https://www.gov.uk/government/publications/international-ai-safety-report-2025 (letzter Zugriff: 01.09.2025).

Beranek, Angelika/Engelhardt, Elisabeth/Rösch, Eike (2024): Medienpädagogik und KI, Themenheft der merz – Medien und Erziehung, in: merz – Medien und Erziehung, Bd. 68. Nr. 3.

Bock, Annekatrin/Franken, Lina/Rau, Franco/Kühn, Jessica/Fehr, Ada (2025): CfP: Beyond Prompting ?! Sozio-technische Systeme, KI und Medienbildung in der Post-Digitalität, online unter: https://www.medienpaed.com/public/journals/1/cfps/CfP_Beyond_Prompting.pdf (letzter Zugriff: 01.09.2025).

Brinda, Torsten/Brüggen, Niels/Diethelm, Ira et al. (2020): Frankfurt-Dreieck zur Bildung in der digital vernetzten Welt. Ein interdisziplinäres Modell, online unter: https://doi.org/10.25656/01:22117 (letzter Zugriff: 01.09.2025).

Bubeck, Sébastien/Chandrasekaran, Varun/Eldan, Ronen et al. (2023): Sparks of Artificial General Intelligence: Early experiments with GPT-4, online unter: https://arxiv.org/abs/2303.12712 (letzter Zugriff: 01.09.2025).

Chollet, François (2024): OpenAI O3 Breakthrough High Score on ARC-AGI-Pub, ARC Prize, online unter: https://arcprize.org/blog/oai-o3-pub-breakthrough (letzter Zugriff: 01.09.2025).

Coeckelbergh, Mark (2020): AI Ethics, Cambridge/Massachusetts: The MIT press.

Dander, Valentin/Grünberger, Nina/Niesyto, Horst/Pohlmann, Horst (Hg.) (2024): Bildung und digitaler Kapitalismus, München: kopaed.

Deci, Edward L./Ryan, Richard M. (2017): Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. https://doi.org/10.25656/01:11173.

Dennett, Daniel C. (1991): Consciousness Explained, Boston: Little, Brown and Co.

Dewey, John (1964): Demokratie und Erziehung, Weinheim: Beltz.

Freud, Sigmund (1947): Gesammelte Werke. Chronologisch geordnet. Band 12: Werke aus den Jahren 1917–1920, London: Imago Publishing Co. Ltd.

Fromm, Erich (1956): Die Kunst des Liebens, Frankfurt am Main: Suhrkamp.

Fuchs, Max (2023): Bildung und Lebensführung: Überlegungen zu einem zeitgemäßen Bildungsbegriff, München: kopaed.

Future of Life Institute (2023): Pause Giant AI Experiments: An Open Letter, online unter: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (letzter Zugriff: 01.09.2025).

Gapski, Harald (2021): Künstliche Intelligenz (KI) und kritische Medienbildung, Projekt Digitales Deutschland, JFF – Institut für Medienpädagogik in Forschung und Praxis, online unter: https://digi-d.jff.de/ki-expertisen/kuenstliche-intelligenz-und-kritische-medienbildung-harald-gapski/ (letzter Zugriff: 01.09.2025).

Giddens, Anthony (1988): Die Konstitution der Gesellschaft. Grundzüge einer Theorie der Strukturierung, Frankfurt am Main: Campus.

Haan, Gerhard de (2008): Gestaltungskompetenz als Kompetenzkonzept der Bildung für nachhaltige Entwicklung, in: Bormann, Inka/Haan, Gerhard de (Hg.): Kompetenzen der Bildung für nachhaltige Entwicklung, Wiesbaden: VS Verlag für Sozialwissenschaften, 23–43.

Haraway, Donna J. (1985): A Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the 1980s, in: Socialist Review 80.

Haraway, Donna J. (2003): The Companion Species Manifesto: Dogs, People, and Significant Otherness, Chicago: Prickly Paradigm Press.

Holzinger, Andreas (2000): Basiswissen Multimedia. Band 2: Lernen, Würzburg: Vogel.

Holzinger, Andreas (2018): Explainable AI (ex-AI), in: Informatik-Spektrum 41 (2), 138–143.

Honneth, Axel (1994): Kampf um Anerkennung: Zur moralischen Grammatik sozialer Konflikte, Frankfurt am Main: Suhrkamp.

Hug, Theo (2023): Education and 4e Cognition: Some Challenges at the Crossroads of Learning and Bildung, in: Constructivist Foundations, 18 (2), 158–168.

Hug, Theo/Missomelius, Petra/Ortner, Heike (2024): Künstliche Intelligenz im Diskurs: Interdisziplinäre Perspektiven zur Gegenwart und Zukunft von KI-Anwendungen, in: Preprint, Innsbruck university press.

Jakubik, Johannes/Vössing, Michael/Kühl, Niklas/Walk, Jannis/Satzger, Gerhard (2024): Data-Centric Artificial Intelligence, online unter: https://link.springer.com/article/10.1007/s12599-024-00857-8 (letzter Zugriff: 01.09.2025).

Jumper, John/Evans, Richard/Pritzel, Alexander et al. (2021): Highly Accurate Protein Structure Prediction with AlphaFold, in: Nature, 596 (7873), 583–589.

Kokemohr, Rainer (2007): Bildung als Welt- und Selbstentwurf im Anspruch des Fremden. Eine theoretisch-empirische Annäherung an eine Bildungsprozesstheorie, in: Koller, Hans-Christoph/Marotzki, Winfried/Sanders, Olaf (Hg.): Theorie Bilden, 1. Aufl., Bielefeld: transcript, 13–68.

Koller, Hans-Christoph (2012): Bildung anders denken: Einführung in die Theorie transformatorischer Bildungsprozesse, Stuttgart: W. Kohlhammer.

Kosinski, Michal (2024): Evaluating Large Language Models in Theory of Mind Tasks, in: Proceedings of the National Academy of Sciences, 121 (45).

Krommer, Axel (2018): Wider den Mehrwert! Oder: Argumente gegen einen überflüssigen Begriff, Bildung unter Bedingungen der Digitalität, online unter: https://axelkrommer.com/2018/09/05/wider-den-mehrwert-oder-argumente-gegen-einen-ueberfluessigen-begriff/ (letzter Zugriff: 01.09.2025).

Krommer, Axel (2024): EdTech aus dem letzten Jahrtausend: kybernetische Pädagogik und künstliche Intelligenz, online unter: https://axelkrommer.com/2024/11/04/edtech-aus-dem-letzten-jahrtausend-kybernetische-paedagogik-und-kuenstliche-intelligenz/ (letzter Zugriff: 01.09.2025).

Li, Yuxiao/Michaud, Eric J./Baek, David D./Engels, Joshua/Sun, Xiaoqing/Tegmark, Max (2024): The Geometry of Concepts: Sparse Autoencoder Feature Structure, online unter: https://arxiv.org/abs/2410.19750 (letzter Zugriff: 01.09.2025).

Lindsey, Jack/Gurnee, Wes/Ameisen, Emmanuel et al. (2025): On the Biology of a Large Language Model. Transformer Circuits Thread, März 2025, online unter: https://transformer-circuits.pub/2025/attribution-graphs/biology.html (letzter Zugriff: 01.09.2025).

Luhmann, Niklas (2018): Soziale Systeme: Grundriß einer allgemeinen Theorie. 17. Auflage. Theorie der Gesellschaft / Niklas Luhmann 666, Frankfurt am Main: Suhrkamp.

MacAskill, William (2022): What We Owe the Future, London: Oneworld.

Maturana, Humberto R./Varela, Francisco J. (2011): Der Baum der Erkenntnis: Die biologischen Wurzeln menschlichen Erkennens, 10. Auflage, Frankfurt am Main: Fischer.

Metzinger, Thomas (2003): Being No One: The Self-Model Theory of Subjectivity. Cambridge, MA: MIT Press.

Metzinger, Thomas (2023): Der Ego Tunnel: eine neue Philosophie des Selbst: von der Hirnforschung zur Bewusstseinsethik. 8. Auflage. Übersetzt von Thorsten Schmidt, Piper 30533, München: Piper.

Morris, Meredith Ringel/Sohl-Dickstein, Jascha/Fiedel, Noah et al. (2024): Levels of AGI for Operationalizing Progress on the Path to AGI, online unter: https://arxiv.org/abs/2311.02462 (letzter Zugriff: 01.09.2025).

Nakkiran, Preetum/Kaplun, Gal/Bansal, Yamini/Yang, Tristan/Barak, Boaz/Sutskever, Ilya (2019): Deep Double Descent: Where Bigger Models and More Data Hurt, online unter: https://arxiv.org/abs/1912.02292 (letzter Zugriff: 01.09.2025).

Novikov, Alexander/Vũ, Ngân/Eisenberger, Marvin et al. (2025): AlphaEvolve: A coding agent for scientific and algorithmic discovery, White paper, Google DeepMind, online unter: https://arxiv.org/abs/2506.13131 (letzter Zugriff: 01.09.2025).

OpenAI (2025): Eine Einführung in Deep Research, OpenAI Blog, online unter: https://openai.com/index/introducing-deep-research/ (letzter Zugriff: 01.09.2025).

Phan, Long/Gatti, Alice/Han, Ziwen et al. (2025): Humanity's Last Exam, online unter: https://arxiv.org/abs/2501.14249 (letzter Zugriff: 01.09.2025).

Radford, Alec/Jozefowicz, Rafal/Sutskever, Ilya (2017): Learning to Generate Reviews and Discovering Sentiment, Preprint, 6. April 2017. https://doi.org/10.48550/arXiv.1704.01444.

Rosa, Hartmut (2016): Resonanz: eine Soziologie der Weltbeziehung, 3. Aufl., Berlin: Suhrkamp.

Russell, Stuart J./Norvig, Peter (2021): Artificial Intelligence: A Modern Approach. Fourth Edition, Boston: Pearson.

Schell, Fred (1989): Aktive Medienarbeit mit Jugendlichen, Opladen: Leske + Budrich.

Schorb, Bernd (1995): Reflexiv-praktische Medienaneignung. Auf dem Weg zur Medienkompetenz, in: Schorb, Bernd: Medienalltag und Handeln, Wiesbaden: VS Verlag für Sozialwissenschaften, 15–48.

Schorb, Bernd (2022): Handlungsorientierte Medienpädagogik, in: Sander, Uwe/Gross, Friederike Von/Hugger, Kai-Uwe (Hg.): Handbuch Medienpädagogik, Wiesbaden: Springer Fachmedien, 1–13.

Searle, John R. (1980): Minds, Brains, and Programs, in: Behavioral and Brain Sciences, 3 (3).

Shiller, Derek (2024): Functionalism, Integrity, and Digital Consciousness, in: Synthese, 203 (2), 47.

Siebert, Horst (2005): Pädagogischer Konstruktivismus: Lernzentrierte Pädagogik in Schule und Erwachsenenbildung, 3. Auflage, Weinheim: Beltz.

Stalder, Felix (2019): Kultur der Digitalität, Berlin: Suhrkamp.

Ständige Wissenschaftliche Kommission der Kultusministerkonferenz (SWK) (2024): Large Language Models und ihre Potenziale im Bildungssystem: Impulspapier, online unter: https://www.kmk.org/dokumentation-statistik/rechtsvorschriften-lehrplaene/uebersicht-lehrplaene.html (letzter Zugriff: 01.09.2025).

Sternberg, Robert J. (1986): A Triangular Theory of Love, in: Psychological Review 93 (2), 119–135. https://doi.org/10.1037/0033-295X.93.2.119.

Tegmark, Max (2017): Life 3.0: Being human in the age of artificial intelligence, London: Allen Lane.

Thumel, Mareike/Kammerl, Rudolf/Irion, Thomas (2020): Digitale Bildung im Grundschulalter. Grundsatzfragen zum Primat des Pädagogischen, München: kopaed. https://doi.org/10.25593/978-3-86736-543-7.

Verbeek, Peter-Paul (2011): Moralizing Technology: Understanding and Designing the Morality of Things, Chicago: University of Chicago Press.

Wing, Jeannette M. (2006): Computational Thinking, in: Communications of the ACM, 49 (3), 3.

Worley, Peter (2011): The If Machine: Philosophical Enquiry in the Classroom, London: Continuum.