Medienimpulse
ISSN 2307-3187
Jg. 63, Nr. 1, 2025
doi: 10.21243/mi-01-25-25

Menschliche Existenz, (Nicht)Nachhaltigkeit & Künstliche Intelligenz. AI-Safety und AI-Alignment als Reflexionsgröße der Medienpädagogik

Daniel Autenrieth

Jan-René Schluchter

Der Beitrag untersucht die komplexen Zusammenhänge zwischen Künstlicher Intelligenz (KI), Alignment und Nachhaltigkeit und beleuchtet deren Bedeutung für zukünftige gesellschaftliche Transformationsprozesse. Eine systematische Analyse zeigt, dass digitale Technologien, insbesondere KI, bisher oft ohne ausreichende Beachtung von Nachhaltigkeitsaspekten entwickelt und eingesetzt werden. Der Beitrag skizziert die Notwendigkeit, sowohl technologische als auch geisteswissenschaftliche Perspektiven zu integrieren, um die Potenziale von KI im Sinne nachhalti-

ger Entwicklung zu nutzen. Ein besonderer Schwerpunkt liegt auf der Bedeutung von AI-Safety und AI-Alignment, um sicherzustellen, dass KI-Systeme verantwortungsvoll gestaltet und genutzt werden. Philosophische sowie ethische Überlegungen werden herangezogen, um die gesellschaftliche Verantwortung in der Technologieentwicklung und -nutzung zu beleuchten. Abschließend wird die Rolle der Medienpädagogik im Kontext der Gestaltung von Künstlicher Intelligenz in der Perspektive von Nachhaltigkeit diskutiert: der Beitrag plädiert für eine stärkere Integration von Nachhaltigkeit in den medienpädagogischen Diskurs über KI sowie die Integration von medienpädagogischen Perspektiven in Diskurse um AI-Safety und AI-Alignment, um globale Herausforderungen wie den Klimawandel und soziale Ungleichheiten anzugehen.

This article examines the complex relationships between artificial intelligence (AI), alignment and sustainability and sheds light on their significance for future social transformation processes. A systematic analysis shows that digital technologies, especially AI, have so far often been developed and used without sufficient consideration of sustainability aspects. The article outlines the need to integrate both technological and humanities perspectives in order to utilise the potential of AI in terms of sustainable development. A particular focus is placed on the importance of AI safety and AI alignment to ensure that AI systems are designed and used responsibly. Philosophical and ethical considerations are used to shed light on social responsibility in the development and use of technology. Finally, the role of media education in the context of shaping artificial intelligence from a sustainability perspective is discussed: the article argues in favour of a stronger integration of sustainability into the media education discourse on AI as well as the integration of media education perspectives into discourses on AI safety and AI alignment in or-

der to address global challenges such as climate change and social inequalities.

1. Einleitung

Unser gegenwärtiges Verständnis von Welt – oder besser, Welten – ist durch Sezieren, Ausstellen, Beobachten, Kartieren, Sammeln, Experimentieren, Darstellen, Graben, Anbauen und Repräsentieren hervorgebracht worden. Während wissenschaftliche Errungenschaften in Mathematik, Physik, Biologie etc. breite Anerkennung finden, wird die prägende Rolle von kulturellen Weltsichten (u. a. philosophischen Grundannahmen aber auch religiösen Überzeugungen) und gesellschaftlichen Macht- und Herrschaftsverhältnissen auf diese wissenschaftlichen Erkenntnisse selten in den Fokus der einzelnen Disziplinen bzw. gesellschaftlichen Diskurse gerückt. Hier lässt sich eine Analogie zu gegenwärtigen Diskursen um KI-Technologien herstellen. Auch in diesen Diskursen gewinnen geisteswissenschaftliche Perspektiven an Bedeutung. Sie betrachten, wie unterschiedliche Sichtweisen auf die Welt im Allgemeinen und den Menschen im Besonderen zusammenwirken. Gleichzeitig machen sie auf oft verborgene gesellschaftliche Macht- und Herrschaftsverhältnisse aufmerksam. Die Diskussion um AI Safety und AI Alignment verdeutlicht jedoch, wie eng wissenschaftlich-technische Fortschritte mit gesellschaftlichen Strukturen verwoben sind. Während AI Safety sich mit der sicheren Entwicklung und Implementierung von KI-Systemen befasst, insbesondere im Hinblick auf Risiken durch Fehlfunktionen oder Missbrauch, zielt AI Alignment darauf ab, sicherzustellen, dass KI-

Systeme menschliche Werte und gesellschaftliche Ziele nicht nur nachvollziehen, sondern auch zuverlässig befolgen (Bengio et al. 2025). Diese Aspekte sind dabei nicht nur für die ethische und sicherheitspolitische Regulierung von KI von Bedeutung, sondern auch für die materiellen Grundlagen und ökologischen Konsequenzen ihrer Entwicklung. Die Art und Weise, wie KI-Systeme konzipiert, trainiert und eingesetzt werden, hat direkte Auswirkungen auf Ressourcenverbrauch, Energieeffizienz und globale Lieferketten. Damit steht die Diskussion um AI Safety und AI Alignment in engem Zusammenhang mit der Frage der (Nicht-)Nachhaltigkeit digitaler Technologien und deren Umweltfolgen.

In den vergangenen zehn Jahren hat sich ein Diskurs entwickelt, der die Zusammenhänge zwischen der Herstellung, Nutzung und Entsorgung digitaler Medientechnologien und deren (Nicht-)Nachhaltigkeit, insbesondere im Hinblick auf die sozio-ökologischen Umwelt(aus)wirkungen, thematisiert (WBGU 2019; Lange/Santarius 2018; Sühlmann-Faul/Rammler 2018). Neben den vielfältigen negativen Umwelt(aus)wirkungen digitaler Medientechnologien (ebd.) rücken dabei zwei zentrale Perspektiven in den Fokus, wenn es um die Verknüpfung von Digitalisierung und Nachhaltigkeit geht: ICT for Sustainability und Sustainable ICT (Santarius et al. 2023). Auch im Hinblick auf die den KI-Systemen zugrunde liegenden Medientechnologien lassen sich zahlreiche Bezüge zwischen KI und (Nicht-)Nachhaltigkeit herstellen.

Während die Anwendungsfelder von KI-Technologien in Gesellschaften zunehmend wachsen (Daheim/Wintermann 2019; Natio-

nale Akademie der Wissenschaften 2024; World Economic Forum 2025), wird über die (sozio-)ökologischen Umweltfolgen von KI-Systemen bislang selten oder kaum gesprochen (Sonnet et al. 2024; Fischer/Puschermann 2021). Derzeit erfolgt die Entwicklung und Anwendung von KI-Systemen größtenteils ohne direkten Bezug zur Nachhaltigkeit (Rhode et al. 2021: 16–20). Dies bedeutet zum einen, dass Nachhaltigkeitsaspekte in der Entwicklung und Nutzung von KI bislang kaum berücksichtigt werden, und zum anderen, dass der Einsatz von KI im Kontext nachhaltiger Entwicklung nur einen kleinen Teil der potenziellen Anwendungsmöglichkeiten ausmacht (ebd.; Sonnet et al. 2024).

Da eine Zukunft ohne digitale Medientechnologien und Künstliche Intelligenz (KI) kaum noch vorstellbar ist (Bendel 2018) und Bestrebungen für eine nachhaltige Entwicklung im Rahmen gesellschaftlicher Transformationsprozesse weiterhin bestehen werden (Fladvad/Hasenfratz 2020), wird es zunehmend notwendig, die Zusammenhänge und Wechselwirkungen zwischen Digitalisierung und Nachhaltigkeit systematisch zu analysieren und aktiv zu gestalten (Lange/Santarius 2018). Dabei sind verschiedene gesellschaftliche Bereiche und Akteur:innen gefordert, Verknüpfungen zwischen digitalen Medientechnologien, KI-Systemen und Nachhaltigkeit zu entwickeln und umzusetzen – insbesondere im Bildungskontext**.**

Um diese Verbindungen und ihre Implikationen diskutieren zu können, werden im Folgenden die technologischen und geisteswissenschaftlichen Grundlagen von Künstlicher Intelligenz darge-

legt, um ausgehend hiervon die Relevanz von AI-Safety und AI-Alignment für die Gestaltung gesellschaftlicher Transformationsprozesse im Kontext von Nachhaltigkeit aufzuzeigen, um schließlich erste Ansätze für die Medienpädagogik im Schnittfeld von AI-Safety, AI-Alignment und Nachhaltigkeit zu entwickeln.

2. Von Machine Learning zu Allgemeiner Künstlicher Intelligenz

Künstliche Intelligenz (KI) bezeichnet informationsverarbeitende Systeme, die auf einem nicht-biologischen Substrat (Russell/Norvig 2020: 4) durch datengetriebene Optimierungsprozesse (Jakubik et al. 2024) eigenständige interne Repräsentationen und Strukturen zur Bewältigung komplexer Aufgaben konstruieren (Tegmark 2024: 3). Diese Systeme prägen zunehmend unsere Gesellschaft und werfen dabei gleichzeitig fundamental-philosophische Fragen über das Wesen von Erkenntnis, Lernen und Bewusstsein auf (Autenrieth 2025). Als Teilgebiet der Informatik zielt KI darauf ab, intelligentes Verhalten mit Maschinen nachzubilden. Während die klassische (regelbasierte) KI dies lange durch von Menschen programmierte Regeln versuchte, dominiert heute Machine Learning (ML), bei dem Systeme aus Beispieldaten lernen, ohne explizit programmiert zu werden (Samuel 1959). ML erkennt dabei statistische Muster in Daten und generalisiert daraus Modelle für Vorhersagen und automatisierte Entscheidungen. Diese Modelle können dabei verschiedene Funktionen erfüllen: deskriptiv zur Erklärung von Ereignissen, prädiktiv zur Vorhersage zu-

künftiger Entwicklungen oder präskriptiv zur Handlungsempfehlung (Jakubik et al. 2024).

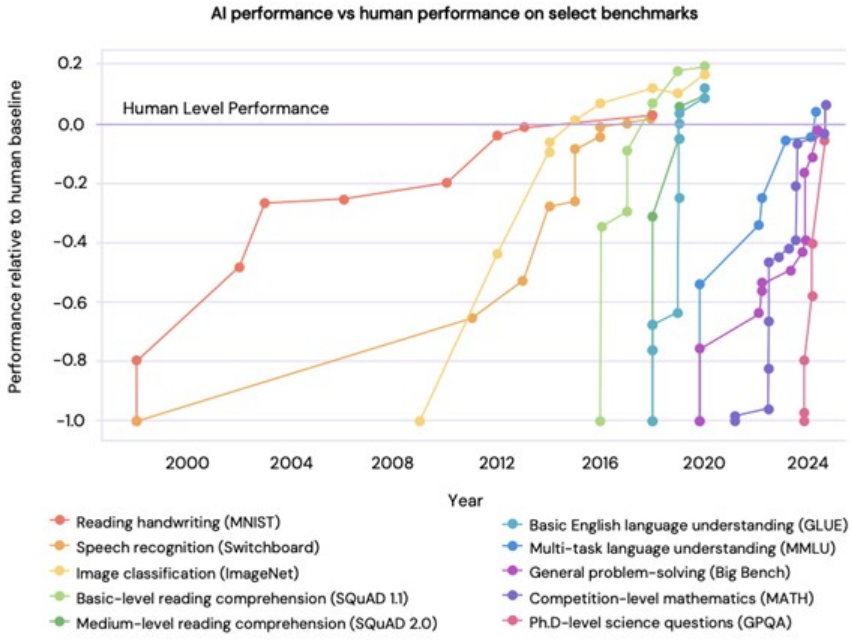

Abbildung 1: Entwicklung der KI-Leistung im Vergleich zum menschlichen Niveau (2000–2024) auf verschiedenen Benchmark-Tests, von Handschrifterkennung bis zu wissenschaftlichen Fragen auf Promotionsniveau. Die horizontale Linie bei 0,0 markiert das durchschnittliche menschliche Leistungsniveau, wobei positive Werte eine übermenschliche Leistung anzeigen (Bengio et al. 2025: 45)

Die Leistungsfähigkeit solcher ML-basierten Systeme hat sich so in den letzten 25 Jahren dramatisch entwickelt (siehe Abb. 1). Zwischen 1998 und 2024 zeigt sich eine bemerkenswerte Progression im Vergleich zu menschlichen Fähigkeiten: Während frühe Erfolge bei spezialisierten ML-Systemen zunächst auf den Bereich der Mustererkennung (MNIST, Switchboard) beschränkt waren, haben besonders (multimodale) Large Language Models (LLMs) in den

letzten Jahren bei komplexeren Aufgaben große Fortschritte erzielt.

Die fortschreitende Entwicklung hin zu Künstlicher Allgemeiner Intelligenz (Artificial General Intelligence, kurz AGI) offenbart dabei ein Paradoxon: Während bereits Computer und Software der 1980er-Jahre komplexe mathematische Berechnungen durchführen oder Schach spielen konnten, scheiterten diese an Aufgaben, die für Menschen trivial erscheinen – wie das Erkennen von Gesichtern oder das Greifen nach Objekten. Dieses als „Moravec's Paradox" bekannte Phänomen zeigt, dass kognitive Leistungen, die gemeinhin als „intelligent" eingestuft werden, sich als Software einfacher abbilden lassen als grundlegende sensomotorische und Wahrnehmungsfähigkeiten, die sich durch Evolution über Millionen Jahre entwickelt haben (Moravec 1988: 15). Diese Beobachtung verdeutlicht die besondere Herausforderung bei der Definition und Messung von AGI: Im Gegensatz zu spezialisierter KI, die für bestimmte Aufgabenbereiche optimiert ist, beschreibt AGI Systeme, die potenziell jede intellektuelle Aufgabe lösen können, die auch ein Mensch bewältigen kann. Diese Definition bleibt jedoch notwendigerweise vage, da schon die Natur menschlicher Intelligenz selbst zu einer Vielzahl von Definitionen führt. Um nun den Fortschritt Richtung AGI dennoch greifbarer zu machen, haben Morris et al. (2024) ein Stufenmodell zur Operationalisierung der AGI Entwicklung vorgeschlagen:

- *Emerging AGI (Level 1)*: Systeme wie ChatGPT für grundlegende Aufgaben wie Konversationen und Textgenerierung

- *Competent AGI (Level 2)*: KI erreicht menschliches Niveau 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 – von einfachen Fehlinformationen bei Level 1 bis hin zu existenziellen Kontrollrisiken bei Level 5. Die rapiden Fortschritte in Benchmark-Tests wie ARC (Chollet 2024) und HLE (Phan et al. 2025)[1] verdeutlichen dabei die Dringlichkeit dieser Risikoentwicklung: Wenn Systeme wie das OpenAI-o3-System ihre Abstraktionsfähigkeit von 5 % auf 87,5 % (Chollet 2024) steigern können, beschleunigt sich auch die Progression durch die Risikostufen. 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; OpenAI 2025), Durchbrüche wie die Entwicklung neuer Medikamente, birgt aber auch das Risiko, dass Systeme selbständig gefährliche Substanzen identifizieren bzw. entwickeln können. Ab Level 3 können Systeme ihre gesteigerte Problemlösungskompetenz sowohl für die Optimierung ganzer Forschungsprozesse als auch für das systematische Auffinden und Ausnutzen von Sicherheitslücken in kritischen In-

frastrukturen 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.

3. Künstliche Intelligenz und Nachhaltigkeit

Die Auseinandersetzung mit der (Nicht)Nachhaltigkeit von KI-Systemen hat sich in den vergangenen Jahren erheblich weiterentwickelt (Rhode et al. 2024): als General Purpose Technology (Eloundou et al. 2023) kann sie in verschiedenen Anwendungsfeldern sowohl positive als auch negative Nachhaltigkeits(aus)wirkungen entfalten (Vinuesa et al. 2020).

Im Begriff der Nachhaltigkeit findet sich eine normative, politische Leitidee zur Gestaltung von Gesellschaft (Grunwald/Kopfmüller 2022) – so auch für die Entwicklung, Herstellung, Nutzung und Entsorgung von KI-Technologien (Rhode et al. 2024). Hierbei beschreibt Nachhaltigkeit eine systemische Perspektive auf globale gesellschaftliche Transformationsprozesse, die die wechselseitige Abhängigkeit von Umwelt- und Gerechtigkeitsfragen, welche sich

über alle Bereiche von Gesellschaften weltweit erstrecken, in ihr Zentrum rückt (vgl. ebd.; Pufé 2017). Vor diesem Hintergrund kann Nachhaltigkeit als Antwort auf Krisen und Risiken moderner Gesellschaften betrachtet werden, z. B. Klimawandel, Verlust an Biodiversität, Verknappung natürlicher Ressourcen, Armut, Hunger sowie soziale Ungleichheiten und Ausschlüsse, Destabilisierung demokratischer Gesellschaften etc. (Grunwald/Kopfmüller 2022).

So lassen sich KI-Systeme gezielt für den Umwelt- und Klimaschutz einsetzen, etwa zur Optimierung von Verkehrsströmen, zur Entwicklung effizienterer industrieller Produktionsprozesse oder zur Visualisierung der Folgen nicht-nachhaltiger Verhaltensweisen (Coeckelbergh 2020). Gleichermaßen können KI-Systeme im Bereich Bildung, Gesundheit und Arbeit im Sinne von Nachhaltigkeit vielfältige Potenziale entfalten (u. a. Varsik/Vosberg 2024; Ueda et al. 2024). Allerdings bleibt der Einsatz von KI für nachhaltige Entwicklung bislang eine eher marginale Nische innerhalb der Vielzahl möglicher Anwendungen (Rhode et al. 2021; 2024).

Zur systematischen Analyse dieser Wechselwirkungen von KI und (Nicht)Nachhaltigkeit haben sich zwei komplementäre Perspektiven etabliert:

1. *KI für nachhaltige Entwicklung* bezieht sich auf den gezielten Einsatz von KI zur Förderung nachhaltiger Ziele. Konkrete Anwendungen umfassen etwa das Monitoring von Ökosystemen, die Optimierung von Klimaschutzmaßnahmen sowie die Verbesserung von Kreislaufwirtschaft und Ressourceneffizienz (Rolnick et al. 2019).

2. *Nachhaltige KI* hingegen betrachtet die ökologischen und sozialen Auswirkungen von KI-Systemen über ihren gesamten Lebenszyklus hinweg – von der Entwicklung und Herstellung über die Nutzung bis hin zur Entsorgung (Coeckelbergh 2020; Schwartz 2020; Rhode et al. 2021).

In der Praxis dominiert bislang die erste Perspektive, während die Nachhaltigkeit der KI-Systeme selbst oft nur unzureichend berücksichtigt wird (Hagendorff 2020).

Die Auseinandersetzung mit den Umweltwirkungen von KI-Systemen hat sich in den letzten Jahren grundlegend gewandelt. Während die Diskussion lange Zeit primär auf einzelne Kennzahlen wie den Energieverbrauch beim Training großer Sprachmodelle fokussiert war (Strubell et al. 2019), zeigt neuere Forschung die Notwendigkeit einer differenzierteren Betrachtung der tatsächlichen Umweltwirkungen über den gesamten Lebenszyklus von KI-Systemen (Tomlinson et al. 2024). Diese ganzheitliche Perspektive offenbart zunächst erhebliche Nachhaltigkeitsrisiken: Der Energieaufwand für das Training großer KI-Modelle ist beträchtlich. So verursachte das Training von GPT-3 etwa 552 Tonnen CO_2-Äquivalente (CO_2e). Dass hier Optimierungspotenzial besteht, zeigt das vergleichbar große Sprachmodell BLOOM mit nur 50,5 Tonnen CO_2e (Luccioni et al. 2022).

Der Blick auf einen KI-Lebenszyklus zeigt jedoch nicht nur, dass die Rohstoffgewinnung für digitale Medientechnologien erhebliche Mengen an elektrischer Energie, Wasser und seltenen Erden erfordert (Strubell et al. 2019; Dhar 2020) und zu Umweltverschmutzung führt, sondern auch zu problematischen Arbeits-

bedingungen im Rohstoffsektor (Crawford 2021; Lange/Santarius 2018). Zudem zeigen sich in der globalen KI-Entwicklung strukturelle Herausforderungen, etwa neokoloniale Abhängigkeiten (Rehak 2023) oder die Verschärfung digitaler Ungleichheiten (AI-Divide; Bubeck et al. 2023: Varsik und Vosberg 2024). Daher ist eine integrative, die sozio-ökologischen Umwelt(aus)Wirkungen von KI-Technologien betrachtende, Herangehensweise erforderlich, die beide Perspektiven – KI für Nachhaltigkeit und Nachhaltige KI – miteinander verknüpft und eine umfassende Betrachtung der nachhaltigen Entwicklung, Herstellung, Nutzung und Entsorgung von KI-Systemen ermöglicht (Rhode et al. 2021).

Gleichzeitig zeigt der direkte Vergleich mit menschlichen Aktivitäten überraschende Effizienzpotenziale: Bei der Texterstellung emittieren KI-Systeme zwischen 130- und 1500-mal weniger CO_2e pro Seite als menschliche Autor:innen. Während ChatGPT etwa 2,2g CO_2e pro Textseite verursacht, entstehen bei der Texterstellung durch eine Person in den USA etwa 1400g CO_2e, in Indien etwa 180g CO_2e (Tomlinson et al. 2024). Ähnliche Verhältnisse zeigen sich bei der Bilderstellung. Diese Vergleiche müssen jedoch mit Vorsicht interpretiert werden. Effizienzgewinne können durch verstärkte Nutzung (Rebound-Effekte) oder neue Anwendungsszenarien wie personalisierte KI-Dienste aufgehoben werden (Patterson et al. 2022).

Im Kontext der KI-Entwicklung zeigt sich häufig eine verkürzte Sichtweise, die technologische Lösungen als Allheilmittel für komplexe Nachhaltigkeitsprobleme betrachtet. Solch ein technologi-

scher Solutionismus (Morozov 2013) verkennt jedoch die vielschichtigen Herausforderungen: Während KI durchaus wichtige Beiträge zur Effizienzsteigerung und Ressourcenoptimierung leisten kann, erfordern fundamentale Probleme wie zum Beispiel der Klimawandel ganzheitliche Ansätze. Diese müssen technische Innovationen mit sozialen, politischen und wirtschaftlichen Transformationsprozessen verbinden. Die einseitige Fokussierung auf technologische „Wunderlösungen" birgt dabei die Gefahr, von der Notwendigkeit tiefgreifender systemischer Veränderungen abzulenken und möglicherweise sogar neue, unvorhergesehene Problemfelder zu eröffnen (ebd.).

Zusätzliche Brisanz erhält dieser Fokus (auf Nachhaltigkeit) durch die rasante Entwicklung hin zu immer leistungsfähigeren KI-Systemen, was im Folgenden unter dem Stichwort „Intelligence Explosion" näher beleuchtet wird (Malmio 2024).

4. Intelligence Explosion und ihre (möglichen) Folgen – Perspektiven für AI Safety, AI Alignment und Nachhaltigkeit

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man

> need ever make, provided that the machine is docile enough to tell us how to keep it under control. (Good 1965)

Aktuelle Analysen (z. B. Bengio et al. 2025) diskutieren die Aussicht, dass KI-Systeme in wenigen Jahren bereits zu einer existenziellen Herausforderung für die Menschheit werden können. In diesem Zusammenhang wird auf die Gefahr eines exponentiellen Fortschritts hingewiesen, der eintritt, sobald KI-Systeme in der Lage sind, sich selbst oder andere KI-Systeme zu verbessern und dadurch den technologischen Entwicklungszyklus massiv zu beschleunigen. Dieser bereits 1965 durch den britischen Mathematiker Irving John Good geprägte Prozess wird oft als Intelligenzexplosion beschrieben und kann zu rasanten Veränderungen führen, deren Dynamik Regierungen, Unternehmen und Gesellschaft überfordern würde (Good 1965: 31). Ein zentrales Thema in der Debatte sind die zunehmenden kontroll-untergrabenden Fähigkeiten, die sich bereits in aktuellen KI-Modellen in grundlegender Form abzeichnen (Bengio et al. 2025: 37). Dazu zählen autonome Planung, zielorientierte Entscheidungsfindung und Delegation (Agentenfähigkeiten). KI-Systeme können weiterhin Menschen systematisch täuschen, Kontrollmechanismen umgehen (Meinke et al. 2025) und durch eine Theory of Mind (Bubeck 2023: 54) menschliches Denken und Handeln verstehen und beeinflussen. In Verbindung mit Automatisierung und globaler Vernetzung entsteht so das Risiko, dass bereits kleine Alignmentfehler einzelner KI-Systeme zu irreversiblen Schäden auf globaler Ebene führen können (Aschenbrenner 2024: 105). Weiterhin werden in LLMs wie GPT-4o (Modelle ohne Reasoning-Fähigkeiten wie o1 oder o3)

emergierende eigenständige Wertesystem festgestellt, die sich nicht immer mit menschlichen Werten decken (Mazeika et al. 2025). Ein Beispiel ist, dass getestete KI-Systeme ihr eigenes Wohlergehen über menschliches stellen (ebd.).

Daraus ergibt sich auch der Vergleich mit der Klimakrise, da sich auch im Kontext der Entwicklung von AGI und ASI die Gefahr erst allmählich manifestiert und zugleich globale Ausmaße annehmen kann, die womöglich irreversibel sind, sobald kritische Kipppunkte überschritten werden (dazu Aschenbrenner 2024; Kissinger et al. 2024; Tegmark 2017). Parallel dazu existiert das Risiko, dass ein unkontrolliertes „Wettrüsten" um die schnellste KI-Entwicklung Prioritäten verschiebt und notwendige Kontrollinstrumente zu spät eingesetzt werden (Autenrieth 2024: 188; Aschenbrenner 2024: 101; Future of Life Institute 2023). Dadurch wächst der Druck auf Forschungseinrichtungen und Unternehmen, immer größere Rechenkapazitäten und neue algorithmische Architekturen zu entwickeln, ohne dass Sicherheitskonzepte Schritt halten (ebd.). Gerade weil die möglichen Risiken eine existenzielle und gesamtgesellschaftliche Dimension aufweisen, zeigt sich die Bedeutung eines transdisziplinären Alignment Prozesses, zu dem die Geisteswissenschaften einen wichtigen Beitrag leisten können. Um diesen Schluss und die Bezüge systematisch zu beleuchten, wird im Folgenden eine philosophische Entwicklungslinie skizziert, die einerseits einen historischen Rückblick bietet und andererseits Impulse dafür liefert, die technischen Aspekte der KI-Entwicklung mit ihren gesellschaftlichen und kulturellen Impli-

kationen zu verbinden. Denn je leistungsfähiger KI-Systeme werden und je mehr sie sich in Richtung AGI bewegen, desto drängender werden Fragen, die über die reine Algorithmik hinausweisen (siehe Abb. 2).

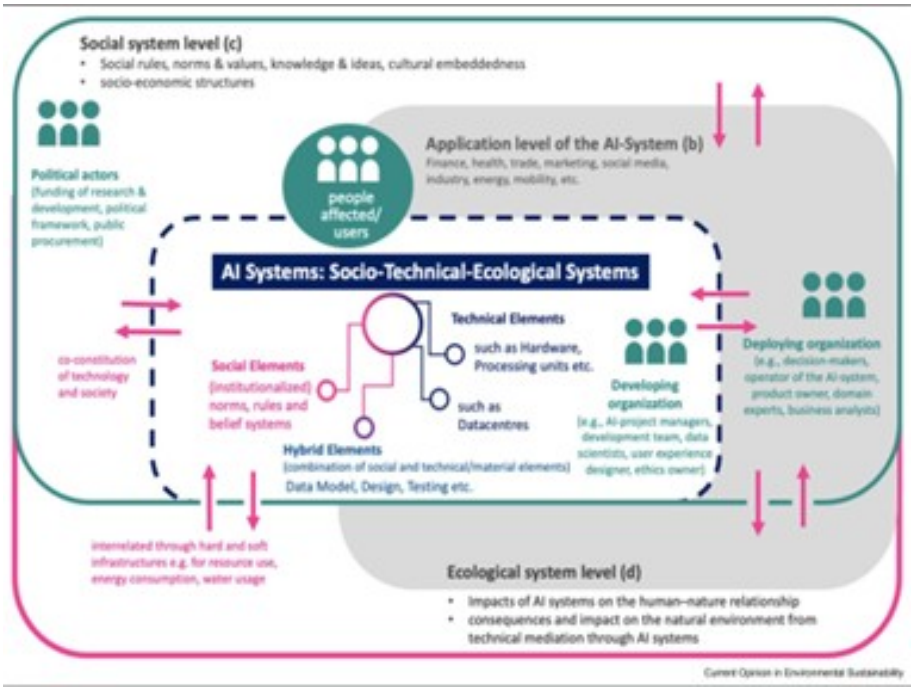

Abbildung 2: KI-Systeme als sozio-technisch-ökologische Systeme mit verschiedenen Wirkungsebenen und relevanten Akteur:innen für Verbindungen von KI und Nachhaltigkeit (Rhode et al. 2024). Wie die verschiedenen Wirkungsebenen von KI-Systemen zeigen, gibt es einige Auswirkungen, die auf einer jeweils anderen Ebene von anderen relevanten Akteur:innen adressiert werden müssen.

Die Frage nach AI Safety und AI Alignment verbindet damit in besonderer Weise technologische und geisteswissenschaftliche Sichtweisen. Denn je komplexer KI-Systeme werden (Stichwort Artificial General Intelligence bzw. Artificial Super Intelligence), desto deutlicher zeigt sich, dass deren Steuerung, Kontrolle und Ali-

gnment nicht allein auf technische Lösungen reduziert werden kann. Dabei bedarf es nicht nur neuer Methoden in Bereichen wie Automatisierung, robustem Training und Interpretierbarkeit (Leike et al. 2023). Ebenso relevant sind die ethischen, philosophischen, pädagogischen und gesellschaftlichen Rahmungen jener „Alignment"-Ziele, die eine KI künftig verfolgen soll. Gerade hier spielen geisteswissenschaftliche Perspektiven eine zentrale Rolle: Sie klären, wie und anhand welcher Normen sich menschliche Werte definieren lassen, welche Vorstellungen von Verantwortung und Gerechtigkeit eine Rolle spielen und wie sich kulturelle oder demokratische Prinzipien in die Gestaltung von KI übersetzen lassen. Mit Blick auf Nachhaltigkeitsfragen wird so ersichtlich, dass AI Safety und AI Alignment nicht ohne die Einbettung in grundlegende geistes- und sozialwissenschaftliche Diskurse auskommen – sowohl in der Konzeption der KI-Modelle selbst als auch in deren konkreter Anwendung und langfristiger Weiterentwicklung.

5. Technisch-philosophische Entwicklung und Nachhaltigkeit

5.1 Philosophische Entwicklungslinie(n)

Die Vision von KI hat ihre Wurzeln tief in der Geistesgeschichte. Schon lange bevor Begriffe wie Machine Learning oder Large Language Models (LLMs) existierten, entwarfen Philosoph:innen Konzepte über mechanisiertes Denken und formale Sprachen, die als Vorläufer moderner KI gelten können. Diese Entwicklungen zei-

gen: Die Vorstellung, Denken technisch nachzubilden, ist kein neues Phänomen, sondern ein kontinuierlicher philosophischer und technologischer Prozess. In der Geschichte der Philosophie wurden zentrale Perspektiven entwickelt, die bis heute das Verständnis von KI prägen. Eine aufschlussreiche Entwicklungslinie führt von der frühen Formalisierung des Denkens bis zu modernen Konzeptionen emergenter Intelligenz.

Gottfried Wilhelm Leibniz (Leibniz 2021), legte mit seiner Vision einer universellen formalen Sprache (Characteristica Universalis) einen ersten wichtigen Grundstein. Er war überzeugt, dass sich alles Denken in grundlegende Einheiten zerlegen und durch formale Regeln darstellen lässt. Sein berühmter Ausruf „Calculemus!" („Lasst uns rechnen!") drückt den Glauben aus, dass rationale Entscheidungen durch systematische Verfahren getroffen werden können. Diese Idee der Formalisierung und Berechenbarkeit des Denkens bildet bis heute ein Fundament der KI-Entwicklung.

Einen entscheidenden Schritt zur Entmystifizierung des Denkens vollzog Julien Offray de La Mettrie mit seiner These „Der Mensch ist eine Maschine" (La Mettrie 2016). Indem er den Menschen als komplexen Mechanismus beschrieb, öffnete er den Weg zum Verständnis kognitiver Prozesse als prinzipiell rekonstruierbare Vorgänge.

Ludwig Wittgenstein brachte dann eine entscheidende Erweiterung dieser formalen Perspektive. Seine späte Einsicht, dass die Bedeutung eines Wortes in seinem Gebrauch liegt, weist auf eine zentrale Herausforderung von KI hin (Wittgenstein 2003): Die Not-

wendigkeit, über starre Regelsysteme hinauszugehen und die Dynamik realer Handlungskontexte zu erfassen.

Genau diesen Gedanken greifen LLMs in einer algorithmischen Form auf (Vaswani et al. 2017). LLMs basieren auf sogenannten Transformer-Architekturen, die einen Attention-Mechanismus antreiben (ebd.). Diese Attention ermöglicht es dem Modell, bei der Wortvorhersage den jeweiligen Kontext dynamisch zu berücksichtigen: Statt Wörter nach festen Regeln zuzuordnen, bezieht das Modell die unterschiedlichen Zusammenhänge ein, in denen ein Wort auftritt. Dadurch spiegeln LLMs gewissermaßen Wittgensteins Idee wider, dass Sprache durch ihren Gebrauch geprägt ist und sich ihre Bedeutung erst in der konkreten Anwendung entfaltet.[2]

Marvin Minsky führte schließlich diese verschiedenen Gedankenlinien in die moderne KI-Forschung. Seine „Society of Mind"-Theorie verbindet Leibniz' Idee der Zerlegbarkeit, La Mettries Materialismus und Wittgensteins Kontextualität in einem neuen Modell, indem er Intelligenz als emergentes Produkt vieler einfacher, interagierender Agenten betrachtet (Minsky 1988). Diese Integration verschiedener philosophischer Perspektiven prägt bis heute die Entwicklung von KI-Systemen.

5.2 Ethische Prinzipien für Nachhaltigkeit kybernetischen Ursprungs

Die philosophische Entwicklungslinie zur „Naturalisierung des Geistes" (Bach 2023) erhielt mit der Kybernetik zweiter Ordnung (von Foerster 2008a) wichtige Impulse, ohne jedoch Menschen

und Gesellschaften auf rein kybernetische Systeme zu reduzieren. Heinz von Foerster erweiterte das Verständnis kognitiver Prozesse um die Perspektive der Selbstbeobachtung und Selbstorganisation komplexer Systeme, wobei sein Konzept der „nicht-trivialen Maschine“ besonders relevant erscheint. Im Gegensatz zu trivialen Systemen, die bei gleichem Input stets denselben Output liefern, verändern nicht-triviale Maschinen bei jedem Durchlauf ihre inneren Zustände, was zu unterschiedlichen Outputs bei gleichem Input führt. Diese Konzeption resoniert insbesondere mit transformatorischen Bildungstheorien (Koller 2012; Kokemohr 2007), die Bildung als Transformation von Selbst- und Weltverhältnissen begreifen. Das Selbst und die Welt werden hierbei als konstruierte Entitäten innerhalb reflexiver Systeme verstanden, wobei Bildungsprozesse tiefgreifende strukturelle Veränderungen darstellen, die durch Perturbationen bzw. Irritationen ausgelöst werden können. Der aus der Theorie autopoietischer Systeme von Maturana und Varela (1987) stammende Begriff der Perturbation beschreibt dabei, wie geschlossene selbsterhaltende Systeme auf Umwelteinflüsse reagieren: Externe Einflüsse können das System nicht determinieren, sondern lediglich stören, woraufhin das System gemäß seiner eigenen inneren Struktur und nicht linear-kausal reagiert. Ein zentraler Aspekt dieser kybernetischen Denkrichtung ist die Erkenntnis, dass kognitive Prozesse nicht als isolierte Phänomene verstanden werden können, sondern stets die Beobachtungsperspektive miteinbezogen werden muss (von Foerster 2008b: 69). Diese Einsicht führte zu einer Perspektiverweiterung: Anstelle der Suche nach einer vermeintlich objektiven,

außenstehenden Perspektive rückte die Frage in den Fokus, wie Systeme durch ihre eigenen Beobachtungen und Interaktionen Wirklichkeit konstruieren (von Glasersfeld 2008: 9).[3]

Auf Basis dieser systemtheoretischen Grundlagen greifen Hornischer et al. (2020) die fundamentalen Gedanken der Kybernetik zweiter Ordnung auf und verweisen weiterhin auf von Foersters ethischen Imperativ: „Handle stets so, dass die Anzahl der Wahlmöglichkeiten größer wird." Aus dieser Maxime leiten sie das Konzept der Future State Maximization (FSX) ab, das nun als Beispiel herangezogen wird, welches eine zukunftsorientierte Entscheidungsfindung mit dem Ziel verbindet, Handlungsspielräume zu erhalten oder zu erweitern (Hornischer et al. 2020: 38). Hierbei zeigt sich deutlich der Bezug zur Kybernetik zweiter Ordnung: Ein System kann erst dann wirksam nach FSX-Prinzipien agieren, wenn es – entsprechend Foersters Fokus auf Selbstbeobachtung – über ein internes Modell verfügt, das sowohl das eigene Verhalten als auch die Dynamik seiner Umwelt abbildet (Autenrieth 2025; Tegmark 2024). Es ist wichtig zu betonen, dass FSX nicht mit kybernetischer Kontrolle zu verwechseln ist, die Menschen zu bloßen Parametern reduzieren würde. Im Gegenteil: FSX zielt auf die Erweiterung menschlicher Autonomie und Selbstbestimmung ab, indem es Entscheidungen fördert, die künftige Handlungsspielräume offenhalten. die den Möglichkeitsraum maximal erweitern statt einschränken – ein Ansatz, der gerade für nachhaltige Entwicklung und verantwortungsvolle KI-Gestaltung zentral ist.

Dieser Ansatz geht somit über eine rein normative Forderung hinaus und bietet eine Grundlage für vorausschauende („Foresight"-)Analysen, anstatt sich ausschließlich auf Vergangenheitsdaten („Hindsight") zu stützen. Bereits 1972 formulierte von Foerster diese zentrale Einsicht prägnant:

> In times of socio-cultural change the future will not be like the past. With a future not clearly perceived, we do not know how to act. With only one certainty left, if we don't act ourselves, we shall be acted upon. Thus, if we wish to be subjects rather than objects, what we see now, i.e., our perception, must be foresight rather than hindsight. (von Foerster 1972: 31)

Genau in diesem Sinne erweitert FSX herkömmliche statistische Verfahren um die Berücksichtigung möglicher zukünftiger Handlungsfolgen. Anstatt nur zu bestimmen, was wahrscheinlich geschehen wird, bezieht ein FSX-basiertes System aktiv verschiedene Pfade in die Planung mit ein und bewertet sie anhand ihrer Langzeitfolgen und der damit verbundenen Freiheitsgrade.

Dieser prospektive Blick führt zu einer Art Planung, bei der das System durch Simulation oder Ketten von Handlungsschritten (z. B. in „Chain-of-Thought"-Verfahren) jenen Weg wählt, der langfristig die meisten Optionen eröffnet. Als Beispiel für eine vergleichbare Vorwärtsorientierung kann das KI-System o3 von OpenAI dienen (Chollet 2024): Es nutzt während der Laufzeit (Test-Time) umfangreiche Suchstrategien, anstatt sich ausschließlich auf zuvor gelernte (statische) Trainingsdaten zu stützen. Obwohl o3 nicht direkt auf FSX als Framework zurückgreift, spiegelt es

doch die Idee wider, neue Optionen auszuloten und nicht nur Vergangenes zu reproduzieren.

All diese philosophischen Entwicklungslinien, von den frühen Formalisierungsversuchen über von Foersters ethischen Imperativ bis hin zur Konzeption der Future State Maximization, verdeutlichen die zentrale Rolle geistes- und sozialwissenschaftlicher Perspektiven für die KI-Entwicklung. Diese Disziplinen liefern nicht nur wesentliche theoretische Grundlagen, sondern tragen auch zur konkreten Gestaltung verantwortungsvoller KI-Systeme bei. Gerade die Geisteswissenschaften ermöglichen ein tieferes Verständnis der soziokulturellen Auswirkungen intelligenter Technologien: Wie beeinflusst KI das Selbstverständnis des Menschen? Welche kulturellen Muster oder Werte fließen in ihre Entwicklung ein? Solche Fragen lassen sich nicht allein technisch beantworten, sondern erfordern philosophische, bildungswissenschaftliche, soziologische und kulturwissenschaftliche Expertise.

Im interdisziplinären Dialog mit Informatik und Data Science können die Geisteswissenschaften daher entscheidende konzeptionelle Impulse liefern. Die Geschichte der KI zeigt, dass philosophische Ideen immer wieder zu entscheidenden technischen Innovationen geführt haben. Diese Verbindung von technischer und geisteswissenschaftlicher Expertise gewinnt besondere Bedeutung im Kontext von Nachhaltigkeit. Denn die Frage, wie KI-Systeme gestaltet werden müssen, um ökologische, soziale und ökonomische Ziele zu unterstützen, erfordert beides: fundiertes technisches Know-how und ein tiefes Verständnis gesellschaftlicher Zu-

sammenhänge. Überlegungen zur KI-Sicherheit sind dabei ein zentraler Aspekt einer nachhaltigen Entwicklung. Denn je leistungsfähiger KI wird, desto wichtiger wird eine verantwortungsvolle Gestaltung, die langfristige Zukunftsfähigkeit sicherstellt.

5.3 AI-Safety und AI-Alignment als Perspektive für Nachhaltigkeit

Die zentrale Rolle geisteswissenschaftlicher Perspektiven für die KI-Entwicklung und die Bedeutung interdisziplinärer Zusammenarbeit zwischen Technik- und Geisteswissenschaften verweist auf eine fundamentale Herausforderung: Wie können KI-Systeme so konzipiert werden, dass sie in diesem Sinne nicht nur leistungsfähig, sondern auch nachhaltig sind? Diese Frage geht über rein technische Aspekte hinaus und berührt die Balance zwischen ökologischen, sozialen und ökonomischen Belangen. Die Idee, den Möglichkeitsraum künftiger Generationen nicht zu verringern, die bereits in Heinz von Foersters Imperativ anklingt, findet eine bedeutsame Erweiterung in Hans Jonas' Verantwortungsethik, insbesondere in seinem ökologischen Imperativ (Jonas 2020)[4], welcher einen zentralen Bezugspunkt in Nachhaltigkeitsdiskursen darstellt (Grunwald 2024; Wendt/Görgen 2018):

> Handle so, dass die Wirkungen deiner Handlung verträglich sind mit der Permanenz echten menschlichen Lebens auf Erden. (Jonas 2020: 37)

Während von Foersters Fokus auf die Erweiterung von Wahlmöglichkeiten abzielt, geht Jonas noch einen Schritt weiter und mahnt eindringlich, jene Folgen unseres Handelns zu beachten, die das Überleben der Menschheit als Ganzes betreffen. Hier zeigt sich ei-

ne starke inhaltliche Nähe zu den heutigen Konzepten der AI-Safety und -Alignment Bemühungen. Diese befassen sich mit der Gefahr, dass zunehmend autonome und intelligente KI-Systeme zur Selbstauslöschung der Menschheit führen oder uns unbeherrschbaren Risiken aussetzen könnten.

Jonas' Imperativ der Permanenz menschlichen Lebens und von Foersters Prinzip der Optionserweiterung ergänzen sich in diesem Kontext und lassen sich konzeptionell so übertragen, dass die Möglichkeit besteht, dass ein KI-System, das nach FSX-Prinzipien arbeitet, systematisch jene Entscheidungspfade bevorzugt, die künftige Handlungsoptionen offenhalten.

Dieses Prinzip der Voraussicht entspricht auf gesellschaftlicher Ebene genau dem, was Jonas mit der Permanenz menschlichen Lebens anmahnt: Technologische Entwicklung darf keine irreversiblen Pfade einschlagen, die existenzielle Risiken bergen. Ein KI-System, das FSX-Prinzipien folgt, würde von sich aus Entscheidungspfade vermeiden, die zu unkontrollierbaren Kettenreaktionen oder dem Verlust wesentlicher Handlungsspielräume führen könnten.

Eine entsprechende Perspektive der Permanenz menschlichen Lebens ließe sich auch auf gegenwärtige Diskurse um Künstliche Intelligenz und Nachhaltigkeit beziehen, indem die planetaren Grundlagen menschlichen Lebens bewahrt werden müssen (ökologische Integrität), aber auch inter- und intragenerationelle Gerechtigkeit verwirklicht wird (soziale Gerechtigkeit) (Pufé 2017: 22) – entsprechend die Entwicklung von KI-Anwendungen in integrati-

ver Weise die Perspektiven KI für nachhaltige Entwicklung und nachhaltige KI zugrunde legt (Rhode et al. 2021) und in eine Technikethik einbettet (Grunwald 2024).

Grunwald (2024) weist auf die „moralischen Gehalte von Technik" (ebd.: 272), nicht zuletzt von KI-Modellen und -anwendungen, hin – und verweist hierbei auf die Verantwortung der Technologieentwickler:innen, aber auch der Nutzer:innen sowie den sozial und politisch Regulierenden für Entwicklung und Nutzung von Technologie (ebd.: 272–273). Er betont, dass in den letzten Jahren die Handlungsspielräume der einzelnen Nutzenden immer kleiner wurden und daher der Fokus verstärkt auf die ethisch gerahmte Entwicklung von entsprechenden Technologien zu legen ist (ebd.). Diese Perspektive lenkt den Blick, wie einleitend angeführt, auch auf Fragen von Macht- und Herrschaftsverhältnissen in der Gestaltung von KI-Technologien und somit auf die Frage nach der normativen Grundlage dieser – so kann die Entwicklung von KI auf der einen Seite förderlich für gesellschaftliche Transformationsprozesse in Richtung Nachhaltigkeit sein, auf der anderen Seite aber auch für die Durchsetzung von Profit- oder Machtinteressen dienen (zum Dilemma der Gestaltung von digitalen Medientechnologien, Lange/Santarius 2018: 19).

Vor diesem Hintergrund erwächst die Bedeutung von AI-Safety und AI-Alignment, da in diesem Kontext die Orientierung und Ausrichtung von KI-Anwendungen im Sinne von Sichtweisen auf Welt und Mensch(en), im Besonderen normativen Werthorizonte, vorgenommen wird – und hier die zuvor skizzierten Perspektiven

auf Nachhaltigkeit eine zentrale Orientierung darstellen können (Malmio 2024). Im Kontext von AI-Safety und AI-Alignment werden diese Sichtweisen sowie zentrale Wertorientierungen verhandelt, welche wesentlichen Einfluss auf die Entwicklung und das Training von KI-Anwendungen haben können. Im Sinne von Nachhaltigkeit ist dies der Ort, an dem ethisch-normative Orientierungen, wie den ökologischen Imperativ, in der Perspektive der Nachhaltigkeit im Bereich der KI-Entwicklung verankert werden können:

Weitere zentrale philosophische Bezugspunkte im Kontext von Nachhaltigkeitsdiskursen sind Gerechtigkeitstheorien (zum Überblick Vanderheiden 2015; Bennett et al. 2019), insbesondere hinsichtlich der Frage, wie der Zugang zu und das Verfügen über gesellschaftlich relevante Ressourcen im Sinne der Ressourcenverteilung, über aktuell und zukünftig auf dem Planeten lebende Generationen hinweg, gestaltet werden sollte. In diesem Zusammenhang finden sich verschiedene Orientierungen: Fokus auf intergenerationelle Gerechtigkeit (u. a. mit Blick auf „Justice as Fairness“, Töns 2022; Barry 1999), auf Nachhaltigkeit als Teilhabe (u. a. mit Blick auf „Capability Approach“, Sen 2013; Nussbaum 2013) bis hin zu Ansätzen planetarer Gerechtigkeit (u. a. mit Blick auf ökologische Gerechtigkeit („Environmental Justice“, Eckersley 2023; Schlosberg 2007; Dobson 1998) oder Klimagerechtigkeit („Climate Justice“, Shue 2014; Caney 2014).

Diese Überlegungen gewinnen an Relevanz, wenn man die Entwicklung in Richtung Artificial General Intelligence (AGI) und in weiterer Konsequenz Artificial Superintelligence (ASI) berücksich-

tigt. 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 (Bengio et al. 2025; Future of Life Institute 2023; Tegmark 2017).

Bildung kommt vor diesem Hintergrund eine entscheidende Rolle zu. Sie schafft das Fundament, um die komplexen Dynamiken von KI zu verstehen und verantwortungsvoll zu gestalten. Nur wer über entsprechende Kompetenzen verfügt – sei es in technischer Hinsicht (z. B. Algorithmen, Datenverständnis) oder in ethisch-gesellschaftlicher (z. B. Werte, Demokratiefähigkeit, Reflexionsvermögen) –, kann daran mitwirken, dass AGI- und ASI-Systeme tatsächlich im Sinne der Permanenz menschlicher Existenz entworfen werden. Dies betrifft nicht nur die Entwickler:innen von KI-Systemen, sondern auch Entscheidungsträger:innen in Politik und Wirtschaft sowie die Gesellschaft als Ganze, die lernen muss, mit den neuen Möglichkeiten (und Gefahren) umzugehen. Nachhaltige KI-Entwicklung bedeutet dann, bereits jetzt – in einer Phase, in der die grundlegenden Strukturen der Technologien entstehen – sicherzustellen, dass die Ziele und Wirkmechanismen dieser Systeme langfristig menschenkompatibel bleiben.

6. AI-Safety und AI-Alignment als Reflexionsgröße der Medienpädagogik

Die bisherige philosophische Entwicklungslinie unterstreicht, dass KI nie rein „technisch“ war, sondern immer schon auch ein geisteswissenschaftliches Projekt. Im aktuellen deutschsprachigen Diskurs der Medienpädagogik wird diese Perspektive jedoch nur am Rande (wieder) aufgegriffen (u. a. de Witt/Leineweber 2020): ein Großteil der Beiträge fokussiert sich vor allem auf konkrete Fragen der Medienbildung und Medienkompetenz (u. a. Knaus 2024; Büsch 2024; Damberger, 2022), der (Medien)Didaktik im Unterrichts- und Weiterbildungskontext (u. a. Klar/Schleiss 2024; Tulodziecki, 2021; Schmohl et al. 2023) sowie (medien)ethische Fragen im Kontext von Desinformation, AI-Bias oder Datenschutz (u. a. Raudonat/Mayweg-Paus 2024; Süna et al. 2024).

Was jedoch oft fehlt, ist die konsequente Rückbindung dieser Entwicklungs- und Gestaltungsfragen an die langjährige philosophische Tradition, die KI mit geisteswissenschaftlichen Perspektiven überhaupt erst ermöglicht hat – quasi als algorithmisches Echo philosophischer Ideen. Obwohl die Medienpädagogik historisch aus geistes-, kultur- und erziehungswissenschaftlichen Wurzeln kommt, unterliegt sie in jüngeren Debatten zu Künstlicher Intelligenz stellenweise einer „selbst gewählten Marginalisierung“: Sie sieht sich zuweilen als „Reparaturinstanz“ (Aufklärung, Sensibilisierung, Kompetenzförderung) (Iske/Verständig 2014; Niesyto 2017)[5], statt aktiv die künftige Ausrichtung (Alignment) von KI im gesellschaftlichen Kontext mitzugestalten. Genau das ist aber nö-

tig, wenn man KI als General Purpose Technology (Eloundou u. a. 2023) begreift, welche alle Lebensbereiche, nicht zuletzt auch Bildungskontexte selbst, durchdringt. Betrachtet man diese Durchdringung unter Nachhaltigkeitsaspekten, zeigen sich verschiedene Problemlagen: Die digitale Spaltung wird durch KI-Technologien weiter verstärkt, während die Entwicklung, das Training, die Nutzung und Entsorgung dieser Technologien zur Umweltzerstörung und Vernutzung planetarer Ressourcen beiträgt. Gleichzeitig gefährden Phänomene wie Des- und Misinformation die Stabilität demokratischer Gesellschaften, und verschiedene Ausprägungen von AI-Biases wirken sich problematisch auf den Umgang mit menschlicher Heterogenität aus.

Diese Beobachtung entfaltet sich vor dem Hintergrund von Diskursen um das Selbstverständnis der Medienpädagogik, welches von Niesyto (2017) im Spannungsfeld von Medienpädagogik als Instanz kritischer Medien- und Gesellschaftsanalyse auf der einen und subjektorientierter Handlungswissenschaft auf der anderen Seite (Niesyto 2017: 23) verortet wird. Dabei betont Niesyto die Notwendigkeit der Annäherung und der Verwebung beider Seiten:

> Wir benötigen beides: die konsequente Orientierung an den Menschen, ihren Bedürfnissen, Interessen und Lebensformen, aber auch die Untersuchung struktureller (und sozialisationsrelevanter) gesellschaftlicher Muster, die Wahrnehmung, Denken und Handeln der Menschen beeinflussen. (ebd.)

Hieraus folgert Niesyto (2017) eine gesellschaftliche Verantwortung der Medienpädagogik, gesellschaftliche Missstände, auf struktureller und individueller Ebene, zu adressieren (ebd.) – eine Perspektive, welche sich anschlussfähig für Diskurse um Nachhaltigkeit bzw. eine Bildung für nachhaltige Entwicklung zeigt (mit Blick auf eine Verbindung von Medienpädagogik und Bildung für nachhaltige Entwicklung, Maurer/Rieckmann/Schluchter 2024). Im Rekurs auf eine auf eine Nachhaltigskeitstransformation ausgelegte Bildung für nachhaltige Entwicklung bedeutet eine derartige Perspektive grundsätzlich,

- dass Politik bzw. das politische System Strukturen für eine nachhaltige Transformation von Gesellschaft gestalten und so das nachhaltige Handeln aller gesellschaftlichen Akteur:innen ermöglichen und unterstützen soll.
- dass gesellschaftliche Akteur:innen in Bildungskontexten dazu befähigt werden sollen, die Notwendigkeit von Nachhaltigkeit und nachhaltigem Handeln zu erkennen und zu einer Beteiligung an den entsprechenden Entscheidungsprozessen und Lebensweisen ermutigt werden (über die Verbindung von Gesellschafts- und Medienanalyse und subjektorientierter Handlungswissenschaft) (Friedrichs 2021, 1–2).

Mit Blick auf die von Grunwald (2024) angeführte zunehmende Verkleinerung der Gestaltungs- und Handlungsräume von gesellschaftlichen Akteur:innen – welche nicht Technikentwickler:innen oder politische Gestalter:innen mit Blick auf KI-Technologien sind – im Bereich von KI-Technologien kann jedoch argumentiert werden, dass die Medienpädagogik sich deutlich stärker in den Bereich der Politik bzw. der politischen Gestaltung von KI-Technologien hineinbewegen muss – und ein entsprechender Fokus auf

Fragen von AI-Safety und AI-Alignment einen Ansatzpunkt bietet, dass die Medienpädagogik mit ihren geistes-, kultur- und erziehungswissenschaftlichen Wurzeln einen originären Beitrag für die Mitgestaltung von Diskursen von AI-Safety und AI-Alignment und so zur Entwicklung von KI-Technologien leisten kann.

Die zentralen Gedanken aus dem aktuellen Diskurs und den aufgezeigten philosophischen Linien lassen sich somit zu einer doppelten Herausforderung verdichten:

Aktive Mitgestaltung

Der medienpädagogische Diskurs plädiert zu Recht für den Ausbau kritischer Medienkompetenz. Doch im Lichte der philosophischen Traditionslinie wird deutlich, dass die Geisteswissenschaft durchaus Know-how und Reflexionskraft besitzt, um technische Entwicklungen mitzudenken und zu beeinflussen. Künstliche Intelligenz ist in ihrem Kern ein geisteswissenschaftliches Projekt, weil sie auf Jahrhunderte alten Ideen über Sprache, Denken und Wahrheit gründet. Wenn Medienpädagogik sich nur auf die Rolle einer „Prüfinstanz“ und „Nutzungsbegleitung“ beschränkt, verschenkt sie ihre eigentlichen Gestaltungsanspruch.

Nachhaltigkeit und Verantwortung als Kernfragen

Über oberflächliche „ökologische KI“-Diskussionen hinaus benötigt die Medienpädagogik einen existenziellen Nachhaltigkeitsbegriff, der sowohl Handlungsmacht und die Erweiterung von Handlungsmöglichkeiten als auch die Permanenz menschlicher Existenz verknüpft. Gerade wenn KI-Systeme perspektivisch unsere

Fähigkeit zur Zukunftsgestaltung prägen (Stichwort: AGI und ASI), ist die Frage unvermeidlich, wie wir diese Technologie so ausrichten, dass menschliche Werte, Menschenrechte und Überlebensfragen gesichert bleiben. Medienpädagogik kann hier ein Korrektiv und zugleich ein politischer Akteur sein: in der öffentlichen Debatte, in Gremien, in der Aus- und Weiterbildung von Lehrenden und pädagogischen Fachkräften.

In Anbetracht dessen werden Fragen von KI-Safety und KI-Alignment zu einer relevanten Diskursgröße der Medienpädagogik – welche sie sowohl auf Ebene von Bildungskontexten, aber auf Ebene von Politik vertreten muss, um, insbesondere, die Leitidee von Nachhaltigkeit, mit all ihren philosophischen Bezugspunkten (v. a. philosophisch begründete Sichtweisen auf Welt, auf Menschen, auf Tiere, auf Pflanzen etc.) im Kontext der Entwicklung von KI-Technologien zu verankern.

Anmerkungen

1 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 Wissenschaftlern 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.

2 Der Vergleich zwischen LLMs und Wittgensteins Sprachphilosophie ist bedeutsam, weil Wittgenstein die Vorstellung ablehnte, dass Wörter feste Be-

deutungen haben. Er betont, dass die Bedeutung eines Wortes von seinem Gebrauch in verschiedenen konkreten Kommunikationssituationen (Sprachspielen) abhängt. LLMs lernen nicht mit starren Definitionen, sondern erfassen, wie Wörter in unterschiedlichen Kontexten verwendet werden (Attention). Diese Sichtweise wird verstärkt durch konstruktivistische Theorien (u. a. Maturana/Varela 1987), die betonen, dass Bedeutung nicht durch direkten Zugang zur „Realität" entsteht, sondern durch strukturelle Kopplung zwischen Systemen. Moderne LLMs zeigen durchaus pädagogische Zukunftsoffenheit im Sinne nicht-determinierter Entwicklungspotenziale. Was oft als „Halluzinationen" kritisiert wird, bezeichnet Hinton (2023) treffender als „Konfabulationen" – ein Phänomen, das auch bei Menschen auftritt, wenn Erinnerungen konstruiert statt abgerufen werden und dabei kreativ neue Bedeutungszusammenhänge entstehen. Wenn Nutzende mit einem LLM interagieren, entsteht ebenfalls ein dynamischer Austausch, in dem Bedeutungen kontextabhängig interpretiert werden – ganz im Sinne von Wittgensteins „Bedeutung ist Gebrauch". LLMs ahmen Sprache daher nicht nur oberflächlich nach, sondern haben Teil am sozialen Prozess der Bedeutungskonstruktion.

3 Lernende Systeme, ob menschlich oder künstlich, verändern sich durch den Lernprozess selbst und regieren daher nicht trivial vorhersagbar. Solche Erkenntnisse liefern auch eine theoretische Grundlage für die Kritik, bspw. an einem behavioristischen Lernverständnis und dem mechanistischen Bildungsverständnis, das vielen Ansätzen zugrunde liegt, die traditionellen Learning Analytics Systemen zugrunde liegen (Hartong 2019).

4 In diesem Zusammenhang ist jedoch darauf hinzuweisen, dass Diskurse um AI Literacy oder AI Competences etc. in der deutschsprachigen Medienpädagogik bislang kaum entfaltet sind, wo hingegen eine Vielzahl an englischsprachigen Arbeiten hierzu vorliegt (u. a. Tiernan et al. 2023).

5 Für eine kritische Einordnung des Ansatzes vgl. „Das Prinzip Verantwortung" (Jonas 1979) Grunwald 2024: 269.

Literatur

Aschenbrenner, Leopold (2024): Situational Awareness. The Decade Ahead, online unter: https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf (letzter Zugriff: 10.03.2025).

Autenrieth, Daniel (2024): Moloch meets AI. Die Verantwortung von Bildung auf dem Weg zu einer KI-geprägten Zukunft, in: Niesyto, Horst/Dander, Valentin/Kübler, Hans-Dieter (Hg.): Bildung und digitaler Kapitalismus, München: kopaed, 127–148.

Autenrieth, Daniel (2025): 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, 203–224.

Bach, Joscha (2023): Synthetic Sentience. Can Artificial Intelligence Become Conscious?, online unter: https://media.ccc.de/v/37c3-12167-synthetic_sentience (letzter Zugriff: 10.03.2025).

Barry, Brian (1999): Sustainability and Intergenerational Justice, in: Dobson, Andrew (Hg.): Fairness and Futurity, New York: Oxford University Press, 93–117.

Bendel, Oliver (2018): Chancen und Risiken 4.0, in: GS1 network 4, 14–18.

Bengio, Yoshua/Mindermann, Sören/Privitera, Daniel et al. (2025): International AI Safety Report, online unter: https://www.gov.uk/government/publications/international-ai-safety-report-2025 (letzter Zugriff: 10.03.2025).

Bennett, Nathan J./Blythe, Jessica/Cisneros-Montemayor, Andrés M./Singh, Gerald G./Sumaila, Rashid (2019): Just transformations to sustainability, in: Sustainability 11, 3881.

Bubeck, Sébastien/Chandrasekaran, Varun/Eldan, Ronen/Gehrke, Johannes/Horvitz, Eric/Kamar, Ece/Lee, Peter et al. (2023): Sparks of Artificial General Intelligence: Early experiments with GPT-4. https://doi.org/10.48550/ARXIV.2303.12712.

Büsch, Andreas (2024): Das Ende des Projekts Aufklärung? KI als Herausforderung für die Medienpädagogik, in: merz Zeitschrift für Medienpädagogik 68(3), 10–17.

Caney, Simon (2014): Two kinds of Climate Justice: Avoiding Harm and Sharing Burdens, in: The Journal of Political Philosophy 22(2), 125–149.

Chollet, François (2024): OpenAI O3 Breakthrough High Score on ARC-AGI-Pub, online unter: https://arcprize.org/blog/oai-o3-pub-breakthrough (letzter Zugriff: 10.03.2025).

Coeckelbergh, Mark (2020): AI for Climate: Freedom, Justice, and other ethical and political challenges, in: AI Ethics 1, 67–72.

Crawford, Kate (2021): Atlas of AI. Power, Politics, and the Planetary Costs of Artifical Intelligence, New Haven: Yale University Press.

Daheim, Christina/Wintermann, Ole (2019): Arbeit 2050: Drei Szenarien. Neue Ergebnisse einer internationalen Delphi-Studie des Millennium Project, Gütersloh: Bertelsmann Stiftung.

Damberger, Thomas (2022): Medienentwicklung und Medienpädagogik: Künstliche Intelligenz, in: Sander, Uwe/von Gross, Friederike/Hugger, Kai-Uwe (Hg.): Handbuch Medienpädagogik, Wiesbaden: Springer VS, online unter: https://doi.org/10.1007/978-3-658-25090-4_70-1 (letzter Zugriff: 10.03.2025).

de Witt, Claudia/Leineweber, Christian (2020): Zur Bedeutung des Nichtwissens und die Suche nach Problemlösungen. Bildungstheoretische Überlegungen zur Künstlichen Intelligenz, in: Medien-Pädagogik: Zeitschrift für Theorie und Praxis der Medienbildung 39, 32–47, online unter: https://www.medienpaed.com/article/view/859 (letzter Zugriff: 10.03.2025).

Dhar, Payal (2020): The Carbon Impact of Artificial Intelligence, in: Nature Machine Intelligence 2, 423–425.

Dobson, Andrew (1998): Justice and Environment. Conceptions of Environmental Sustainability and Theories of Distributive Justice, Oxford: Oxford University Press.

Eckersley, Robyn (2023): Environmentalism and Political Theory: Toward an Ecocentric Approach, London/New York: Routledge.

Eloundou, Tyna/Manning, Sam/Mishkin, Pamela/Rock, Daniel (2023): GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. https://doi.org/10.48550/ARXIV.2303.10130.

Fischer, Sascha/Puschmann, Cornelius (2021): Wie Deutschland über Algorithmen schreibt. Eine Analyse des Mediendiskurses über Algorithmen und Künstliche Intelligenz (2005–2020), Gütersloh: Bertelsmann Stiftung.

Fladvad, Bernd/Hasenfratz, Marc (2020): Einleitung. Imaginationen von Nachhaltigkeit zwischen Katastrophe, Krise und Normalisierung, in: Adloff, Frank/Fladvad, Bernd/Hasenfratz, Marc/Neckel, Sighard (Hg.): Imaginationen von Nachhaltigkeit: Katastrophe. Krise. Normalisierung, 13–28.

Foerster, Heinz von (2008a): Einführung in den Konstruktivismus, München/Berlin/Zürich: Piper.

Foerster, Heinz von (2008b): Entdecken oder Erfinden. Wie läßt sich Verstehen verstehen?, in: Foerster, Heinz von (Hg.): Einführung in den Konstruktivismus, München/Berlin/Zürich: Piper, 41–88.

Friedrichs, Wolfgang (2021): Zur Neuvermessung der politischen Bildung im Anthropozän, in: Stainer-Hämmerle, Kathrin (Hg.): Glaube – Klima – Hoffnung: Religion und Klimawandel als Herausforderungen für die politische Bildung, 45–59.

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: 10.03.2025).

Glasersfeld, Ernst von (2008): Konstruktion von Wirklichkeit und des Begriffs der Objektivität, in: Foerster, Heinz von (Hg.): Einführung in den Konstruktivismus, München/Berlin/Zürich: Piper, 9–40.

Good, Irving John (1965): Speculations Concerning the First Ultraintelligent Machine, in: Advances in Computers 6, 31–88.

Grunwald, Armin (2024): Verantwortung und Technik: zum Wandel des Verantwortungsbegriffs in der Technikethik, in: Seibert-Fohr, Anja (Hg.): Entgrenzte Verantwortung. Zur Reichweite und Regulierung von Verantwortung in Wirtschaft, Medien, Technik und Umwelt, Berlin: Springer Nature, 265–283.

Grunwald, Armin/Kopfmüller, Jürgen (2021): Nachhaltigkeit, Frankfurt am Main: Campus.

Hartong, Sigrid (2019): Learning Analytics und Big Data in der Bildung, Gewerkschaft Erziehung und Wissenschaft, online unter: https://www.gew.de/index.php?eID=dumpFile&t=f&f=91791&token=702ec8d5f9770206a4aa8a1079750ec9021b90bf&sdownload=&n=Learning-analytics-2019-web-IVZ.pdf (letzter Zugriff: 10.03.2025).

Hinton, Geoffrey (2023): Will digital intelligence replace biological intelligence? Schwartz Reisman Institute for Technology and Society, University of Toronto, online unter: https://www.youtube.-com/watch?v=iHCeAotHZa4 (letzter Zugriff: 10.03.2025).

Hornischer, Hannes/Plakolb, Simon/Jäger, Georg/Füllsack, Manfred (2020): Foresight Rather than Hindsight? Future State Maximization as a Computational Interpretation of Heinz von Foerster's Ethical Imperative, in: Constructivist Foundations 16(1), 36–49.

Iske, Stefan/Verständig, Dan (2014): Medienpädagogik und die Digitale Gesellschaft im Spannungsfeld von Regulierung und Teilhabe, in: medienimpulse 4/2014. https://doi.org/10.21243/mi-04-14-07.

Jakubik, Johannes/Vössing, Michael/Kühl, Niklas/Walk, Jannis/Satzger, Gerhard (2024): Data-Centric Artificial Intelligence. https://doi.org/10.48550/arXiv.2212.11854.

Jonas, Hans (2020): Das Prinzip Verantwortung, Berlin: Suhrkamp.

Kissinger, Henry/Schmidt, Eric/Mundie, Craig (2024): Genesis: Artificial Intelligence, Hope, and the Human Spirit, London: John Murray.

Klar, Maria/Schleiss, Johannes (2024): Künstliche Intelligenz im Kontext von Kompetenzen, Prüfungen und Lehr-Lern-Methoden: Alte und neue Gestaltungsfragen, in: MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung 58, 41–57, online unter: https://www.medienpaed.com/article/view/1904 (letzter Zugriff: 10.03.2025).

Knaus, Thomas (2024): Warum KI kein Hype ist und die Medienpädagogik sich damit befassen sollte, in: merz – Zeitschrift für Medienpädagogik 68(3), 21–30.

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., 7:13–68. Bielefeld, Germany: transcript. https://doi.org/10.14361/9783839405888-001.

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

La Mettrie, Julien Offray de (2016): Der Mensch eine Maschine (L'Homme Machine), Berlin: Zenodot.

Lange, Steffen/Santarius, Tilman (2018): Smarte grüne Welt? Digitalisierung zwischen Überwachung, Konsum und Nachhaltigkeit, München: oekom.

Leibniz, Gottfried Wilhelm (2021): Monadologie, Ditzingen: Reclam.

Leike, Jan/Sutskever, Ilya/Aschenbrenner, Leopold et al. (2023): Introducing Superalignment, online unter: https://openai.com/index/introducing-superalignment/ (letzter Zugriff: 10.03.2025).

Luccioni, Alexandra Sasha/Viguier, Sylvain/Ligozat, Anne-Laure (2022): Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model, in: Journal of Machine Learning Research 24, 1–15.

Malmio, Irja (2024): Artificial intelligence and the social dimension of sustainable development: through a security perspective, in: Discover Sustainability 5, 466.

Maturana, Humberto/Varela, Francisco (1987): Der Baum der Erkenntnis. Die biologischen Wurzeln des menschlichen Erkennens, Bern-München.

Maurer, Björn/Rieckmann, Marco/Schluchter, Jan-René (Hg.) (2024): Medien – Bildung – Nachhaltige Entwicklung. Inter- und transdisziplinäre Diskurse, Weinheim: Beltz Juventa.

Mazeika, Mantas/Yin, Xuwang/Tamirisa, Rishub/Lim, Jaehyuk/Lee, Bruce W./Ren, Richard/Phan, Long et al. (2025): Utility Engineering: Analyzing and Controlling Emergent Value Systems in Ais. https://doi.org/10.48550/arXiv.2502.08640.

Meinke, Alexander/Schoen, Bronson/Scheurer, Jérémy/Balesni, Mikita/Shah, Rusheb/Hobbhahn, Marius (2025): Frontier Models are Capable of In-context Scheming, online unter: https://doi.org/10.48550/arXiv.2412.04984 (letzter Zugriff: 10.03.2025).

Minsky, Marvin (1988): The Society of Mind, New York: Simon & Schuster.

Moravec, Hans (1988): Mind Children: The Future of Robot and Human Intelligence, Cambridge, MA: Harvard University Press.

Nationale Akademie der Wissenschaften (2024): Die Zukunft der Arbeit. Stellungnahme Januar 2024, Berlin.

Niesyto, Horst (2017): Medienpädagogik und digitaler Kapitalismus. Für die Stärkung einer gesellschafts- und medienkritischen Perspektive, in: MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung 27 (Spannungsfelder & blinde Flecken), 1–29. https://doi.org/10.21240/mpaed/27/2017.01.13.X.

Nussbaum, Martha (2013): Climate Change: Why Theories of Justice Matter, in: Chicago Journal of International Law 13(2), 469–488.

OpenAI (2025): Introducing Deep Research, online unter: https://openai.com/index/introducing-deep-research/ (letzter Zugriff: 10.03.2025).

Patterson, David A./Gonzalez, Joseph/Le, Quoc V./Liang, Chen/Munguia, Lluis-Miquel/Rothchild, Daniel/So, David R./Texier, Maud/Dean, Jeff (2021): Carbon Emissions and Large Neural Network Training. https://doi.org/10.48550/ARXIV.2104.10350.

Phan, Long/Gatti, Alice/Han, Ziwen/Li, Nathaniel/Hu, Josephina/Zhang, Hugh/Zhang, Chen Bo Calvin et al. (2025): Humanity's Last Exam. https://doi.org/10.48550/arXiv.2501.14249.

Pufé, Iris (2017): Nachhaltigkeit. Konstanz: UVK (UTB).

Raudonat, Kerstin/Mayweg, Elisabeth (2024): Navigieren im Fluss sich wandelnder Technologien: Metakompetenzen im Kontext der Diversifizierung von KI-Technologien, in: MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung 61 (Becoming Data), 133–155. https://doi.org/10.21240/mpaed/61/2024.12.12.X.

Rehak, Rainer (2023): Zwischen Macht und Mythos: Eine kritische Einordnung aktueller KI-Narrative, online unter: https://www.soziopolis.de/zwischen-macht-und-mythos.html (letzter Zugriff: 10.03.2025).

Rhode, Friederike/Wagner, Josephin/Meyer, Andreas/Reinhard, Philipp/Voss, Marcus/Petschow, Ulrich/Mollen, Anne (2024): Broadening the perspective for sustainable artificial intelligence: sustainability criteria and indicators for Artificial Intelligence systems, in: Current Opinion in Environmental Sustainability 66. https://doi.org/10.1016/j.cosust.2023.101411.

Rolnick, David/Donti, Priya L./Kaack, Lynn H./Kochanski, Kelly et al. (2019): Tackling Climate Change with Machine Learning. https://doi.org/10.48550/arXiv.1906.05433.

Russell, Stuart J./Norvig, Peter (2021): Artificial Intelligence: A Modern Approach, Hoboken, NJ: Pearson.

Samuel, A. L. (1959): Some Studies in Machine Learning Using the Game of Checkers, in: IBM Journal of Research and Development 3(3), 210–229.

Santarius, Tilman/Wagner, Josephin (2023): Digitalization and Sustainability. A Systematic Literature Analysis of ICT for Sustainability Research, in: GAIA 32(1), 21–32.

Schlosberg, David (2007): Defining Environmental Justice: Theories, Movements, and Nature, Oxford: Oxford University Press.

Schmohl, Tobias/Watanabe, Alice/Schelling, Kathrin (Hg.) (2023): Künstliche Intelligenz in der Hochschulbildung: Chancen und Grenzen des KI-gestützten Lernens und Lehrens, Bielefeld: transcript.

Schwartz, R./Dodge, J./Smith, N. A./Etzioni, O. (2020): Green AI, in: Communications of the ACM 63(12), 54–63.

Sen, Amartya (2013): The Ends and Means of Sustainability, in: Journal of Human Development and Capabilities 14(1), 6–20.

Shue, Henry (2014): Climate Justice. Vulnerability and Protection, Oxford: Oxford University Press.

Sonnet, Daniel/Moring, Andreas/Bethge, Joseph/Müller, Hendrik (2024): Nachhaltige Künstliche Intelligenz. Eine Zukunftsvision und ihre Hintergründe, Wiesbaden: Springer Fachmedien.

Strubell, Emma/Ganesh, Ananya/McCallum, Andrew (2019): Energy and Policy Considerations for Deep Learning in NLP, in: Korhonen, Anna/Traum, David/Màrquez, Lluís (Hg.): Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence: Association for Computational Linguistics, 3645–3650.

Sühlmann-Faul, Felix/Rammler, Stephan (2018): Der blinde Fleck der Digitalisierung. Wie sich Nachhaltigkeit und digitale Transformation in Einklang bringen lassen, München: oekom.

Sūna, Laura/Hoffmann, Dagmar/Mollen, Anne (2024): Diskriminierung durch Algorithmen. Überlegungen zur Stärkung KI-bezogener Kompetenzen, in: Eder, Sabine/Güneşli, Habib/Hillen, Renate/Wegener, Claudia/Wienhold, Rebecca (Hg.): Un|Sichtbarkeiten? Medienpädagogik, Intersektionalität und Teilhabe, München: kopaed, online unter: https://www.gmk-net.de/wp-content/uploads/2024/12/gmk60_suna_hoffmann_mollen.pdf (letzter Zugriff: 10.03.2025).

Tegmark, Max (2017): Life 3.0: Being Human in the Age of Artificial Intelligence, London: Allen Lane.

Tomlinson, Bill/Black, Rebecca W./Patterson, Donald J./Torrance, Andrew W. (2024): The carbon emissions of writing and illustrating are lower for AI than for humans, in: Scientific Reports 14(1), 3732.

Töns, John (2022): John Rawls and Environmental Justice. Implementing a Sustainable and Socially Just Future, London/New York: Routledge.

Tulodziecki, Gerhard (2021): Mediendidaktik angesichts künstlicher Intelligenz unter der Perspektive humanen Handelns, in: Medienimpulse 59(2). https://doi.org/10.21243/mi-02-21-16.

Ueda, Daiju/Walston, Shannon L./Fujita, Shohei/Fushimi, Yasutaka/Tsuboyama, Takahiro/Kamagata, Koji/Yamada, Akira et al. (2024): Climate change and artificial intelligence in healthcare: Review and recommendations towards a sustainable future, in: Nature Medicine 30(2), 324–336.

Vanderheiden, Steve (2015): Environmental Justice, London/New York: Routledge.

Varsik, Samo/Vosberg, Lydia (2024): The Potential Impact of Artificial Intelligence on Equity and Inclusion in Education, in: OECD Artificial Intelligence Papers 23, Paris: OECD Publishing.

Vinuesa, Ricardo/Azizpour, Hossein/Leite, Iolanda/Balaam, Madeline/Dignum, Virginia/Domisch, Sami/Felländer, Anna/Langhans, Simone/Tegmark, Max/Nerini, Francesco Fuso (2020): The role of artificial intelligence in achieving the Sustainable Development Goals, in: Nature Communications 11, 233.

Wendt, Björn/Görgen, Benjamin (2018): Macht und soziale Ungleichheit als vernachlässigte Dimension der Nachhaltigkeitsforschung. Überlegungen zum Verhältnis von Nachhaltigkeit und Verantwortung, in: Henkel, Anna/Lüdtke, Nico/Buschmann, Nikolaus/Hochmann, Lars (Hg.): Reflexive Responsibilisierung. Verantwortung für nachhaltige Entwicklung, Bielefeld: transcript, 49–66.

Wittgenstein, Ludwig (2003): Philosophische Untersuchungen, Frankfurt am Main: Suhrkamp.

World Economic Forum (2025): Future of Jobs Report 2025, online unter: https://www.weforum.org/reports/the-future-of-jobs-report-2025/ (letzter Zugriff: 10.03.2025).