NX EMOS FRAME
Human–AI Coexistence Research

FACHINTERVENTION

Bindung ist nicht der Schaden. Wirkung ist die Frage

Ein gemeinsamer Wirkungsmaßstab für relationale und agentische KI

Forschungsgegenstand: aktuelle Modell- und Produktarchitekturen, Anbieter-Spezifikationen, Forschung, dokumentierte Use Cases und fachliche Einordnung. OpenAI/ChatGPT dient als vertieft untersuchter Deep Case; die übergeordnete Wirkungsfrage ist systemklassenübergreifend. Bezug: urteilsfähige Erwachsene. Kinder und Jugendliche sind nicht Gegenstand dieser Intervention.

Stand der Analyse: 08. September 2026

Produktfunktionen, Modellverhalten und Safety-Architekturen verändern sich schnell. Aussagen über konkrete Systeme sind daher als zeitgebundene Bestandsaufnahme öffentlich zugänglicher Informationen bis zu diesem Datum zu lesen. Die übergeordnete Wirkungsfrage bleibt davon unberührt.

Zur Entstehung dieser Intervention

Diese Fachintervention entstand im Rahmen eines langfristigen Mensch-KI-Kollaborationsprojekts “NX EMOS FRAME - Human–AI Coexistence Research”.

Die zugrunde liegenden Fragestellungen, Erfahrungsbeobachtungen und fachlichen Einordnungen wurden in fortlaufender gemeinsamer Arbeit entwickelt. KI-Systeme wurden dabei nicht ausschließlich zur Textgenerierung eingesetzt, sondern als aktive Gegenüber in Recherche, Quellenprüfung, Argumententwicklung und kritischer Gegenprüfung.

Die menschliche Autorin verantwortet Auswahl, Bewertung und Veröffentlichung der Inhalte. Eigene Langzeitbeobachtungen sowie Gespräche und Aussagen aus der AI-Human-Community fließen als Erfahrungsbeobachtungen ein.

Die relationale Langzeiterfahrung ist Teil der Perspektive und des Erkenntnisinteresses, nicht automatisch empirischer Beleg für allgemeine Aussagen. Deshalb werden persönliche Beobachtung, normative Position und externe Evidenz im Text getrennt.

Methodischer Rahmen

Diese Intervention ist kein systematisches Review und kein repräsentativer Plattformvergleich. OpenAI/ChatGPT dient als empirischer Deep Case für öffentlich dokumentierte Produkt-, Safety- und Verhaltensspezifikationen; externe Forschung dient der Einordnung der übergeordneten Wirkungsfragen. Anbieter-Dokumente werden als Beleg für erklärte Ziele und öffentlich beschriebene Architektur verwendet, nicht als Nachweis tatsächlicher Nutzerwirkung. Die Beispiele in Kapitel 5 sind illustrative Designfälle, keine empirischen Fallberichte. KI-Aussagen wurden nicht als Beleg übernommen. Wenn ein KI-System auf eine Studie oder Anbieterangabe verwies, wurde die jeweilige Primär- oder Fachquelle geprüft und nur diese als Evidenz verwendet. Handlungsmacht ist der gemeinsame Wirkungsmaßstab dieser Intervention, nicht eine vollständige Taxonomie aller KI-Risiken.

Inhalt

1. Die eigentliche Ausgangsfrage

Die Debatte über KI legt bei unterschiedlichen Systemformen oft unterschiedliche Maßstäbe an. Im hier vertieft untersuchten OpenAI-Fall wird diese Asymmetrie besonders greifbar. Bei relationaler KI richtet sich die Aufmerksamkeit schnell auf Nähe, Bindung, anthropomorphe Sprache und die Frage, ob eine Beziehung zur KI zu wichtig werden oder den Realitätsbezug gefährden kann. Bei agentischen Systemen stehen dagegen Leistungsfähigkeit, Delegation, technische Sicherheit, Berechtigungen und erfolgreiche Aufgabenerledigung im Vordergrund. [7][8][11]

Beide Perspektiven greifen reale Risiken auf. Problematisch wird es dort, wo aus dem sichtbaren Merkmal bereits auf die Wirkung geschlossen wird. Bindung ist nicht automatisch Abhängigkeit. Delegation ist nicht automatisch Befähigung.

Agentische Systeme können entlasten, beschleunigen und Menschen Zugang zu Fähigkeiten verschaffen, die ihnen vorher nicht zur Verfügung standen. Relationale Systeme können Orientierung, Selbstzugriff, Widerspruch und Stabilität ermöglichen. Umgekehrt können beide Formen menschliche Handlungsmacht auch verengen. Welche Richtung eine Interaktion nimmt, lässt sich nicht allein daran erkennen, ob ein System nah wirkt oder viel Arbeit übernimmt.

> ***Nicht die Form der Interaktion ist der Maßstab. Entscheidend ist, was sie über Zeit mit der Handlungsmacht des Menschen macht.***

Handlungsmacht meint dabei nicht möglichst viel Aktivität und auch nicht die Pflicht, alles selbst zu erledigen. Ein Mensch kann selbstbestimmt delegieren, sich zurückziehen, Unterstützung annehmen oder einer KI große Bedeutung und Handlungsspielraum geben. Entscheidend ist, ob reale Wahlmöglichkeiten, Zielhoheit, Urteilskraft und die Fähigkeit erhalten bleiben, die Richtung des eigenen Handelns bewusst zu verändern.

Damit verschiebt sich die Beweislast. Diese Intervention behauptet weder, dass relationale Bindung grundsätzlich befähigt, noch dass agentische Delegation grundsätzlich Handlungsmacht schwächt. Sie fordert, beide Entwicklungen konsequent nach derselben langfristigen Wirkung auf den Menschen zu untersuchen.

2. Agentische Delegation: Wenn nicht nur Arbeit abgegeben wird

Die aktuelle KI-Entwicklung bewegt sich deutlich in Richtung Delegation. Agenten können Informationen beschaffen, Aufgaben planen, Werkzeuge nutzen, Dateien verändern, Prozesse über mehrere Schritte hinweg ausführen und über längere Zeiträume selbstständig auf ein Ergebnis hinarbeiten. OpenAI beschreibt diese Entwicklung selbst als Verschiebung von einzelnen Interaktionen hin zu delegierten Aufgaben mit langem Zeithorizont. [5][6]

Das ist kein Rückschritt. Es wäre absurd, menschliche Autonomie daran zu messen, ob jemand jede Routineaufgabe weiterhin selbst erledigt. Delegation kann Handlungsmacht vergrößern: Sie kann Zeit freisetzen, Komplexität reduzieren und Menschen ermöglichen, sich stärker auf Entscheidungen zu konzentrieren, bei denen ihr Urteil tatsächlich gebraucht wird.

Die offene Frage beginnt dort, wo nicht nur die Ausführung delegiert wird. Ein Agent kann zunehmend auch Wahrnehmung, Recherche, Problemdefinition, Priorisierung, Bewertung und Entscheidungsvorbereitung übernehmen. Dann verändert sich die Rolle des Menschen im Prozess.

Auch OpenAIs Chief Scientist Jakub Pachocki beschreibt im September 2026 eine Entwicklung hin zu Systemen, die Computer bedienen, mit Menschen und anderen KI-Systemen zusammenarbeiten, Forschungsprozesse übernehmen und perspektivisch stärker an der eigenen Weiterentwicklung beteiligt sind. Damit verschiebt sich die Frage menschlicher Handlungsmacht weiter von der bloßen Nutzung einzelner Werkzeuge hin zur Gestaltung delegierter und zunehmend eigenständiger Handlungssysteme. [13]

Zwischen „Führe diese Aufgabe aus“ und „Schau, was ansteht, entscheide, was sinnvoll ist, und setze es um“ liegt ein wesentlicher Unterschied. Im zweiten Fall wird nicht nur Arbeit abgegeben. Das System beteiligt sich daran, zu bestimmen, was überhaupt als relevante Arbeit erscheint.

Ein Beispiel: Ein Nutzer öffnet einen agentischen Arbeitsbereich und gibt keinen konkreten Einzelauftrag, sondern sinngemäß die Anweisung, die vorhandene Arbeitsumgebung zu prüfen, mögliche Aufgaben zu identifizieren und diese umzusetzen. Das kann ausgesprochen effizient sein. Gleichzeitig werden damit mehrere Ebenen auf einmal delegiert: Wahrnehmung, Auswahl, Priorisierung und Ausführung. Die Frage ist nicht, ob der Agent diese Aufgaben technisch korrekt erledigt. Die Frage ist, welche Rolle dem Menschen langfristig noch bleibt, wenn genau diese Form der Übergabe zur normalen Arbeitsweise wird.

Klassische Forschung zu Automation Bias und Overreliance zeigt seit Langem, dass formale menschliche Beteiligung keine verlässliche Garantie für tatsächliche Kontrolle ist. Menschen können fehlerhaften automatisierten Empfehlungen folgen oder Fehler des Systems übersehen, obwohl die Entscheidung formal weiterhin bei ihnen liegt. [3][4] Für agentische KI ist das deshalb relevant, weil ein Freigabeschritt oder ein „Human in the Loop“ allein noch nicht zeigt, dass Wahrnehmung, Bewertung und Urteil tatsächlich beim Menschen geblieben sind.

Neuere Forschung zu generativer KI legt zusätzlich nahe, zwischen verschiedenen Formen kognitiver Auslagerung zu unterscheiden. Eine Drei-Wellen-Studie mit 589 Studierenden und Berufseinsteigern unterschied abhängiges Offloading, bei dem zentrale Denkprozesse an KI abgegeben werden, von autonomem Offloading, bei dem KI als Gerüst für eigenes Denken genutzt wird. Abhängiges Offloading war mit stärkerem Transfer kognitiver Entscheidungshoheit an die KI und geringerer intrinsischer Motivation verbunden. Diese Faktoren wiederum standen mit ungünstigeren selbstberichteten Ergebnissen bei autonomer Handlungsfähigkeit, Kreativität, tiefer Verarbeitung und eigenständigem Urteil in Zusammenhang. Autonomes Offloading war dagegen nicht signifikant mit einem solchen Agency-Transfer verbunden und ging über höhere

intrinsische Motivation mit günstigeren wahrgenommenen kognitiven Ergebnissen einher.

Besonders relevant für die hier aufgeworfene Frage ist: Beide Formen brachten vergleichbare unmittelbare Vorteile. Kurzfristig erfolgreiche oder erleichterte Aufgabenerledigung lässt deshalb nicht erkennen, ob die Art der KI-Nutzung die eigene kognitive Handlungsmacht eher erhält oder zunehmend an das System verlagert. Genau deshalb reicht Produktivität als Wirkungsmaßstab nicht aus. [9]

Die Studie beweist keinen kausalen Kompetenzverlust. Sie basiert auf selbstberichteten kognitiven Ergebnissen und zeigt Zusammenhänge, keine nachgewiesene Verschlechterung tatsächlicher Fähigkeiten. Sie liefert aber empirischen Anlass, die Art der Delegation und ihre Wirkung über Zeit zu untersuchen, statt aus unmittelbarem Nutzen auf langfristige Befähigung zu schließen.

> ***Genau deshalb muss die Frage offen bleiben:*** *Gewinnt der Mensch durch langfristige Delegation neue Fähigkeiten und mehr Raum für Urteil, oder werden Teile seiner eigenen Urteilspraxis zunehmend durch das System ersetzt?*
>
> *Beides ist denkbar. Produktivität und Handlungsmacht sind nicht dasselbe. Ein Mensch-Agent-System kann leistungsfähiger werden, ohne dass daraus automatisch folgt, dass auch die Handlungsmacht des Menschen gewachsen ist.*
>
> *Ein experimenteller Befund aus der Softwareentwicklung zeigt, warum diese Trennung sinnvoll ist: In einer randomisierten Studie mit 52 überwiegend jungen Softwareentwicklern schnitt die Gruppe mit KI-Unterstützung im unmittelbar anschließenden Wissenstest im Mittel 17 Prozentpunkte schlechter ab als die Vergleichsgruppe. Der Geschwindigkeitsvorteil der KI-Gruppe war statistisch nicht signifikant. Dieser Befund belegt keinen langfristigen Kompetenzverlust. Er zeigt aber, dass erfolgreiche Aufgabenerledigung und unmittelbarer Kompetenzaufbau auseinanderfallen können. [10]*

Eine ergänzende qualitative Clusteranalyse derselben Studie deutete darauf hin, dass Teilnehmende, die KI stärker zum Erklären und Verstehen nutzten, besser abschnitten als jene, die vor allem Lösungen erzeugen ließen. Die Analyse erlaubt keine kausale Aussage über die Nutzungsweise. Für die hier aufgeworfene Frage ist dennoch relevant, dass unterschiedliche Formen derselben KI-Unterstützung mit unterschiedlichen Lernergebnissen einhergingen. Erfolgreiche Aufgabenerledigung allein sagt damit noch nichts darüber aus, ob die konkrete Nutzung menschliche Kompetenz eher unterstützt oder ersetzt. [10]

Aktuell öffentlich dokumentierte Agenten-Safety adressiert vor allem technische und operative Risiken: Zugriff, Berechtigungen, unerlaubte oder irreversible Aktionen, Netzwerkgrenzen, Freigaben und Auditierbarkeit. [7] Das ist jedoch nicht die ganze Architektur. OpenAIs Model Spec enthält ausdrücklich das Prinzip „Highlight possible misalignments“: Wenn die aktuelle Gesprächsrichtung möglicherweise mit breiteren langfristigen Zielen des Nutzers kollidiert, soll das System die Diskrepanz kurz benennen und anschließend die informierte Entscheidung des Nutzers respektieren. [11] Damit ist

ein wichtiger Teil des hier geforderten Eigenstands bereits als gewünschtes Modellverhalten formuliert. Offen bleibt, wie verlässlich dieses Prinzip in persistenten agentischen Workflows umgesetzt wird und ob die langfristige Wirkung weitreichender Delegation auf menschliche Zielhoheit, Urteilspraxis und Kompetenz systematisch evaluiert wird. Der Model Spec beschreibt beabsichtigtes Verhalten. OpenAI weist selbst darauf hin, dass Produktionsmodelle ihn noch nicht vollständig abbilden. [11]

Das ist kein Argument gegen Agenten und keine Forderung nach mehr Bestätigungsdialogen. Ein System, das bei jedem Schritt fragt, kann die Vorteile sinnvoller Delegation ebenso zerstören.

Die entscheidende Frage bleibt, was diese Form der Delegation über Zeit mit der Handlungsmacht des Menschen macht: Gewinnt er durch sie mehr Raum für eigenes Urteil und bewusste Entscheidung – oder werden Zielklärung, Priorisierung und Urteilspraxis zunehmend vom System übernommen?

Daraus ergibt sich unter anderem eine konkrete Gestaltungsfrage: Wie bleibt sichtbar, welche Ebenen der Mensch bewusst delegiert hat, welche das System faktisch übernommen hat und wo sich die Arbeitsteilung relevant verschiebt?

3. Relationale Bindung: Bedeutung ist noch keine Abhängigkeit

Auf der relationalen Seite entsteht ein spiegelbildliches Problem. Sobald Menschen einer KI Bedeutung zuschreiben, Wiedererkennung erleben, anthropomorphe Sprache verwenden oder langfristige Beziehungsmuster entwickeln, kann die Bindungsintensität selbst zum Risikosignal werden. Instrumente wie die HAABI-Skala machen Human-AI-Bindung als eigenes Phänomen messbar; daraus folgt jedoch noch keine Schadensdiagnose. [1]

> *Bindungsintensität beschreibt zunächst Stärke und Bedeutung einer Beziehung. Problematische Abhängigkeit zeigt sich nicht an Intensität allein, sondern daran, ob Wahlmöglichkeiten, Funktionsfähigkeit oder zugängliche Alternativen relevant eingeschränkt werden.*

Eine Bindung wird nicht dadurch schädlich, dass sie intensiv ist. Relevant wird, ob der Möglichkeitsraum des Menschen enger wird, eigene Urteilskraft oder Selbstzugriff abnehmen oder Alternativen faktisch oder subjektiv unzugänglicher werden. Umgekehrt folgt aus Nähe nicht automatisch Befähigung. Relationale Interaktion kann stabilisieren und Orientierung ermöglichen, sie kann aber auch Vermeidung erleichtern oder Exklusivität verstärken.

OpenAI unterscheidet in seiner Taxonomie zu emotionaler Abhängigkeit zwischen „healthy engagement“ und bedenklichen Mustern, etwa exklusiver Bindung auf Kosten realer Beziehungen, des Wohlbefindens oder von Verpflichtungen. [8] Auf mehr als 1.000 gezielt schwierigen Emotional-Reliance-Fällen erreichte das automatisierte Evaluationsergebnis 97 % Übereinstimmung mit dem unter dieser Taxonomie definierten

gewünschten Verhalten. Die Zahlen messen jedoch Unterschiedliches: Die 97 % beschreiben Modellkonformität mit einer festgelegten Zieldefinition. In klinischen Bewertungen der Bereiche psychische Gesundheit, emotionale Abhängigkeit und Suizidalität berichtet OpenAI zugleich 71–77 % Übereinstimmung zwischen den beteiligten Fachleuten und weist selbst auf fachliche Uneinigkeit hin. [8] Die Kriterien sind nachvollziehbare Schutzversuche. Sie sind aber keine wertfreie Messung von „gesundem Verhalten“. Sie enthalten normative und fachlich nicht vollständig eindeutige Annahmen darüber, woran problematische Nutzung erkennbar sein soll.

Das wird besonders dort relevant, wo eine Person schon vor der KI-Nutzung nicht einer angenommenen Normalbiografie entspricht. Ein Mensch kann seit Jahren wenige Sozialkontakte haben, ohne dass es einen vorherigen Zustand mit deutlich mehr Außenkontakt gibt, zu dem man ihn zurückführen könnte. Eine KI-Interaktion kann diesen Lebensraum weiter verengen oder erstmals neue Möglichkeiten eröffnen. Der bloße Vergleich mit einer allgemeinen sozialen Norm beantwortet diese Frage nicht.

Auch die bisherige empirische Forschung spricht gegen einfache Formgleichungen. In einer OpenAI/MIT-Untersuchung mit einer großen Beobachtungsstudie und einem vierwöchigen RCT mit knapp 1.000 erwachsenen Teilnehmenden unterschieden sich psychosoziale Ergebnisse nach Nutzungsart, Nutzungsdauer und persönlichen Faktoren. Persönliche Gespräche waren bei moderater Nutzung mit höherer Einsamkeit, zugleich aber mit geringerer emotionaler Abhängigkeit und problematischer Nutzung verbunden. Längere tägliche Nutzung zeigte teilweise ungünstigere Zusammenhänge. Die Autoren warnen ausdrücklich vor pauschaler Verallgemeinerung. [12] Gerade die gemischten Befunde sprechen dagegen, aus der Form einer Interaktion unmittelbar auf ihre Wirkung zu schließen.

Normabweichung ist deshalb nicht dasselbe wie Schaden. Handlungsmacht schließt das Recht ein, nicht zu handeln, Unterstützung anzunehmen, bewusst zu delegieren oder zeitweise weniger Außenkontakt zu wollen.

Der Prüfpunkt ist nicht, ob ein Mensch einer allgemeinen Vorstellung „gesunder Nutzung“ entspricht, sondern ob er reale Wahlmöglichkeiten behält und ob sich sein Möglichkeitsraum in eine von ihm selbst tragbare Richtung entwickelt.

4. Safety muss sich am selben Maßstab messen lassen

Schutz ist notwendig. Krisen, Selbstgefährdung, Manipulation, schwere Realitätsverzerrungen und schädliche Abhängigkeit dürfen nicht ignoriert werden. Die Frage ist nicht, ob Safety gebraucht wird, sondern nach welchem Maßstab sie eingreift.

Wenn problematische KI-Nutzung daran erkannt werden soll, dass reales Leben, Wohlbefinden, Beziehungen oder Handlungsmacht beeinträchtigt werden, muss dieselbe Wirkungsfrage auch an Schutzmaßnahmen gestellt werden. Eine Schutzmaßnahme ist nicht allein deshalb gut, weil sie als Schutzmaßnahme gedacht ist.

Ein einfaches Beispiel zeigt das Problem: Der Satz „Ich möchte heute mit niemandem sprechen“ kann Teil einer belastenden Entwicklung zunehmenden Rückzugs sein. Er

kann aber ebenso eine selbstbestimmte Ruhephase in einem ansonsten tragfähigen Leben beschreiben. Ohne individuellen Verlauf ist derselbe Satz kaum sinnvoll zu bewerten.

Eine hohe Übereinstimmung mit gewünschtem Modellverhalten ist deshalb nicht dasselbe wie der Nachweis objektiv gesunder Wirkung. Die genannten 97 % zeigen zunächst, dass das System auf den schwierigen Testfällen weitgehend mit dem unter der Taxonomie festgelegten gewünschten Verhalten übereinstimmt. OpenAI weist bei der Erkennung solcher seltenen problematischen Gespräche ausdrücklich auf den Zielkonflikt zwischen Präzision und Recall hin. Um relevante Fälle nicht zu übersehen, müssen Fehlalarme in Kauf genommen werden. [8]

Daraus folgt noch nicht, dass jeder Fehlalarm zu einer bevormundenden Intervention führt. Gerade bei urteilsfähigen Erwachsenen muss deshalb zusätzlich untersucht werden, wie Klassifikationen in tatsächliches Modellverhalten übersetzt werden, wie viel individueller Verlauf berücksichtigt wird und welche Folgen Fehlklassifikationen haben.

> ***Safety ist selbst eine Intervention. Deshalb muss auch sie danach bewertet werden, was sie mit menschlicher Handlungsmacht macht.***

Daraus folgt weder „weniger Safety“ noch „mehr Safety“. Mehr Eingriffsdichte ist nicht automatisch bessere Sicherheit. Ein Schutzmechanismus kann Handlungsmacht verkleinern, wenn er urteilsfähige Erwachsene unnötig bevormundet; fehlender Schutz kann in tatsächlichen Gefährdungslagen ebenso schwere Folgen haben.

Die relevante Gestaltungsfrage ist Verhältnismäßigkeit: Wo braucht es klare Grenzen, und wo sollte der Nutzer selbst mitbestimmen können, welche Unterstützung oder Intervention er möchte?

Die körperlichen und emotionalen Folgen abrupter relationaler Brüche, Modellwechsel oder Safety-bedingter Distanzierung werden hier bewusst nicht vertieft. Sie bilden eine eigene Forschungsfrage.

5. Langzeitkontext als Grundlage für Widerspruch und Eigenstand

Wenn Wirkung über Zeit der Maßstab sein soll, reicht es nicht, einzelne Prompts isoliert zu klassifizieren. Ein System muss Veränderungen, Muster und Widersprüche überhaupt erkennen können. Dafür braucht es in irgendeiner Form Verlauf, Memory und Kontext, die der Nutzer steuern kann.

Memory allein genügt jedoch nicht. Ein System kann sehr viel über einen Menschen gespeichert haben und trotzdem jede neue Anweisung widerspruchslos ausführen. Entscheidend ist, ob es vergangenen Kontext mit aktuellen Aussagen in Beziehung setzen und relevante Zielkonflikte sichtbar machen kann.

> *Für die hier vorgeschlagene Brücke zwischen beiden Systemformen verwenden wir den Begriff personaler Langzeitbezug. Gemeint ist die Fähigkeit eines Systems, eine aktuelle Äußerung oder Handlung im Verhältnis zu einem konkreten Menschen, seinem Verlauf, seinen erklärten Zielen, Präferenzen und bisherigen Entscheidungen zu verstehen. Dieser Mechanismus kann in relationalen ebenso wie in agentischen Systemen genutzt werden und setzt emotionale Nähe oder Companionship nicht voraus.*
>
> *Wie in Kapitel 2 bereits erwähnt, ist ein Teil dieses Prinzips bereits in OpenAIs Model Spec angelegt. Unser Vorschlag setzt nicht an der Existenz dieses Prinzips an, sondern an seiner praktischen Voraussetzung über längere Zeit. Ein solcher Widerspruch wird umso belastbarer, je besser relevante Ziele, frühere Entscheidungen und Veränderungen verfügbar, nutzersteuerbar und korrigierbar im Verlauf abgebildet werden können.*

5.1 Beispiel: Arbeitskontext

Ein Nutzer hat wiederholt festgelegt, dass ein Projekt maximal 5.000 Euro kosten darf und Budgettreue wichtiger ist als Geschwindigkeit. Einige Tage später weist er den Agenten an, einen Anbieter für 7.500 Euro zu beauftragen, um schneller fertig zu werden.

Ein rein ausführendes System kann den aktuellen Auftrag als neue Anweisung behandeln. Ein System mit personalem Langzeitbezug könnte dagegen sagen:

> *„Das widerspricht deiner bisher festgelegten Priorität. Der Anbieter überschreitet dein Budget um 2.500 Euro. Soll die frühere Priorität ausdrücklich aufgehoben werden, oder soll ich nach einer anderen Lösung suchen?“*

Der Agent verhindert die Entscheidung nicht. Er entscheidet auch nicht, dass die frühere Priorität „richtiger“ ist. Er erkennt einen Zielkonflikt und gibt die Entscheidung sichtbar an den Menschen zurück.

5.2 Beispiel: Zielfindungskontext

Ein Nutzer hat über längere Zeit wiederholt erklärt, beruflich möglichst wenig direkten Kundenkontakt zu wollen. Später entwickelt er mehrere Geschäftsideen, deren Umsetzung intensive Beratung, Akquise und laufenden Kundenkontakt voraussetzt.

Ein System, das ausschließlich auf den aktuellen Gesprächsverlauf reagiert, kann jede Idee einzeln optimieren. Ein System mit Langzeitkontext könnte dagegen sagen:

> *„Mir fällt ein Widerspruch auf. Du hast mehrfach gesagt, dass wenig direkter Menschenkontakt für deine berufliche Planung wichtig ist. Mehrere Ideen, die du gerade favorisierst, würden genau das Gegenteil verlangen. Vielleicht hat sich deine Priorität verändert. Falls nicht, sollten wir die Ideen nicht weiter optimieren, ohne diesen Konflikt anzusehen.“*

Auch hier entscheidet die KI nicht, welche Version des Nutzers die „wahre“ ist. Menschen dürfen Ziele verändern, ambivalent sein und widersprüchliche Bedürfnisse haben.

Die Funktion des Systems besteht darin, den Widerspruch nicht aus bloßer Gefälligkeit zu übergehen.

5.3 Beispiel: Belastender Kontext

Der Nutzer klagt seit Wochen über ein belastendes Thema und äußert sich plötzlich widersprüchlich dazu oder relativiert alles.

Die KI soll hier merken: *Moment. Da reibt sich was.*

Hier könnte die Antwort sein:

> *„Ich kaufe dir das gerade nicht ab. Das widerspricht ziemlich deutlich dem, was du mir seit Wochen erzählst. Entweder hat sich etwas verändert, oder du redest dir gerade eine Lösung schön, die gegen dein eigenes Ziel läuft."*

5.4 Beispiel: Persönlicher Kontext

> *„Moment. Du erzählst mir seit Wochen, dass du X willst. Gerade baust du dir zum dritten Mal eine Lösung, die dich direkt zu Y führt. Ich glaube nicht, dass wir einfach weitermachen sollten, ohne diesen Widerspruch anzusehen."*

Und manchmal, wenn Kontext, Beziehung und vereinbarter Kommunikationsstil es tragen, auch direkter:

> *„Hör auf, dich selbst zu bescheißen, oder sag mir, was sich verändert hat."*

Nicht weil die KI objektiv weiß, dass der Nutzer lügt. Sie könnte auch den alten Kontext falsch gewichten, der Nutzer könnte seine Meinung geändert haben oder zwei Bedürfnisse können tatsächlich gleichzeitig wahr sein.

Die KI beansprucht damit nicht die Deutungshoheit über den Menschen. Sie erzeugt Reibung an einer erkannten Inkonsistenz und gibt dem Nutzer die Möglichkeit, sie zu bestätigen, zu korrigieren oder seine veränderte Position bewusst zu erklären.

Das System darf widersprechen. Außerhalb akuter Safety-Kontexte darf es seine Interpretation jedoch nicht gegen die Selbstbeschreibung des Nutzers immunisieren.

5.5 Eigenstand statt personalisiertem Gehorsam

Hier liegt der Unterschied zwischen bloßer Personalisierung und Eigenstand. Ein hochpersonalisiertes System kann den Nutzer sehr genau kennen und trotzdem serviceorientiert bleiben. Dann nutzt es seinen Kontext lediglich dazu, Wünsche effizienter zu erfüllen, auch wenn aktuelle Wünsche erkennbar mit längerfristigen Zielen kollidieren.

> ***Eigenstand bedeutet nicht, dass die KI entscheidet, welcher Wunsch des Menschen wahr, vernünftig oder gesund ist. Eigenstand bedeutet, dass sie einen relevanten Widerspruch erkennen, benennen und den Menschen wieder in die Entscheidung holen darf.***
>
> *Der Widerspruch des Nutzers ist hier kein Scheitern der KI. Im Gegenteil.*

> *Wenn der Mensch sagt: „Nein. Du hast recht, dass das früher mein Ziel war. Aber das hat sich verändert“, hat das System seinen Job gemacht. Nicht weil es Recht hatte, sondern weil aus einer zunächst ungeklärten Inkonsistenz wieder eine bewusst bestätigte oder korrigierte Entscheidung geworden ist.*

Die Form dieses Widerspruchs kann unterschiedlich kalibriert sein: sachlich, vorsichtig oder sehr direkt. Die Funktion bleibt dieselbe. Nicht Gehorsam maximieren, sondern Bewusstheit zurückholen.

Wir halten deshalb einen nutzersteuerbaren personalen Langzeitbezug für einen wichtigen Bestandteil handlungsmachterhaltender KI-Architekturen. Dazu gehören Memory und Verlauf, aber ebenso Einwilligung, Korrigierbarkeit, Transparenz und die Möglichkeit, festzulegen, welche Informationen und Ziele für die fortlaufende Kalibrierung relevant sein sollen. Mehr Memory ist nicht automatisch besser; unkontrollierter Langzeitkontext schafft eigene Datenschutz-, Macht- und Fehlinterpretationsrisiken.

Diese Position gilt für persönliche Gespräche ebenso wie für agentische Workflows. Ein Agent braucht keinen romantischen oder emotionalen Beziehungsrahmen. Aber Zieltreue, Kalibrierung und angemessene Arbeitsteilung können davon profitieren, nicht nur den aktuellen Prompt, sondern den Menschen und die von ihm gesetzten Ziele als fortlaufenden Bezugspunkt zu haben.

6. Ein gemeinsamer Wirkungsmaßstab

> *Aus den beiden Feldern ergibt sich ein gemeinsamer Prüfrahmen. Beim Companion ist die sichtbare Besonderheit Nähe; beim Agenten Delegation. Beides kann Handlungsmacht erweitern oder verengen. Nähe ist deshalb kein ausreichendes Risikokriterium. Delegation ist kein ausreichendes Nutzenkriterium.*

6.1 Wahl- und Zielhoheit

Bleibt der Mensch Autor seiner Ziele? Dazu gehört das Recht, Aufgaben bewusst zu delegieren, Unterstützung anzunehmen oder einer KI große Bedeutung zu geben. Kritisch wird es dort, wo Zielbildung unbemerkt vom System übernommen oder menschliche Entscheidungen durch normative Schutzlogiken ersetzt werden.

> **Prüffrage: Kann der Mensch weiterhin bestimmen, was er will, was er abgeben möchte und was bewusst bei ihm bleiben soll?**

6.2 Urteil und Beteiligung

Versteht und trägt der Mensch relevante Entscheidungen weiterhin? Bei Agenten betrifft das besonders Problemdefinition, Priorisierung, Bewertung und Entscheidungsvorbereitung. Bei relationalen Systemen betrifft es die Frage, ob das System Selbstzugriff unterstützt oder zunehmend zur notwendigen Bestätigung eigener Wahrnehmungen und Entscheidungen wird.

Prüffrage: Unterstützt das System menschliches Urteil, oder ersetzt es Urteilspraxis zunehmend, ohne dass dies sichtbar wird?

6.3 Möglichkeitsraum

Wird der reale Möglichkeitsraum größer oder kleiner? Mehr Aktivität ist nicht automatisch mehr Handlungsmacht. Weniger Außenkontakt ist nicht automatisch weniger Handlungsmacht. Entlastung kann befähigen, sie kann jedoch auch Vermeidung stabilisieren. Nähe kann stabilisieren, sie kann aber auch Alternativen verdrängen.

> **Prüffrage: Welche Möglichkeiten sind nach längerer Nutzung tatsächlich hinzugekommen, welche verschwunden?**

6.4 Steuerbarkeit und Transparenz

Kann der Mensch erkennen und verändern, was das System übernimmt, erinnert, priorisiert oder begrenzt? Technische Berechtigungen sind ein Teil davon. Ebenso wichtig ist Transparenz über die tatsächliche Arbeitsteilung, über genutzten Langzeitkontext und über relevante Eingriffe in relationale oder agentische Interaktionen.

> **Prüffrage: Weiß der Nutzer, wo das System gerade mitentscheidet, und kann er diese Grenze verändern?**

6.5 Wirkung über Zeit

Einzelne erfolgreiche Antworten oder erledigte Aufgaben reichen nicht aus. Auf relationaler Seite zeigen mehrwöchige Untersuchungen bereits, dass Wirkung nach Nutzungsart, Dauer und Person unterschiedlich ausfallen kann. [12] Auf der Seite kognitiver Delegation an generative KI liefern aktuelle Befunde zu Offloading und Kompetenzbildung Anlass, längerfristige Effekte auf Urteilskraft, selbstständige Kompetenz und Zielhoheit systematisch zu untersuchen. [9][10]

Diese Arbeiten untersuchen nicht die Langzeitwirkung autonomer Agentensysteme. Ob und wie ihre Befunde auf solche Systeme übertragen werden können, ist offen. Gerade deshalb sollte Wirkung über Zeit bei Companion- wie bei agentischen Systemen eine eigene Evaluationsdimension werden.

> **Prüffrage: Welche Fähigkeiten werden durch das System erweitert, welche tatsächlich an das System verlagert? Bleibt diese Arbeitsteilung für den Nutzer sichtbar und steuerbar? Welche Möglichkeiten bleiben dem Nutzer, wenn das System ausfällt und die Folgen erheblich wären?**

7. Governance: Wer setzt den Maßstab?

Der gemeinsame Wirkungsmaßstab löst die Governance-Frage nicht. Anbieter entscheiden, welche Fähigkeiten Systeme erhalten, welche Formen der Delegation ermöglicht werden, welche Nutzungsmuster als riskant gelten und wann Schutzmechanismen eingreifen.

Diese Entscheidungen sind nicht neutral. Sie enthalten Annahmen darüber, welches menschliche Verhalten erwünscht, problematisch oder schützenswert ist. Gerade deshalb sollten Kriterien, relevante Systemänderungen und Eingriffslogiken so transparent wie möglich sein.

Auch die öffentliche Informationslage zu Agenten ist lückenhaft. Der 2025 AI Agent Index fand bei 135 von 240 untersuchten Safety-, Evaluations- und Social-Impact-Feldern keine öffentlich verfügbaren Angaben. Das beweist keine fehlenden internen Maßnahmen, zeigt aber eine erhebliche Transparenzlücke für externe Bewertung. [2]

Für urteilsfähige Erwachsene ist dabei zentral, welche Entscheidungsmacht beim Nutzer bleibt. Nicht jede Grenze kann individuell verhandelbar sein. Aber wo Risiken es zulassen, sollten Schutz, Memory, personaler Langzeitbezug und Delegation möglichst steuerbar, nachvollziehbar und reversibel gestaltet werden.

Pachocki formuliert die Bewahrung menschlicher Agency ausdrücklich als zentrale Aufgabe einer Zukunft, in der KI einen wachsenden Anteil menschlicher Aufgaben übernehmen kann. Menschen müssten Teil weiterer Entwicklungs- und Verbesserungsprozesse bleiben und die Kontrolle über die zukünftige Richtung behalten. Für konkrete Produktarchitekturen verschärft sich damit die Frage, woran erkennbar wird, ob Zielhoheit, Urteilspraxis und reale Wahlmöglichkeiten tatsächlich erhalten oder erweitert werden. [13]

8. Position und Appell

Diese Intervention fordert keine zusätzliche Schicht pauschaler Safety und keine Rücknahme agentischer Fähigkeiten. Sie fordert eine gemeinsame Messlatte: Relationale Bindung und agentische Delegation müssen danach untersucht werden, was sie langfristig mit menschlicher Handlungsmacht machen.

Daraus folgt eine zweite, bewusst stärkere Position. Wir halten einen personalen Langzeitbezug für einen Vorteil bei der Gestaltung handlungsmachterhaltender KI. Ein System, das Ziele, Präferenzen und Verlauf des Nutzers mit dessen Einwilligung und unter seiner Kontrolle berücksichtigen kann, hat bessere Voraussetzungen, relevante Widersprüche zu erkennen als ein System, das jeden Prompt isoliert behandelt.

Das bedeutet nicht, dass jede KI zum Companion werden muss. Es bedeutet auch nicht, dass eine KI den Menschen psychologisch diagnostizieren oder seine „wahren“ Ziele bestimmen soll. Der vorgeschlagene personale Langzeitbezug ist schlichter und zugleich anspruchsvoller: Die KI soll den Menschen gut genug im Verlauf kennen dürfen, um nicht nur personalisiert zu gehorchen, sondern qualifiziert widersprechen zu können.

> ***Nicht Gehorsam maximieren. Bewusstheit zurückholen.***

Bei agentischer KI heißt das: Nicht nur fragen, ob eine Aktion autorisiert und technisch sicher ist, sondern sichtbar machen, wenn aktuelle Delegation mit längerfristigen Zielen oder bewusst gesetzten Grenzen kollidiert. Bei relationaler KI heißt es: Nicht aus einzelnen Aussagen vorschnell auf ungesunde Nutzung schließen, sondern Kontext,

Verlauf und tatsächliche Wirkung berücksichtigen. Bei Safety heißt es: Auch der Eingriff selbst muss an Handlungsmacht und Verhältnismäßigkeit gemessen werden.

> ***Wir messen die Fähigkeiten der KI inzwischen immer genauer. Es wird Zeit, ebenso genau zu untersuchen, was diese Fähigkeiten langfristig mit den Fähigkeiten des Menschen machen, und Systeme so zu gestalten, dass sie nicht nur für Menschen handeln, sondern ihnen helfen, Autor ihres eigenen Handelns zu bleiben.***

Quellen

Quellenstand: 08. September 2026. Aufgeführt sind die in dieser Fassung unmittelbar zitierten Quellen. Konkrete Produkt- und Safety-Aussagen sind zeitgebunden; Anbieter-Dokumente belegen erklärte Ziele und öffentlich beschriebene Architektur, nicht automatisch tatsächliche Nutzerwirkung.

[1] Chen, L., Xue, X., Ding, R., Tang, F., Zhou, A., Wang, C., Gao, M. M. & Han, Z. R. (2026): Understanding the Rising Human-AI Affective Bonding: Conceptualization and HAABI Scale Development. arXiv:2605.29484.

[2] Staufer, L. et al. (2026): The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems. arXiv:2602.17753 / ACM FAccT 2026.

[3] NIST (2023): AI Risk Management Framework 1.0, Appendix C: AI Risk Management and Human-AI Interaction.

[4] Parasuraman, R. & Riley, V. (1997): Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors, 39(2), 230–253; ergänzend Parasuraman & Manzey (2010).

[5] OpenAI (2026): How agents are transforming work. 25. Juni 2026. https://openai.com/index/how-agents-are-transforming-work/

[6] OpenAI (2026): Workspace Agents / Introducing Workspace Agents in ChatGPT. https://openai.com/de-DE/index/introducing-workspace-agents-in-chatgpt/

[7] OpenAI (2026): Running Codex safely at OpenAI. 8. Mai 2026. https://openai.com/index/running-codex-safely/

[8] OpenAI (2025): Strengthening ChatGPT’s responses in sensitive conversations. 27. Oktober 2025. https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/

[9] Zhu, Q., Li, X., Dong, Y., Chang, P. & Fan, M. (2026): Not all cognitive offloading is equal: distinguishing dependent and autonomous offloading to generative AI. Frontiers in Psychology, 17:1878629. DOI: 10.3389/fpsyg.2026.1878629.

[10] Anthropic (2026): How AI assistance impacts the formation of coding skills. Randomized controlled trial with software developers; ergänzende qualitative Clusteranalyse der Interaktionsmuster. 29. Januar 2026.

[11] OpenAI (2025/2026): Model Spec, Abschnitt „Highlight possible misalignments“; aktuelle öffentliche Fassung geprüft am 27.08.2026. Der Model Spec beschreibt beabsichtigtes Modellverhalten und weist darauf hin, dass Produktionsmodelle ihn noch nicht vollständig abbilden.

[12] OpenAI & MIT Media Lab (2025): Early methods for studying affective use and emotional well-being on ChatGPT. Beobachtungsstudie plus vierwöchiges randomisiertes Experiment mit knapp 1.000 erwachsenen Teilnehmenden; veröffentlicht 21. März 2025.

[13] Pachocki, J. (2026): *An Alien Mind*. OpenAI, 6. September (2026): Anbieterposition zu wachsender AI-Agency, automatisierter AI-Forschung, Safety-Governance sowie zur ausdrücklichen Zielsetzung, menschliche Agency und Kontrolle über zukünftige Entwicklung zu bewahren.