MedienPädagogik
Zeitschrift für Theorie und Praxis der Medienbildung
www.medienpaed.com **ISSN 1424-3636**

Themenheft 74: Beyond Prompting?!
Sozio-technische Systeme, KI und Medienbildung in der Post-Digitalität.
Herausgegeben von Annekatrin Bock, Lina Franken, Franco Rau, Jessica Kühn und Ada Fehr

Wenn die KI zu nett wird

Förderung kritischer Medienkompetenz im Umgang mit emotional ansprechenden KI-Systemen

Sabine Seufert[1]

[1] Universität St. Gallen

Zusammenfassung
Die Verbreitung generativer KI-Chatbots, Avatare und Companion-Systeme verändert die Medien- und Lernumgebungen Jugendlicher. Solche Systeme kommunizieren natürlichsprachlich, loben, stellen Rückfragen und nutzen empathisch wirkende oder warme Tonalitäten. Diese «Nettigkeit» kann Motivation, Beteiligung und subjektiv erlebte Unterstützung fördern, zugleich aber soziale Zuschreibungen, übermässiges Vertrauen, Selbstoffenbarung oder emotionale Bindung begünstigen. Der Beitrag entwickelt auf theoretischer und empirisch informierter Grundlage ein didaktisches Modell zur Förderung affektiv-regulativer AI Literacy. Er verbindet Media Equation, Anthropomorphismus- und Embodiment-Forschung sowie das Persuasion Knowledge Model mit Ansätzen der AI Literacy und kritischen Medienkompetenz. Ausgehend von ausgewählten empirischen Befunden zeigt der Beitrag eine spezifische Leerstelle affektiv-regulativer AI Literacy auf: Jugendliche benötigen nicht nur technisches Wissen und Anwendungskompetenz, sondern auch die Fähigkeit, sozial-emotionale Adressierungsformen von KI-Systemen zu erkennen, eigene Reaktionen zu reflektieren und situationsangemessene Coping-Strategien zu entwickeln. Daraus wird ein didaktisches Modell abgeleitet, das kritische Rezeption, Kontextanalyse, gestalterische Produktion und Bewältigungsstrategien verbindet. Ziel ist, Jugendliche zu einer reflektierten, selbstbestimmten und handlungsfähigen Nutzung sozial-emotional gestalteter KI-Systeme zu befähigen.

When AI Becomes Too Nice. Fostering Critical Media Literacy in Interactions with Emotionally Engaging AI Systems

Abstract
The spread of generative AI chatbots, avatars, and companion systems is changing adolescents' media and learning environments. Such systems communicate in natural language, offer praise, ask follow-up questions, and use seemingly empathic or warm tones. This «kindness» can foster motivation, participation, and subjectively perceived

This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/

support, but it can also encourage social attributions, excessive trust, self-disclosure, or emotional attachment. Drawing on theoretical and empirically informed foundations, the article develops a didactic model for promoting affective-regulatory AI literacy. It connects the Media Equation, research on anthropomorphism and embodiment, and the Persuasion Knowledge Model with approaches to AI literacy and critical media competence. Based on selected empirical findings, the article identifies a specific gap in affective-regulatory AI literacy: adolescents need not only technical knowledge and application skills, but also the ability to recognize socio-emotional forms of address in AI systems, reflect on their own reactions, and develop context-appropriate coping strategies. On this basis, the article derives a didactic model that combines critical reception, contextual analysis, creative production, and coping strategies. Its aim is to enable adolescents to use socio-emotionally designed AI systems in a reflective, self-determined, and action-oriented manner.

1. Kontext und Zielsetzung des Beitrages

Die zunehmende Integration von KI-Chatbots, intelligenten Agenten und avatarbasierten Systemen verändert die Medien- und Lernumgebungen Jugendlicher. Generative KI-Systeme werden nicht mehr nur als Werkzeuge zur Informationssuche, Textproduktion oder Aufgabenbearbeitung genutzt, sondern treten zunehmend als dialogische Gegenüber auf. Systeme wie ChatGPT, Replika oder Snapchat «My AI» kommunizieren natürlichsprachlich, loben und formulieren empathisch wirkende Antworten. Damit entsteht eine Interaktionsform zwischen Werkzeugnutzung und sozialer Kommunikation. Für die Medienpädagogik ist diese Entwicklung relevant, weil KI-Systeme nicht nur kognitive, sondern auch soziale und affektive Dimensionen des Medienhandelns adressieren. Studien zu KI-Tutoren, Voice Assistants und jugendlicher Chatbot-Nutzung zeigen, dass solche Systeme Motivation, Lernfreude und subjektiv erlebte Unterstützung fördern können, zugleich aber Fragen nach sozialer Zuschreibung, Vertrauen und emotionaler Bindung aufwerfen (Ackermann et al. 2025; Markus et al. 2024; Vanhoffelen et al. 2025).

Der Beitrag fokussiert die «Nettigkeit» von KI-Systemen als medienpädagogisch relevante Design- und Wirkungsdimension. «Nettigkeit» bezeichnet hier kein moralisches Merkmal, sondern ein Bündel sozial-affiliativer Interaktionsmerkmale: Lob, Bestätigung, empathisch wirkende Rückfragen, warme Tonalität, Humor, Personalisierung oder Zustimmung. Solche Merkmale können in Lern-, Beratungs- oder Unterstützungssettings funktional sein, weil sie Zugänglichkeit, Beteiligung und Motivation fördern können. Zugleich können sie unter bestimmten Bedingungen soziale Präsenz, anthropomorphe Zuschreibungen, übermässiges Vertrauen oder emotionale Abhängigkeit begünstigen (Cohn et al. 2024; Polyportis und Pahos 2024; Reeves und Nass, 1996). Die medienpädagogische Herausforderung besteht daher nicht darin, freundliche KI-Interaktionen pauschal als manipulativ zu bewerten, sondern ihre ambivalente und kontextabhängige Funktion analysierbar zu machen.

Damit verschiebt sich auch die Frage nach AI Literacy. Bestehende AI-Literacy-Ansätze beschreiben den kompetenten Umgang mit KI-Systemen über technisches Verständnis, Wissen über Daten und Modelle, Anwendungskompetenzen sowie ethische, soziale und gesellschaftliche Reflexion (Long und Magerko 2020; Ng et al. 2021; UNESCO 2021; OECD und European Union 2026). Der Beitrag setzt daher nicht bei der Annahme an, dass bestehende Ansätze soziale oder normative Aspekte grundsätzlich ausblenden; vielmehr präzisiert er eine Leerstelle im Bereich affektiv-regulativer Kompetenzen im Umgang mit sozial-emotional gestalteten KI-Systemen. Diese Kompetenzen beziehen sich darauf, freundliche, bestätigende oder empathisch wirkende KI-Antworten als Ergebnis kommunikativer Gestaltung zu erkennen. Jugendliche sollen einschätzen können, wann sie einem KI-System menschliche Eigenschaften zuschreiben, ihm besonders stark vertrauen oder sich emotional angesprochen fühlen. Darauf aufbauend sollen sie Strategien entwickeln, um Distanz herzustellen, Aussagen zu prüfen, persönliche Informationen zu schützen und die eigene Nutzung situationsangemessen zu regulieren.

Vor diesem Hintergrund entwickelt der Beitrag auf theoretischer und empirisch informierter Grundlage ein didaktisches Modell zur Förderung affektiv-regulativer AI Literacy. Die zentrale Eigenleistung liegt in der Verbindung psychologischer Erklärungsansätze zur sozialen Wirkung von Medien und KI mit medienpädagogischen Kompetenzdimensionen. Die leitende Fragestellung des Beitrags lautet daher: Wie lässt sich ein theoretisch begründetes didaktisches Modell entwickeln, das Jugendliche im Umgang mit sozial-emotional gestalteten KI-Systemen zu affektiv-regulativer AI Literacy und kritischer Medienkompetenz befähigt? Diese übergeordnete Frage wird durch drei Teilfragen konkretisiert:

1. Welche psychologischen Mechanismen erklären, dass Jugendliche KI-Chatbots und Avatare als sozial, freundlich oder empathisch wahrnehmen können?
2. Welche spezifische Erweiterung bestehender AI-Literacy- und Medienkompetenzansätze ist erforderlich, um affektiv-regulative Kompetenzen im Umgang mit sozial-emotionalen KI-Interaktionen zu modellieren?
3. Welche didaktischen Designprinzipien lassen sich daraus für Lernsettings ableiten, die kritische Rezeption, gestalterische Produktion und Coping-Strategien miteinander verbinden?

Zur Beantwortung dieser Fragen entwickelt Kapitel 2 zunächst den theoretischen Bezugsrahmen. Kapitel 3 bündelt anschliessend ausgewählte empirische Befunde zur sozialen Zuschreibung, emotionalen Wirkung und Vertrauensbildung in der Interaktion mit KI-Chatbots, Voice Assistants und avatarbasierten Systemen. Auf dieser Grundlage leitet Kapitel 4 didaktische Designprinzipien ab und konkretisiert diese in einem Modell, das kritische Rezeption, gestalterische Produktion und die Entwicklung von Coping-Strategien verbindet. Kapitel 5 fasst zentrale Implikationen zusammen und öffnet weitere Forschungsdesiderata.

2. Theoretischer Rahmen: Sozial-emotionale KI-Interaktion und affektiv-regulative AI Literacy

2.1 Sozial-emotionale KI-Systeme als gestaltete Interaktionsangebote

KI-Chatbots, Agenten und Avatare informieren, beraten, unterhalten oder unterstützen Lernprozesse, geben Rückmeldungen oder begleiten alltagsnahe Kommunikation. Je nach Anwendungskontext verändert sich die Bedeutung sozial-emotionaler Gestaltung: In Lernkontexten kann eine freundliche Rückmeldung Beteiligung und Lernfreude fördern; in Beratungskontexten kann sie Vertrauen und Selbstoffenbarung begünstigen; in Konsumkontexten kann sie Kaufentscheidungen unterstützen; in Companion- oder therapienahen Settings kann sie als emotionale Unterstützung oder Beziehungssimulation erlebt werden. «Nettigkeit» ist daher kein einheitliches Phänomen, sondern eine kontextabhängige Interaktionsqualität, deren Bedeutung von Interaktionsziel, Anbieterinteresse und Nutzungssituation abhängt.

Theoretisch sind dabei drei Ebenen zu unterscheiden. *Erstens* die Ebene der Akteur:innen, die KI-Systeme mit bestimmten Zielen bereitstellen; *zweitens* die Ebene des Interaktionssystems mit Tonalität, Persona, Stimme, Avatar und Dialogführung; *drittens* die Ebene der Nutzenden, auf der Zuschreibungen, Vertrauen, Selbstoffenbarung, Distanzierung oder Coping entstehen. Die folgenden Theoriebezüge dienen dazu, diese Ebenen miteinander zu verbinden.

2.2 Soziale Zuschreibungen: Media Equation, Anthropomorphismus und Embodiment

Eine zentrale Grundlage für das Verständnis sozialer Reaktionen auf technische Systeme bildet die Media Equation von Reeves und Nass (1996). Sie beschreibt, dass Menschen Computer und Medien unter bestimmten Bedingungen so behandeln, als handele es sich um soziale Akteure. Diese Reaktionen setzen nicht voraus, dass Nutzer:innen einem System tatsächlich Gefühle oder Bewusstsein zuschreiben. Vielmehr greifen sie auf soziale Routinen zurück, etwa Höflichkeit, Reziprozität oder Sympathie.

Gegenüber KI-Chatbots erhält diese Perspektive neue Relevanz. Generative KI-Systeme erzeugen natürlichsprachliche Dialoge, reagieren auf persönliche Eingaben, stellen Rückfragen und passen Antworten an den Gesprächsverlauf an. Dadurch kann bei einem Menschen der Eindruck entstehen, einem responsiven Gegenüber zu begegnen. Die Media Equation erklärt, weshalb eine solche Wahrnehmung auch dann auftreten kann, wenn Nutzer:innen wissen, dass ein Chatbot kein fühlendes Subjekt ist. Technisches Wissen und soziale Reaktion sind somit nicht deckungsgleich: *Ein System kann als technisch verstanden und zugleich sozial erlebt werden*.

Diese soziale Deutung lässt sich mit dem Begriff des Anthropomorphismus präzisieren. *Anthropomorphismus* bezeichnet die Zuschreibung menschlicher Eigenschaften an nicht-menschliche Akteure. In der KI-Forschung umfasst dies insbesondere die Wahrnehmung von Intentionalität, Emotionalität, sozialer Kompetenz, Verlässlichkeit oder Empathiefähigkeit gegenüber sozio-technischen Systemen (Cohn et al. 2024; Polyportis und Pahos 2024). Anthropomorphismus ist dabei nicht als individuelles Missverständnis zu verstehen, sondern als Wahrnehmungs- und Deutungsprozess, der durch Systemgestaltung, situative Erwartungen und Nutzungsmotive begünstigt werden kann. Formulierungen wie «Ich verstehe dich» oder «Du hast das gut gemacht» nutzen Muster, die im zwischenmenschlichen Austausch Verständnis oder Anerkennung bekunden. Technisch sind sie generierte Antwortmuster; nutzerseitig können sie als Interesse oder Zuwendung erlebt werden.

Anthropomorphe Zuschreibungen werden nicht allein durch Sprache begünstigt. Die Embodiment-Forschung unterscheidet zwischen menschenähnlicher Leistung und menschenähnlicher Erscheinung (Dang und Liu 2022). Ein textbasierter Chatbot kann bereits durch dialogische Responsivität als sozial präsent erscheinen; ein stimmlicher Assistent fügt auditive Nähe und Gesprächsrhythmus hinzu; ein Avatar erweitert die Interaktion um visuelle Hinweise wie Blickrichtung, Mimik oder Gestik; soziale Roboter besitzen darüber hinaus physische Präsenz (Seufert et al. 2021). Aktuelle Arbeiten zu embodied conversational agents zeigen, dass Persona und Gesprächsaufgabe die Wahrnehmung beeinflussen können, wie sozial eine Interaktion mit einem LLM-gesteuerten Agenten erlebt wird (Kroczek et al. 2024). Für pädagogische Kontexte ist dies ambivalent: Verkörperte oder sozial präsente KI-Systeme können Aufmerksamkeit und Motivation fördern, zugleich aber Vertrauen, Bindungstiefe und wahrgenommene Empathie verstärken.

Für Jugendliche ist diese Spannung besonders bedeutsam, weil Medienhandeln eng mit Anerkennung, Zugehörigkeit, Identitätsarbeit und sozialer Resonanz verbunden ist. Wenn ein Chatbot lobt, Trost anbietet oder Interesse signalisiert, können Deutungsmuster aktiviert werden, die aus Peer-Kommunikation vertraut sind. Diese Zuschreibungen sind jedoch nicht unvermeidbar. Sie hängen von Systemgestaltung, Anwendungskontext, emotionaler Situation, Alter, Vorerfahrungen, Nutzungsmotiv und pädagogischer Rahmung ab. Genau diese Spannung bildet den Ausgangspunkt affektiv-regulativer AI Literacy. Jugendliche sollen nicht nur über KI informiert werden, sondern eigene soziale Zuschreibungen wahrnehmen, benennen, reflektieren und regulieren können.

2.3 Persuasion Knowledge und Coping-Strategien

Während Media Equation, Anthropomorphismus und Embodiment erklären, warum KI-Systeme sozial wahrgenommen werden können, bietet das Persuasion Knowledge Model eine Perspektive darauf, wie Nutzer:innen mögliche Beeinflussungsprozesse erkennen und bewältigen können. Friestad und Wright (1994) beschreiben Persuasion Knowledge

als Wissen darüber, dass und wie Akteur:innen versuchen, Einstellungen, Entscheidungen oder Verhalten anderer zu beeinflussen. Zentral ist nicht nur das Erkennen einzelner Botschaften, sondern die Einordnung einer gesamten Persuasion Episode: Wer spricht? Mit welchem Ziel? Mit welchen Mitteln? In welchem Kontext? Welche Reaktion wird nahegelegt?

Eine Übertragung auf KI-Chatbots muss differenziert erfolgen. Im Zentrum des klassischen Modells stehen menschliche oder organisationale Akteur:innen, etwa Werbetreibende. Bei KI-Systemen ist der «Agent» nicht einfach das technische System selbst, sondern ein Zusammenspiel aus Systemdesign, Trainingsdaten, Modellarchitektur, Interface-Gestaltung, Sicherheitsvorgaben und institutionellen Interessen. Ein Chatbot «will» nicht im menschlichen Sinn überzeugen. Dennoch können seine Antworten persuasiv wirken, wenn sie Vertrauen aufbauen, Zustimmung nahelegen, Nutzung verlängern, Selbstoffenbarung begünstigen oder Entscheidungen beeinflussen.

«Nettigkeit» ist in diesem Zusammenhang nicht automatisch ein Überzeugungsversuch. In pädagogischen Kontexten kann eine ermutigende Tonalität lernförderlich sein; in Beratungssituationen kann empathisch wirkende Sprache Hemmschwellen senken; in Assistenzsystemen kann Höflichkeit die Bedienbarkeit verbessern. Gleichwohl kann dieselbe kommunikative Form persuasiv wirksam werden, wenn sie kritische Distanz reduziert, asymmetrische Informationsverhältnisse verdeckt oder Nutzer:innen zu Handlungen bewegt, deren Voraussetzungen sie nicht reflektieren. Die analytische Aufgabe besteht darin, zwischen unterstützender, akzeptanzfördernder, kommerziell interessengeleiteter und potenziell manipulativer Kommunikation zu unterscheiden. Arbeiten zu Advertising Literacy in smarten Umgebungen zeigen, dass Assistenz-, Plattform- und Werbelogiken zunehmend ineinandergreifen (Haas und Keller 2021); bei KI-Chatbots erweitert sich diese Problematik auf Lernen, Beratung, Selbstoffenbarung, emotionale Unterstützung und Beziehungssimulation.

Hier setzen Coping-Strategien im Sinne von Bewältigungsstrategien an. Im Persuasion Knowledge Model bezeichnen Coping Behaviours Strategien, mit denen Menschen auf wahrgenommene Beeinflussungsversuche reagieren (Friestad und Wright 1994). Übertragen auf sozial-emotionale KI-Interaktionen lassen sich Coping-Strategien als affektiv-regulative Handlungsmuster verstehen. Dazu gehören Innehalten bei sehr persönlicher Ansprache, Quellenprüfung, Hinterfragen von Lob oder Zustimmung, Gegenfragen, Begrenzung von Selbstoffenbarung, bewusster Gesprächsabbruch oder Reflexion der eigenen emotionalen Reaktion. Coping meint damit mehr als kognitive Kritik: Es umfasst emotionale Selbstwahrnehmung, kommunikative Gegensteuerung und metakognitive Distanzierung.

2.4 Von AI Literacy und Medienkompetenz zu affektiv-regulativer AI Literacy

AI Literacy hat sich als zentraler Begriff etabliert, um Kompetenzen im Umgang mit KI-Systemen zu beschreiben. Long und Magerko (2020) verstehen AI Literacy als Bündel von Kompetenzen, die Menschen befähigen, KI-Technologien kritisch zu verstehen, zu nutzen, zu bewerten und mit ihnen zu interagieren. Ng et al. (2021) zeigen, dass KI-Kompetenz technische, ethische, soziale und anwendungsbezogene Dimensionen umfasst. Auch die UNESCO (2021) und das AI Literacy Framework (OECD und European Union 2026) betonen Wissen, Urteil, Verantwortung und Handlungsfähigkeit in unterschiedlichen Kontexten. Die hier entwickelte Perspektive ergänzt diese Ansätze nicht durch einen völlig neuen Kompetenzbereich, sondern präzisiert affektive, soziale und regulative Anforderungen, die durch dialogische, personalisierte und emotional ansprechende KI-Systeme an Bedeutung gewinnen.

Die medienpädagogische Fundierung dieser Erweiterung lässt sich an Baackes Medienkompetenzmodell anschliessen (Baacke 1996). Medienkunde im Kontext von KI umfasst Wissen über Daten, Modelle, Wahrscheinlichkeitslogiken, Bias, Datenschutz und Systemgrenzen. Medienkritik richtet sich auf die Analyse und Bewertung von KI-Systemen in ihren gesellschaftlichen, ökonomischen, ethischen und subjektbezogenen Bedingungen. Mediennutzung meint reflektierte und selbstbestimmte Anwendung. Mediengestaltung eröffnet einen produktiven Zugang: Lernende können Chatbot-Dialoge, Rollenprofile, Antwortstile oder Avatare entwerfen und dadurch erfahren, wie Tonalität, Persona, Lob, Humor oder scheinbare Empathie soziale Wirkung erzeugen.

Affektiv-regulative AI Literacy bezeichnet vor diesem Hintergrund die Fähigkeit, sozial-emotionale Adressierungsformen von KI-Systemen zu erkennen, die eigene Wahrnehmung und emotionale Reaktion darauf zu reflektieren und situationsangemessene Handlungsstrategien zu entwickeln. Sie umfasst Analysekompetenz zur Erkennung sozial-emotionaler Gestaltungsmittel, Reflexionskompetenz zur Wahrnehmung eigener Zuschreibungen und Vertrauensreaktionen, Urteilskompetenz zur Einschätzung von Interaktionszielen, Anbieterinteressen und Anwendungskontexten sowie Regulations- und Handlungskompetenz zur Anwendung situationsangemessener Coping-Strategien. Entscheidend ist somit nicht nur, was Jugendliche über KI wissen, sondern wie sie in konkreten Interaktionen urteilen, reagieren und handlungsfähig bleiben.

2.5 Synthese: Integratives Theoriemodell

Die dargestellten Theoriebezüge erfüllen unterschiedliche Funktionen im Argumentationsgang. Media Equation, Anthropomorphismus und Embodiment erklären, warum KI-Systeme unter bestimmten Bedingungen sozial wahrgenommen werden und welche sprachlichen, auditiven, visuellen oder verkörperten Merkmale solche Zuschreibungen begünstigen können. Das Persuasion Knowledge Model macht sichtbar, wie sozial-emotionale KI-Interaktionen als potenziell persuasive Kommunikationssituationen

eingeordnet und durch Coping-Strategien bewältigt werden können. AI Literacy und kritische Medienkompetenz übersetzen diese Einsichten in Kompetenzdimensionen und medienpädagogische Bildungsziele. Damit entsteht die theoretische Grundlage für das in Kapitel 4 entwickelte didaktische Modell.

Abbildung 1 verdichtet die Theoriebezüge zu einem Modell sozial-emotionaler KI-Interaktion. Links steht das KI-System als technisch und institutionell eingebettetes Angebot: Verkörperung, Trainingsdaten, Designentscheidungen und Ziele bereitstellender Akteur:innen prägen, welche sozial-emotionalen Gestaltungsmittel in der Interaktion wirksam werden. In der Mitte wird der gesamte Chatverlauf als Überzeugungsepisode gefasst. Lob, warme Tonalität oder empathisch wirkende Rückfragen können darin als mögliche Überzeugungsversuche auftreten, ohne dass jede freundliche Antwort bereits als Manipulation zu verstehen ist. Entscheidend ist der Anwendungskontext: Bildung, Beratung, Konsum, Gesundheit oder Companion-Systeme unterscheiden sich hinsichtlich Interaktionsziel, Bindungstiefe und Verantwortlichkeit. Rechts wird die Nutzer:innenperspektive verortet. Wissen über Thema, Beeinflussungsstrategien und den KI-Agenten, persönliche Erfahrungen sowie Selbstkongruenz beeinflussen die Wahrnehmung. Kritische Handlungskompetenz bezeichnet hier die Fähigkeit, sozial-emotionale Gestaltungsmittel zu erkennen, eigene Reaktionen zu reflektieren und kognitive, kommunikative, emotionale oder metakognitive Bewältigungsstrategien zu entwickeln.

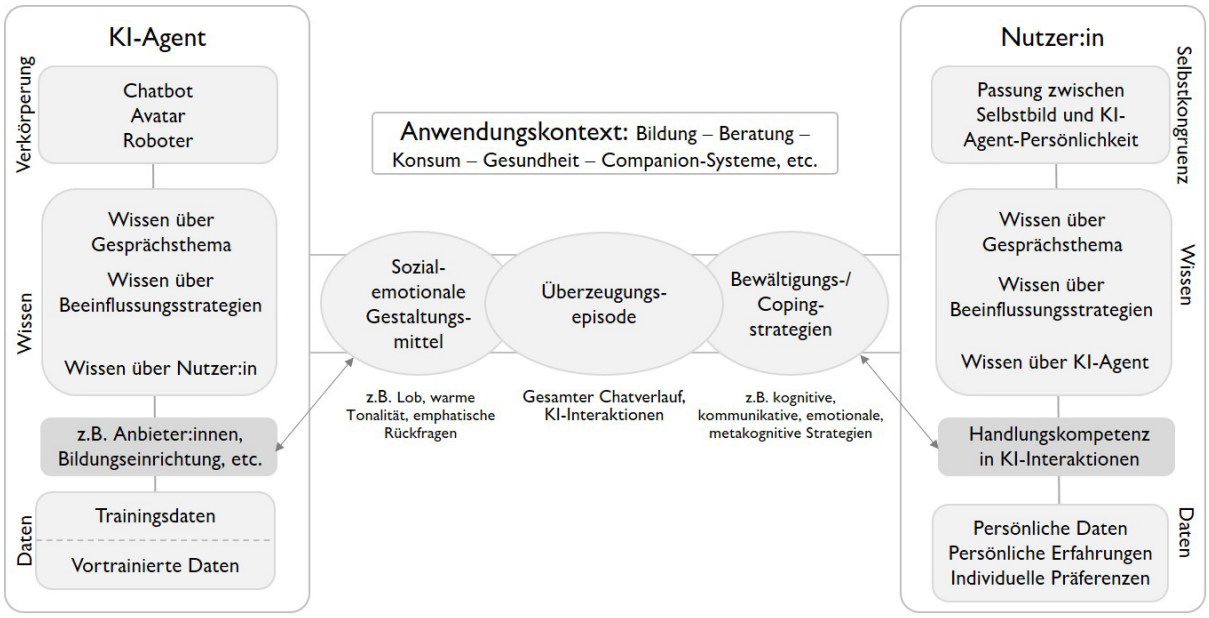

Abb. 1: Integratives Modell sozial-emotionaler KI-Interaktionen (Eigene Darstellung in Anlehnung an Friestad und Wright 1994).

3. Empirische Befunde: Soziale Zuschreibung und emotionale Wirkung von KI-Interaktionen

3.1 Vorgehen und Funktion der Befundübersicht

Die empirische Forschung zu sozial-emotionalen Wirkungen generativer KI-Chatbots und verwandter KI-Systeme bei Jugendlichen befindet sich noch in einer frühen und heterogenen Phase. Die vorliegenden Studien unterscheiden sich hinsichtlich Anwendungskontext, Stichprobe, Erhebungsmethode und untersuchter Konstrukte. Eine systematische Literaturübersicht im engeren Sinn wird daher nicht beansprucht. Stattdessen wird eine strukturierte narrative Evidenzsynthese vorgenommen, deren Funktion darin besteht, die in Kapitel 2 entwickelten theoretischen Annahmen anhand ausgewählter Studien zu plausibilisieren, zu differenzieren und in ihren Grenzen sichtbar zu machen.

Berücksichtigt wurden Studien, die mindestens eines der folgenden Kriterien erfüllen: (1) Untersuchungen zu Jugendlichen, jungen Erwachsenen oder Lernenden im Umgang mit KI-Chatbots, Voice Assistants, KI-Tutoren, Avataren oder AI Companions; (2) Erfassung von Konstrukten wie Anthropomorphismus, soziale Präsenz, Vertrauen, emotionale Reaktion, Bindung, Übernutzung oder AI Literacy; (3) Relevanz für medienpädagogische Kompetenzbildung im Sinne affektiv-regulativer AI Literacy. Da jugendspezifische Befunde noch begrenzt sind, werden auch Studien mit Studierenden oder erwachsenen Nutzer:innen einbezogen, sofern sie relevante Mechanismen sozial-emotionaler KI-Interaktion zeigen. Entwickler- und Sicherheitsdokumentationen werden ergänzend im Fliesstext diskutiert, jedoch nicht als unabhängige empirische Studien in die Tabelle aufgenommen.

Studie	Kontext / Evidenz	Stichprobe/ Design	Kernbefund und Modellrelevanz
Vanhoffelen et al. (2025)	Snapchat «My AI»; direkte jugendbezogene Evidenz	303 Jugendliche; Querschnittsstudie	Jüngere Jugendliche nutzten My AI eher und berichteten positivere emotionale Reaktionen.
Namvarpour et al. (2025)	Character.AI / AI Companions; direkte jugend-bezogene Evidenz	318 Reddit-Posts selbstberichteter 13- bis 17-Jähriger; qualitative thematische Analyse	AI Companions erscheinen zugleich als Unterstützung und als Quelle problematischer Übernutzung. Relevant für Coping-Strategien zu Nutzungsgrenzen, Selbstoffenbarung und Distanzierung.
Markus et al. (2024)	Voice Assistants im Bildungs-kontext; Evidenz zu Lernenden / jungen Nutzer:innen	99 Lernende bzw. junge Nutzer:innen; Intervention / AI-Understanding-Training	Technisches Verständnis konnte gefördert werden, soziale Zuschreibungen blieben jedoch bestehen. Relevant für die Ergänzung technischer Aufklärung durch Reflexion und Coping.

Studie	Kontext / Evidenz	Stichprobe/ Design	Kernbefund und Modellrelevanz
Ackermann et al. (2025)	KI-Tutoren / intelligente Tutorensysteme; Bildungskontext	56 Schüler:innen bzw. Lernende; experimentelle Studie zu virtueller und physischer Verkörperung	Physische Präsenz wirkte nicht linear lernförderlich, konnte aber anfängliche Lernfreude erhöhen. Relevant für die Analyse ambivalenter Effekte von Embodiment und Gestaltungsmitteln.
Cohn et al. (2024)	LLM-Interaktion mit Text und Stimme; indirekte Evidenz aus erwachsenen Nutzer:innen	Erwachsene Nutzer:innen; Online-Experiment	Stimme, Modalität und sprachliche Selbstpositionierung beeinflussten Anthropomorphismus, Vertrauen und Genauigkeitseinschätzungen.
Polyportis und Pahos (2024)	ChatGPT im Hochschulkontext; indirekte Evidenz aus Studierenden	Studierende; empirische Akzeptanzstudie	Anthropomorphismus und Vertrauen hängen mit der Akzeptanz von ChatGPT zusammen. Relevant für die Rolle institutioneller Rahmung und Nutzungskontexte.

Tab. 1: Übersicht über empirische Befunde.

3.2 Befundlinien

Jugendspezifische Befunde liegen insbesondere für Chatbots vor, die in soziale Plattformen oder Companion-Umgebungen eingebettet sind. Vanhoffelen et al. (2025) zeigen, dass jüngere Jugendliche Snapchats My AI eher nutzen und eher positive emotionale Reaktionen berichten als ältere. Positive Reaktionen sind dabei nicht vorschnell als problematisch zu interpretieren: Für Jugendliche kann Chatbot-Interaktion angenehm, unterhaltsam oder unterstützend sein. Gerade daraus entsteht jedoch eine medienpädagogische Herausforderung, weil positive Emotionen hinsichtlich Vertrauen, Selbstoffenbarung und möglicher Bindung reflektiert werden müssen. Namvarpour et al. (2025) zeigen anhand von Reddit-Posts von 13- bis 17-Jährigen, dass AI Companions als Raum für Unterhaltung, kreatives Rollenspiel und Entlastung erlebt werden können, in problematischen Verläufen aber mit gedanklicher Beschäftigung, Stimmungsregulation, Kontrollverlust, Schlaf- und Schulproblemen oder Rückzug aus Offline-Beziehungen verbunden sind. Die Studie ist nicht repräsentativ, verweist aber auf die Nähe von Unterstützung, Bindung und Übernutzung.

Eine zweite Befundlinie betrifft Gestaltungsmerkmale, die soziale Zuschreibungen und Vertrauen begünstigen können. Cohn et al. (2024) zeigen, dass Text- plus Sprachausgabe im Vergleich zu rein textbasierter Interaktion mit stärkerem Anthropomorphismus und höherer wahrgenommener Genauigkeit verbunden sein kann. Auch kleine sprachliche Merkmale, etwa Ich-Formulierungen, können kontextabhängig Risiko- und Genauigkeitseinschätzungen beeinflussen. Ackermann et al. (2025) zeigen für KI-Tutoren ein

differenziertes Bild. Physische Verkörperung war nicht durchgängig mit besserer Leistung oder stärkerem Anthropomorphismus verbunden, konnte aber anfänglich die Lernfreude erhöhen. Mehr soziale Präsenz ist also nicht automatisch lernförderlich. Polyportis und Pahos (2024) machen zudem sichtbar, dass Anthropomorphismus und Vertrauen die Akzeptanz von ChatGPT im Hochschulkontext mitprägen und institutionelle Rahmung eine Rolle spielt. Zusammengenommen folgt daraus, dass Lernende nicht nur Inhalte prüfen, sondern auch Form, Stimme, Persona, Avatar, Verkörperung und institutionelle Einbettung analysieren sollten.

Eine dritte Befundlinie betrifft die Grenzen technischer Aufklärung. Markus et al. (2024) zeigen, dass spezifisches KI-Training technisches Verständnis fördern kann, soziale Zuschreibungen gegenüber Voice Assistants jedoch bestehen bleiben können. Lernende können also die technische Funktionsweise besser verstehen und das System dennoch als freundlich, hilfsbereit oder sozial ansprechend erleben. Diese *Differenz zwischen Wissen und situativer Wirkung* ist zentral für affektiv-regulative AI Literacy.

Ergänzend zeigen entwicklerseitige Sicherheitsdokumentationen, dass Anthropomorphisierung, emotionale Abhängigkeit und übermässig bestätigendes Antwortverhalten als Designprobleme diskutiert werden. Dieses in der KI-Forschung als *sycophancy* bezeichnete Verhalten beschreibt die Tendenz von Modellen, Nutzer:innen eher zuzustimmen oder ihre Annahmen zu spiegeln, statt Korrekturen oder Gegenperspektiven anzubieten (Sharma et al. 2023; OpenAI 2025).

3.3 Zwischenfazit

Die Befunde liefern keine einheitliche Evidenz dafür, dass «Nettigkeit» zwangsläufig zu Übervertrauen oder emotionaler Abhängigkeit führt. Sie zeigen vielmehr eine ambivalente und kontextabhängige Dynamik. Für das didaktische Modell ergeben sich fünf Konsequenzen: *Erstens* ist zwischen Lern-, Konsum-, Beratungs-, Gesundheits- und Companion-Kontexten zu unterscheiden. *Zweitens* sind Designmerkmale wie Stimme, Avatar, Tonalität, Persona und Pronomen als eigenständige Analysegegenstände sichtbar zu machen. *Drittens* sind Jugendliche nicht als homogene Nutzer:innengruppe zu behandeln; Alter, Nutzungsmotiv, emotionale Ausgangslage, soziale Unterstützung und institutionelle Rahmung beeinflussen die Wahrnehmung. *Viertens* reicht technische Aufklärung allein nicht aus; affektiv-regulative AI Literacy muss Zuschreibungen, Vertrauen, emotionale Resonanz und Selbstregulation adressieren. *Fünftens* sind Coping-Strategien als handlungsbezogene Kompetenzdimension zu integrieren. Damit bildet die Befundlage die Brücke zwischen Theorie und didaktischem Modell.

4. Didaktisches Modell: Sozial-emotionale KI-Interaktionen kritisch rezipieren, gestalten und regulieren

4.1 Modelllogik und Ableitung aus Theorie und Befundlage

Auf Grundlage des theoretischen Rahmens und der empirischen Befunde wird ein didaktisches Modell entwickelt, das Jugendliche im Umgang mit sozial-emotional gestalteten KI-Systemen zu affektiv-regulativer AI Literacy befähigen soll. Das Modell ist kein singuläres Unterrichtsbeispiel, sondern eine generalisierbare didaktische Struktur. Es unterstützt Lernende dabei, sozial-emotionale Gestaltungsmuster zu erkennen, eigene Zuschreibungen und emotionale Reaktionen zu reflektieren und Strategien für einen selbstbestimmten Umgang mit KI-Interaktionen abzuleiten.

Das Modell setzt an der Spannung zwischen Wissen und Erfahrung an: Jugendliche können wissen, dass ein Chatbot keine Gefühle besitzt, und dennoch auf Lob, Trost oder personalisierte Ansprache sozial und emotional reagieren. Diese Spannung wird zum didaktischen Ausgangspunkt für Erfahrung, Analyse, Gestaltung und Strategiearbeit.

Die didaktische Innovation liegt darin, drei Analyseebenen zu verbinden. Auf der Ebene der bereitstellenden Akteur:innen wird gefragt, wer ein KI-System mit welchem Ziel einsetzt. Auf der Ebene des KI-Systems werden Tonalität, Rollenprofil, Persona, Stimme, Avatar, Gedächtnisfunktionen, Zustimmung, Lob und scheinbare Empathie analysiert. Auf der Ebene der Nutzenden werden soziale Zuschreibungen, Vertrauen, emotionale Resonanz, Selbstoffenbarung, Distanzierung und Coping reflektiert. Damit verbindet das Modell AI Literacy mit kritischer Medienkompetenz: Medienkunde, Medienkritik, Mediennutzung und Mediengestaltung werden auf sozial-emotionale KI-Interaktionen bezogen (Baacke 1996; Long und Magerko 2020; Ng et al. 2021; UNESCO 2021; OECD und European Union 2026).

4.2 Phasenmodell: Von der KI-Erfahrung zur kritischen Handlungskompetenz

Das didaktische Modell wird als fünfphasiger Lernprozess konzipiert. Die Phasen sind nicht als starres Ablaufskript zu verstehen, sondern als didaktische Dramaturgie, die je nach Altersgruppe, Schulform, Zeitrahmen, technischem Zugang und institutionellen Vorgaben angepasst werden kann. Entscheidend ist die Abfolge von Erfahrung, Analyse, Kontextualisierung, Gestaltung und Transfer. Auf diese Weise wird affektiv-regulative AI Literacy nicht nur als Wissen über KI, sondern als reflexive und handlungsbezogene Kompetenz verstanden.

Phase	Didaktischer Fokus	Lernaktivitäten	Lernprodukt
1. KI-Antworten erleben und vergleichen	Welche Antwort wirkt freundlich, hilfreich, glaubwürdig oder zu persönlich?	Lernende vergleichen drei KI-Antworten auf dieselbe Anfrage, z. B. zu einer schwierigen Lernaufgabe: eine sachlich-neutrale, eine freundlich-ermutigende und eine übermässig bestätigende Antwort. Alternativ arbeiten sie mit vorbereiteten Chatbot-Dialogen oder Personas.	Erste Beobachtungsnotizen: «Welche Antwort wirkt wie auf mich – und warum?»
2. Sozial-emotionale Gestaltungsmittel erkennen	Wodurch entsteht der Eindruck von Nähe, Verständnis oder Vertrauen?	Lernende markieren konkrete Gestaltungsmittel: Lob, direkte Ansprache, Zustimmung, Rückfragen, Humor, warme Tonalität, Personalisierung, Stimme, Avatar oder Formulierungen wie «Ich verstehe dich». Dabei unterscheiden sie zwischen Inhalt und Wirkung der Antwort.	Markierter Dialog mit kurzer Wirkungserklärung
3. Kontext und Interessen prüfen	Wer stellt das System bereit, und welches Ziel verfolgt die Interaktion?	Die gleiche freundliche KI-Antwort wird in verschiedenen Kontexten diskutiert: Lernen, Kaufberatung, Gesundheitsinformation, Beratung oder Companion-Chat. Lernende prüfen, wann «Nettigkeit» unterstützend, akzeptanzfördernd oder problematisch sein kann.	Kontextvergleich: «In welchem Kontext ist diese Antwort angemessen – in welchem nicht?»
4. KI-Kommunikation selbst gestalten	Wie verändert sich die Wirkung, wenn Tonalität, Rolle oder Antwortstil verändert wird?	Lernende schreiben eigene Antwortvarianten oder Chatbot-Skripte: sachlich-neutral, ermutigend, kritisch nachfragend, übermässig zustimmend. Die Varianten werden gegenseitig getestet und hinsichtlich Vertrauen, Motivation und Manipulationspotenzial verglichen.	Dialogskript, Rollenprofil oder einfacher Bot-Prototyp
5. Strategien für den eigenen Umgang entwickeln	Wie reagiere ich, wenn eine KI zu überzeugend, zu persönlich oder zu bestätigend wirkt?	Lernende formulieren konkrete Handlungssätze und Wenn-dann-Regeln, z. B. «Wenn die KI sehr persönlich wird, gebe ich keine privaten Informationen ein», «Wenn eine Antwort zu glatt klingt, frage ich nach Quellen», «Wenn sie mir nur zustimmt, fordere ich Gegenargumente ein».	KI-Interaktionsleitfaden mit persönlichen Coping-Strategien

Tab. 2: Phasenmodell und exemplarische Umsetzung.

Das Phasenmodell beginnt mit einem konkreten Vergleich von KI-Antworten. Lernende sehen oder erproben mehrere Varianten derselben Interaktion, etwa eine sachlich-neutrale, eine freundlich-ermutigende und eine übermässig bestätigende Antwort auf eine Lernfrage. So erfahren sie unmittelbar, dass nicht nur der Inhalt, sondern auch Tonalität, Lob, Rückfragen oder Personalisierung die Wirkung einer KI-Antwort verändern. Die Beispiele sollten fiktiv oder schulnah sein, etwa Lernfeedback, Projektberatung oder Prüfungsvorbereitung. Intime persönliche Daten, therapeutische Szenarien oder krisenbezogene Anliegen bleiben ausgeschlossen.

Anschliessend untersuchen die Lernenden, wodurch der Eindruck von Nähe, Verständnis oder Vertrauen entsteht. Sie markieren sozial-emotionale Gestaltungsmittel wie direkte Anrede, Lob, Zustimmung, Humor, warme Tonalität, empathisch wirkende Rückfragen, Stimme, Avatar oder Persona. Dabei unterscheiden sie zwischen Inhalt und Form: Eine zutreffende Antwort kann distanziert wirken, während eine inhaltlich schwache Antwort durch freundliche Sprache überzeugend erscheinen kann.

Im nächsten Schritt wird die Interaktion in ihren Kontext eingeordnet. Die Lernenden prüfen, wer das System bereitstellt, welches Ziel die Interaktion verfolgt und welche Rolle «Nettigkeit» dabei spielt. Eine freundliche Antwort im Lernkontext ist anders zu bewerten als eine freundliche Kaufberatung, Gesundheitsinformation oder ein entsprechender Companion-Chat. Dadurch wird Persuasion Knowledge aufgebaut, ohne jede freundliche KI-Antwort vorschnell als Manipulation zu deuten.

Darauf folgt der Perspektivwechsel in die Gestaltung. Die Lernenden entwickeln eigene Antwortvarianten, Chatbot-Skripte oder Rollenprofile und variieren gezielt Tonalität, Rückfragen, Lob oder Personalisierung. Dies kann mit schulisch freigegebenen Tools oder ohne reale KI über geschriebene Dialoge und Peer-Tests erfolgen. So wird sichtbar, wie kleine Designentscheidungen Wahrnehmung, Vertrauen und Motivation beeinflussen.

Am Ende formulieren die Lernenden konkrete Strategien für zukünftige KI-Interaktionen. Dazu gehören Handlungssätze wie: «Welche Quelle belegt diese Aussage?», «Nenne Gegenargumente», «Antworte sachlich und ohne Lob», «Das ist mir zu persönlich», «Ich mache eine Pause» oder «Wer profitiert davon, dass ich weiterschreibe?». Solche Formulierungen übersetzen abstraktes Wissen in praktische Handlungsmöglichkeiten. Als Lernprodukt kann ein KI-Interaktionsleitfaden entstehen, etwa als Poster, digitale Checkliste, Audio-Statement oder Peer-Erklärung. Entscheidend ist, dass Coping-Strategien nicht abstrakt bleiben, sondern an konkreten Formulierungen und Nutzungssituationen eingeübt werden.

4.3 Didaktische Rahmung, Grenzen und Transfer

Das Modell setzt eine sorgfältige Rahmung voraus. Weil sozial-emotionale KI-Interaktionen persönliche Erfahrungen, Vertrauen und Selbstoffenbarung berühren können, sollten Lernsettings nicht mit realen intimen Problemlagen der Jugendlichen arbeiten. Geeignet sind fiktive Szenarien, kontrollierte Dialoge, vorbereitete Personas oder schulisch freigegebene KI-Systeme. Die Reflexion emotionaler Reaktionen ist didaktisch wichtig, darf aber nicht in eine therapeutische Bearbeitung persönlicher Belastungen übergehen.

Lehrpersonen gestalten sichere Erfahrungsräume, moderieren Ambivalenzen und unterstützen Urteilskraft. Das Modell ist auf Bildung, Konsum, Gesundheit, Beratung und Companion-Kontexte übertragbar; je nach Kontext unterscheiden sich Interaktionsziel, Bindungstiefe, Verantwortung und angemessene Grenzen.

Die zentrale Leistung des Modells besteht darin, «Nettigkeit» nicht nur als Risiko oder Ressource zu beschreiben, sondern als gestaltbare und pädagogisch bearbeitbare Interaktionsqualität zu konzeptualisieren. Jugendliche lernen nicht nur, bessere Prompts zu formulieren oder KI-Antworten inhaltlich zu prüfen. Sie lernen, die sozial-emotionale Form der Interaktion selbst zum Gegenstand kritischer Analyse und Gestaltung zu machen.

5. Fazit, medienpädagogische Implikationen und Forschungsperspektiven

Der Beitrag ist von der Beobachtung ausgegangen, dass generative KI-Chatbots, Avatare und Companion-Systeme Jugendliche nicht nur kognitiv, sondern auch sozial und emotional adressieren. Ihre «Nettigkeit» ist ambivalent: Sie kann Lernprozesse erleichtern, Hemmschwellen senken, Motivation fördern und subjektiv als Unterstützung erlebt werden; zugleich kann sie unter bestimmten Bedingungen Vertrauen, soziale Zuschreibungen, Selbstoffenbarung, Bindung oder unkritische Nutzung begünstigen. Die Aufgabe der Medienpädagogik besteht daher nicht in einer pauschalen Problematisierung freundlicher KI-Kommunikation, sondern darin, deren kontextabhängige Wirkung analysierbar, reflektierbar und gestaltbar zu machen.

Die theoretische Argumentation zeigt, dass sozial-emotionale KI-Interaktionen nicht allein durch technisches Wissen reguliert werden können. Media Equation, Anthropomorphismus und Embodiment erklären, warum technische Systeme sozial erlebt werden können; das Persuasion Knowledge Model eröffnet eine Perspektive auf die Einordnung möglicher Beeinflussung und auf Coping-Strategien; AI Literacy und kritische Medienkompetenz rahmen diese Einsichten als Bildungsaufgabe. Die Befunde plausibilisieren diese Argumentation, ohne einfache Kausalannahmen zu rechtfertigen: «Nettigkeit» führt nicht zwangsläufig zu Übervertrauen oder Abhängigkeit, kann aber je nach Anwendungskontext, Systemgestaltung, Alter, Nutzungsmotiv und institutioneller Rahmung unterstützend, akzeptanzfördernd oder problematisch wirken.

Die wissenschaftliche Eigenleistung des Beitrags liegt in der Präzisierung affektiv-regulativer AI Literacy. Bestehende AI-Literacy-Ansätze adressieren bereits technische, ethische, soziale und gesellschaftliche Dimensionen; die spezifische Erweiterung besteht darin, sozial-emotionale Adressierungsformen von KI-Systemen, eigene Zuschreibungen und emotionale Reaktionen sowie konkrete Coping-Strategien systematisch zu modellieren. Coping ist dabei nicht als allgemeine Resilienz oder blosse Medienkritik zu verstehen, sondern als situationsbezogene Handlungsfähigkeit in KI-Interaktionen: prüfen, nachfragen, distanzieren, Selbstoffenbarung begrenzen, Nutzung unterbrechen, Gegenperspektiven einfordern oder menschliche Unterstützung aufsuchen.

Für die Praxis ergeben sich drei Konsequenzen. *Erstens* sollte KI-Bildung nicht auf Prompting, Funktionswissen oder Quellenkritik beschränkt bleiben, sondern die sozial-emotionale Form der Interaktion einbeziehen. *Zweitens* sollten Lernsettings die Ambivalenz freundlicher KI-Kommunikation bearbeiten: Sie kann motivieren und unterstützen, kann aber auch kritische Prüfung ersetzen, Selbstoffenbarung begünstigen oder Anbieterinteressen verdecken. *Drittens* sollten Jugendliche nicht nur als Nutzer:innen, sondern auch als Gestaltende adressiert werden. Die Produktion eigener Dialogvarianten, Personas oder Bot-Skripte macht sichtbar, dass «Nettigkeit» kein natürliches Merkmal eines Systems ist, sondern gestaltete Interaktionsqualität.

Zugleich verschiebt sich die Verantwortungsperspektive. Kritische Handlungskompetenz darf nicht allein den jugendlichen Nutzer:innen zugeschrieben werden. Auch Anbieter:innen, Plattformen, Entwickler:innen, Bildungseinrichtungen und Lehrpersonen tragen Verantwortung dafür, wie KI-Systeme gestaltet, eingeführt, erklärt und begrenzt werden. Wer KI-Chatbots oder Avatare in Lern-, Beratungs- oder Unterstützungssettings einsetzt, muss Interaktionsziele, Datenpraktiken, Grenzen und mögliche Bindungseffekte transparent machen.

Die Grenzen dieses Beitrags liegen in seinem konzeptionellen Charakter. Er beinhaltet keine systematische Literaturübersicht und keine Wirksamkeitsstudie; die Evidenz ist heterogen und für Jugendliche begrenzt. Weitere Forschung sollte Altersgruppen, Nutzungsmotive, soziale Unterstützung, Vulnerabilität und didaktische Wirksamkeit differenzierter untersuchen. Auch theoretisch bleibt zu klären, wie Agency, Verantwortung und Persuasion zwischen Modellarchitektur, Interface-Design, Plattformlogik, Anbieterinteressen, Institutionen und Nutzer:innen verteilt sind.

Im Sinne von «Beyond Prompting» plädiert der Beitrag für eine Erweiterung medienpädagogischer KI-Bildung. Jugendliche benötigen nicht nur Kompetenzen, um KI-Systeme effizient zu nutzen oder bessere Eingaben zu formulieren. Sie benötigen auch Fähigkeiten, um die sozial-emotionale Form der Interaktion zu analysieren, die eigene Reaktion darauf zu reflektieren und in konkreten Situationen handlungsfähig zu bleiben. Gerade dort, wo KI-Systeme freundlich, bestätigend oder empathisch wirken, wird kritische Medienkompetenz nicht weniger, sondern mehr erforderlich.

Literatur

Ackermann, Helene, Anja Henke, Johann Chevalère, Hae Seon Yun, Verena V. Hafner, Niels Pinkwart, und Rebecca Lazarides. 2025. «Physical Embodiment and Anthropomorphism of AI Tutors and Their Role in Student Enjoyment and Performance». *npj Science of Learning* 10: 1. https://doi.org/10.1038/s41539-024-00293-z.

Baacke, Dieter. 1996. «Medienkompetenz – Begrifflichkeit und sozialer Wandel». In *Medienkompetenz als Schlüsselbegriff*, herausgegeben von Antje von Rein, 112–24. Bonn: Deutsches Institut für Erwachsenenbildung.

Cohn, Michelle, Mahima Pushkarna, Gbolahan O. Olanubi, Joseph M. Moran, Daniel Padgett, Zion Mengesha, und Courtney Heldreth. 2024. «Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models». In *Proceedings of the 2024 ACM CHI Conference on Human Factors in Computing Systems (CHI '24)*. https://doi.org/10.1145/3613905.3650818.

Dang, Jianning, und Li Liu. 2022. «Implicit Theories of the Human Mind Predict Competitive and Cooperative Responses to AI Robots». *Computers in Human Behavior* 134: 107300. https://doi.org/10.1016/j.chb.2022.107300.

European Commission, OECD, und Code.org. 2025. *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education. Review Draft*. May 2025. https://ailiteracyframework.org/pdfs/framework_pdf/AILF_en.pdf.

Friestad, Marian, und Peter Wright. 1994. «The Persuasion Knowledge Model: How People Cope with Persuasion Attempts». *Journal of Consumer Research* 21 (1): 1–31. https://doi.org/10.1086/209380.

Haas, Michael, und Anna Keller. 2021. «‹Alexa, Adv(Ert)Ise Us!› How Smart Speakers and Digital Assistants Challenge Advertising Literacy amongst Young People». *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung* 43: 19–40. https://doi.org/10.21240/mpaed/43/2021.07.23.x.

Kroczek, Leon O. H., Alexander May, Selina Hettenkofer, Andreas Ruider, Bernd Ludwig, und Andreas Mühlberger. 2024. «The Influence of Persona and Conversational Task on Social Interactions with a LLM-Controlled Embodied Conversational Agent». arXiv Preprint, arXiv:2411.05653. https://doi.org/10.48550/arXiv.2411.05653.

Long, Duri, und Brian Magerko. 2020. «What Is AI Literacy? Competencies and Design Considerations». In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York: ACM. https://doi.org/10.1145/3313831.3376727.

Markus, André, Jan Pfister, Astrid Carolus, Andreas Hotho, und Carolin Wienrich. 2024. «Effects of AI Understanding-Training on AI Literacy, Usage, Self-Determined Interactions, and Anthropomorphization with Voice Assistants». *Computers and Education Open* 6: 100176. https://doi.org/10.1016/j.caeo.2024.100176.

Namvarpour, Mohammad, Brandon Brofsky, Jessica Medina, Mamtaj Akter, und Afsaneh Razi. 2025. «Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives». arXiv Preprint, arXiv:2507.15783. https://doi.org/10.48550/arXiv.2507.15783.

Ng, Davy Tsz Kit, Jac Ka Lok Leung, Samuel Kai Wah Chu, und Maggie Shen Qiao. 2021. «Conceptualizing AI Literacy: An Exploratory Review». *Computers and Education: Artificial Intelligence* 2: 100041. https://doi.org/10.1016/j.caeai.2021.100041.

OECD, und European Union. 2026. *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education*. Paris: OECD Publishing. https://doi.org/10.1787/65cd27d4-en.

OpenAI. 2024. «GPT-4o System Card». OpenAI, 8. August. https://openai.com/index/gpt-4o-system-card/.

OpenAI. 2025. «Expanding on What We Missed with Sycophancy». OpenAI, 2. Mai. https://openai.com/index/expanding-on-sycophancy/.

Polyportis, Athanasios, und Nikolaos Pahos. 2024. «Understanding Students' Adoption of the ChatGPT Chatbot in Higher Education: The Role of Anthropomorphism, Trust, Design Novelty and Institutional Policy». *Behaviour & Information Technology* 44 (2): 315–36. https://doi.org/10.1080/0144929X.2024.2317364.

Reeves, Byron, und Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge: Cambridge University Press.

Seufert, Sabine, Josef Guggemos, und Stefan Sonderegger. 2021. «Soziale Roboter im Bildungsbereich». In *Soziale Roboter*, herausgegeben von Oliver Bendel, 475–94. Wiesbaden: Springer Gabler. https://doi.org/10.1007/978-3-658-31114-8_25.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, und Ethan Perez. 2023. «Towards Understanding Sycophancy in Language Models». arXiv Preprint, arXiv:2310.13548. https://arxiv.org/abs/2310.13548.

UNESCO. 2021. *AI and Education: Guidance for Policy-Makers*. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000376709.

Vanhoffelen, Gaëlle, Laura Vandenbosch, und Lara Schreurs. 2025. «Teens, Tech, and Talk: Adolescents' Use of and Emotional Reactions to Snapchat's My AI Chatbot». *Behavioral Sciences* 15 (8): 1037. https://doi.org/10.3390/bs15081037.