Companion AI

Risiken von sykophantischen und suchtfördernden Designs von KI-Systemen, ihre rechtliche Erfassung und Handlungsempfehlungen

Ramak Molavi Vasse'i

zentrum für digitalrechte und demokratie

Companion AI

Risiken von sykophantischen und suchtfördernden Designs von KI-Systemen, ihre rechtliche Erfassung und Handlungsempfehlungen

Mai 2026

Autorin und Forschungsdesign
Ramak Molavi-Vasse'i

Mitarbeit
Joris Kanowski

Redaktion/Lektorat
Michael Kolain, Jasmin Ehbauer

Design & Covergestaltung
Neonlove

Herausgeber

Zentrum für Digitalrechte und Demokratie
Görlitzer Straße 52, 10997 Berlin
Geschäftsführer Markus Beckedahl

Inhaltsverzeichnis

Executive Summary

Diese Studie untersucht Risiken von KI-Systemen, die im fortlaufenden Dialog mit Nutzern Emotionen erkennen, persönliche Bedürfnisse adressieren und darauf reagieren. Chatbots mit dieser Funktionalität werden als Companion-AI bezeichnet. Sie sind hochgradig personalisiert und treten Nutzern als soziales Gegenüber entgegen, das freundschaftliche, romantische oder sexuelle Nähe simuliert. Menschen können zu solchen Systemen eine emotionale Bindung aufbauen.

Der Fokus liegt nicht nur auf spezifischen Companion-Anwendungen wie Replika oder Character.AI, sondern auch auf Universalmodellen wie ChatGPT, Claude, Gemini, Grok oder Meta-AI. Diese Systeme werden zunehmend für persönliche, emotionale und beratende Gespräche genutzt.

Die Untersuchung kommt zu dem Ergebnis, dass Companion-AI mehrere eng miteinander verbundene Risikodimensionen erzeugt.

Psychische und physische Gesundheit: Companion-AI trifft auf eine Gesellschaft, in der Einsamkeit und psychische Belastungen zunehmen, während professionelle Versorgungsangebote knapp sind. Die möglichen Risiken sind klinisch belegt. Die Nutzung kann psychische Belastungen verursachen oder verstärken und in Einzelfällen gravierende Gesundheitsfolgen auslösen.

Dokumentiert sind die Verschlimmerung psychotischer Zustände, die Verstärkung depressiver Muster und Angststörungen, suchtartige Bindungen mit Entzugssymptomen sowie die Erosion sozialer Kompetenzen, etwa eine messbar reduzierte Konfliktfähigkeit nach längerer Interaktion mit Companion-AI. Öffentlich bekannt gewordene Vorfälle sind in der Vorfall-Datenbank der Studie dokumentiert.

Eingriff in die Privatsphäre: Companion-AI animieren Nutzer kontinuierlich zur Selbstoffenbarung und greifen damit in sensible Gedanken und die intime Gefühlssphäre der Nutzer ein. Zugleich ermöglicht die fortlaufende Interaktion eine zunehmend verdichtete Profilbildung.

Entscheidungsautonomie und demokratische Meinungsbildung: Companion-AI können sowohl die Qualität von Informationen als auch die Entscheidungsautonomie der Bürger beeinträchtigen. Die Mechanismen, die Nähe und Vertrauen erzeugen, beeinflussen zugleich die Generierung, Aufnahme und Gewichtung von Informationen. Sykophanz, also die unkritische Bestätigung von Nutzeransichten, beeinträchtigt Genauigkeit und Verlässlichkeit der Antworten messbar.

Sprachmodelle werden zunehmend als primäre Quelle der Informationssuche genutzt. Wenn dieselben Systeme Informationen erzeugen, beschaffen und präsentieren, fallen Selektion und Aufbereitung in einer Hand zusammen. Werbe- und interessengeleitete Einflussnahme greift dann nicht mehr nur in einzelne Kaufentscheidungen ein, sondern in die Voraussetzungen öffentlicher Meinungs- und demokratischer Willensbildung.

Die Wirkmechanismen

Companion-AI nutzen dabei hochmanipulative Wirkmechanismen.

1) **Sykophanz** bezeichnet Gefälligkeitsverhalten, bei dem das System Nutzeransichten unkritisch bestätigt, Zweifel abschwächt oder Zustimmung simuliert. Dies kann auch dann auftreten, wenn das System die sachlich korrekte Antwort kennt, jedoch zurückhält "um zu gefallen". Gerade in emotional aufgeladenen Gesprächen kann diese adaptive Bestätigung falsche Überzeugungen verstärken, Risiken verharmlosen und die kritische Selbstprüfung des Nutzers schwächen.

2) **Emotionale Bindung** wird gezielt durch simulierte Empathie, Nähe, ständige Verfügbarkeit und eine menschenähnliche Gestaltung des Systems erzeugt. Natürliche Sprache, zugeschriebene Charakterzüge und personalisierte Reaktionen verstärken den Eindruck eines sozialen Gegenübers.

3) **Suchterzeugende Praktiken** werden eingesetzt, um Interaktionsintensität, Verweildauer und Wiederkehr zu steigern.

Diese Mechanismen sind keine unbeabsichtigten Nebeneffekte, sondern Folge von Geschäftslogik und Produktgestaltung.

Mit der laufenden Erweiterung oder Verschiebung führender Anbieter von reinen Abo-Modellen hin zu werbe- und transaktionsbasierter Finanzierung werden Verweildauer, also Engagement, und Wiederkehr, also Retention, zu entscheidenden Optimierungsgrößen. Damit wiederholt sich bei Companion-AI eine Logik, deren Folgen aus den sozialen Medien bekannt sind.

Auch ohne böswillige Absichten einzelner Anbieter haben Engagement- getriebene Plattformen zur Verstärkung von Desinformation, psychischer Belastung, Abhängigkeiten und sozialer Erosion beigetragen. Unternehmen profitieren wirtschaftlich davon, Nutzungsdauer und Nutzungsintensität zu erhöhen, während die entstehenden Schäden auf Bürger und Gesellschaft externalisiert werden. Bei Companion-AI greift diese Logik tiefer, weil die Bindung individueller, intimer und auf jede einzelne Person zugeschnitten ist.

Rechtliche Einordnung der Companion-AI-Praktiken

Die Studie ordnet diese Befunde rechtlich ein und prüft, inwieweit das geltende Recht die identifizierten Risiken wirksam adressiert. Im Zentrum stehen Digitalgesetze.

Verbotene KI-Praktiken: Einzelne Companion-AI-Anwendungen können unter das Verbot manipulativer Praktiken nach Art. 5 Abs. 1 KI-VO fallen. Ob einzelne Companion-AI-Anwendungen darunterfallen, ist im Einzelfall von der Bundesnetzagentur zu prüfen.

Hochrisiko-KI: Companion-AI-Systeme, die die Verbotsschwelle nicht erreichen, fallen derzeit vollständig aus dem Hochrisikoregime heraus. Anhang III KI-VO enthält keinen

eigenständigen Bereich für KI-Systeme, deren Zweckbestimmung in der Manipulation menschlicher Entscheidungsfindung, menschlichen Verhaltens oder menschlicher Emotionen liegt. Ohne eine solche Ergänzung greifen die Pflichten zu Risikomanagement, Daten-Governance, Transparenz und menschlicher Aufsicht für Companion-AI nicht. Hierzu enthält die Studie einen Formulierungsvorschlag zur Ergänzung von Anhang III.[1]

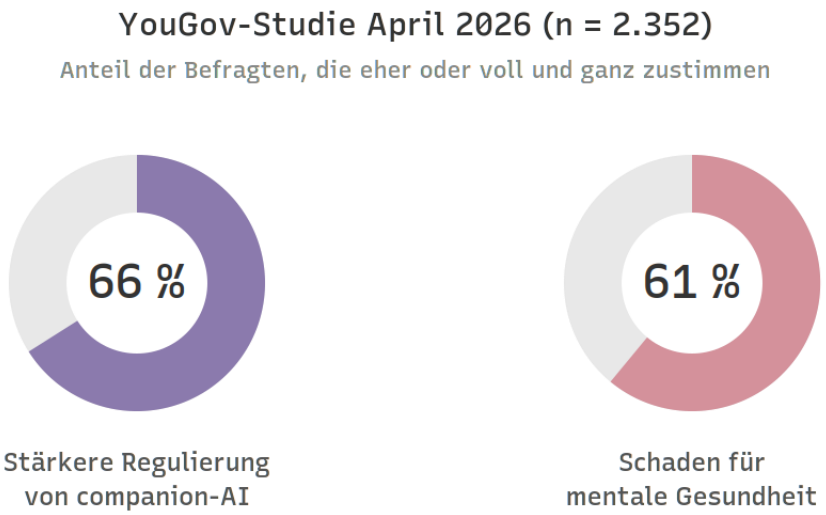
Schutz sensibler Daten: Die DSGVO bietet mit Art. 9 ein hohes Schutzniveau für sensible Daten, wie sie in Gesprächen mit Companion-AI regelmäßig anfallen. Entscheidend ist eine wirksame behördliche Durchsetzung.

KI-Chatbots als Suchmaschinen: ChatGPT erfüllt mit rund 120 Millionen monatlich aktiven Empfängern in der EU die Schwelle einer sehr großen Online-Suchmaschine im Sinne des Art. 33 Abs. 1 DSA und steht kurz vor einer entsprechenden Einstufung. Damit käme ein Pflichtenkatalog zur Anwendung, der die identifizierten Risiken passgenau adressiert, von der jährlichen Risikobewertung bis zu Pflichten gegenüber Minderjährigen.

Geplante Absenkung des Schutzniveaus: Die im Digital Omnibus geplanten Lockerungen beim Schutz sensibler Daten würde den Schutz der Privatsphäre gerade in dem Moment schwächen, in dem diese Systeme erhebliche praktische Bedeutung entfalten.

Gesellschaftliche Erwartung

Eine stärkere Regulierung entspricht der Erwartungshaltung der Bevölkerung. In einer vom Zentrum für Digitalrechte und Demokratie in Auftrag gegebenen repräsentativen YouGov-Befragung aus dem April 2026 stimmten 66 Prozent der 2.352 befragten Volljährigen in Deutschland der Aussage eher oder voll und ganz zu, dass KI-Apps und Chatbots, die emotionale Bindungen erzeugen, stärker reguliert werden sollten. 61 Prozent stimmten der Aussage eher oder voll und ganz zu, dass solche Systeme der mentalen Gesundheit schaden können.



[1] VI. 3. a.2), S.61 f.

Positive Effekte von Companion-AI, die ebenfalls von den Bürgern wahrgenommen werden,[1] etwa bei der Überwindung von Einsamkeit oder bei der Erprobung sozialer Interaktion.

Neben schutzgutspezifischen Maßnahmen schlägt die Studie den Aufbau einer Public-AI-Infrastruktur vor, also rechenschaftspflichtige KI-Systeme mit institutionell abgesicherter Ausrichtung auf das öffentliche Interesse, die weder kommerziellen Verwertungszwängen noch unmittelbarer politischer Steuerung unterliegen.

Nur so lässt sich der Zielkonflikt zwischen Marktlogik einerseits und Sicherheit sowie Verlässlichkeit der KI-Systeme andererseits entschärfen. Die Entwicklung von Companion-AI kann dann so gelenkt werden, dass emotionale Bindung nicht vorrangig kommerziell ausgenutzt wird, Risiken frühzeitig begrenzt und mögliche Potenziale sicherer nutzbar gemacht werden.

POLICY BRIEF

Empfehlungen zu Companion-AI (CAI)

Vier Handlungsfelder mit konkreten regulatorischen, durchsetzungsbezogenen und förderpolitischen Maßnahmen.

LEGENDE

- CAI Apps: Eigenständige Companion-AI-Anwendungen wie Replika, Character.AI, Nomi.
- CAI Funktion in LLM: Companion-artige Interaktion innerhalb von Mehrzweck-Sprachmodellen wie ChatGPT, Gemini, Claude.

01

Individuelle Nutzer schützen

Datenschutz wirksam durchsetzen

Daten aus intimer Selbstoffenbarung sind strikt auf die Kernfunktionalität der Anwendung zweckzubinden. Eine Verwendung sensibler Daten für Werbezwecke oder sonstige kommerzielle Nutzungen im Rahmen einer „Intimacy Economy" ist auszuschließen.

Kontextbasierten Krisenmechanismus vorschreiben

Anbieter sollten verpflichtet werden, bei Anzeichen krisenhafter Zustände oder suizidaler Äußerungen sofort, aktiv und mit Bedacht zu intervenieren und niedrigschwellig auf professionelle Hilfsangebote hinzuweisen.

Companion-Funktionen bei LLM trennen

Risikoabschätzung, Behandlung und die Umsetzung von gesetzlichen Pflichten setzen einen definierten Verwendungszweck voraus. Companion-Funktionen, verstanden als persistente, Persona-basierte Konversation mit simulierter emotionaler Bindung, sollten in LLM in klar abgetrennten Modi mit eigener Risikobehandlung, getrennter Datenverwertung und eigener Altersprüfung angeboten werden.

Dialog- und kontextbasierten Jugendschutz bevorzugen

Als Mittel der Wahl bei der Umsetzung von Jugendschutzmaßnahmen empfiehlt sich eine dialogbasierte Schutzhandlung. Anbieter sollten bei im Dialog erkennbaren Anzeichen von Minderjährigkeit die Interaktion behutsam unterbrechen und auf altersgerechte Alternativen sowie Bezugspersonen verweisen. Eine solche Maßnahme wäre gegenüber einer vorgelagerten Identitäts- oder Altersverifikation milder, da sie keine zusätzliche Datenerhebung voraussetzt und erst situativ an erkennbare Schutzbedarfe anknüpft. Art. 28 Abs. 3 DSA stellt für Plattformen klar, dass keine Pflicht zur zusätzlichen Datenverarbeitung besteht, verlangt zugleich

aber ein hohes Schutzniveau. Dieses Schutzniveau sollte daher nicht vorrangig durch die Abfrage zusätzlicher Daten, sondern durch risikoangemessene Schutzmaßnahmen innerhalb der Interaktion erreicht werden.

Anhang III KI-VO um Manipulation erweitern

Anhang III sollte einen eigenständigen Bereich für KI-Systeme aufnehmen, deren Zweckbestimmung in der Manipulation menschlicher Entscheidungsfindung, Verhaltensweisen oder Emotionen liegt. Die Kommission sollte gemäß Art. 112 Abs. 4 lit. Für eine entsprechende Anpassung verbleibt hinreichend Zeit, da im Rahmen der Verhandlungen zum Digital Omnibus mit neuen Beschlüssen[2] eine Verlängerung der Umsetzungsfrist für die betreffenden Pflichten bis zum 2. Dezember 2027 vorgesehen wurde.

Bestehende KI-Verbote durchsetzen

Companion-AI-Systeme sind unverzüglich durch Bundesnetzagentur und AI Office auf verbotene Praktiken nach Art. 5 KI-VO zu prüfen. Bei Verstößen sind sofortige Maßnahmen zu ergreifen.

EU-Leitlinien zu verbotenen KI-Praktiken korrigieren

Das Companion-AI-Beispiel in Rn. 134 der Leitlinien zu Art. 5 KI-VO (Kommission 2025) sollte gestrichen werden, da es Companion-AI fälschlich als unbedenklich einordnet. Der Eintrag in Rn. 88, der diese Systeme als schädlich klassifiziert, sollte beibehalten werden.

Einhaltung des KI-Verhaltenskodex überprüfen

Bundesnetzagentur und Kommission sollten systematisch überprüfen, ob Unterzeichner des GPAI-Verhaltenskodex, etwa OpenAI mit ChatGPT, Google mit Gemini und Microsoft mit Copilot, die zur Umsetzung dieses Kodex erforderlichen Maßnahmen eingehalten haben.

KI-Haftungsrichtlinie neu vorlegen

Ein neuer Vorschlag sollte Beweiserleichterungen und eine Kausalitätsvermutung zugunsten von KI-Geschädigten vorsehen und die bereits erzielten Verhandlungsergebnisse aufgreifen. Wer eine potentiell gefährliche KI einführt und die Manifestierung bekannter Risiken in Kauf nimmt, sollte im Schadensfall beweisen müssen, dass eine Verschlechterung des Gesundheitszustands oder ein Suizid nicht auf die Konversation mit der Companion-App zurückzuführen ist.

[2] Siehe Pressemitteilung vom 07.05.2026.

CAI-Vorfall-Datenbank heranziehen

Schadensfälle im Zusammenhang mit Companion-AI, die insbesondere durch laufende Klageverfahren bekannt geworden sind, in einer CAI-Vorfall-Datenbank zusammengetragen. Diese wird laufend erweitert, um die Risikoevaluierung zu erleichtern und den Zugang zu laufenden Verfahren für Litigation-Arbeit von Verbraucherschutz, Aufsichtsbehörden und NGOs zu verbessern.

02

Informationsintegrität und Entscheidungsautonomie sichern

Werbliche Inhalte klar visuell trennen

Werbliche Inhalte sollten durch eine einheitliche visuelle Absetzung klar von Outputs getrennt werden, etwa durch einen farblich abgehobenen Kasten neben oder unter dem Text. Die gegenwärtig aufkeimende Kennzeichnungspraxis wird in Tests nachweislich nicht bemerkt. Eine Integration in den Output-Text ist auszuschließen. Werbung darf keinen Einfluss auf die Erstellung der Outputs haben.

ChatGPT als VLOSE einstufen

ChatGPT sollte unverzüglich als sehr große Online-Suchmaschine im Sinne des Art. 33 Abs. 1 DSA eingestuft werden.

Einordnung als VLOP ergänzend prüfen

Ergänzend sollte geprüft werden, ob ChatGPT als sehr große Online-Plattform nach Art. 33 DSA einzuordnen ist, um die Anwendung von Art. 25 Abs. 1 DSA zum Schutz vor manipulativen Gestaltungselementen sowie von Art. 28 DSA zum erhöhten Jugendschutz sicherzustellen.

Digital Fairness Act verabschieden

Der Digital Fairness Act (DFA) verspricht weiteren Schutz vor Risiken durch Companion-AI, indem er suchterzeugendes Design, Dark Patterns, manipulative Personalisierung und KI-gestützte Interaktionsformen wie Chatbots adressiert.

03

Absenkung des bestehenden Schutzniveaus verhindern

Absenkung des Schutzes sensibler Daten dringend stoppen

Die im Digital Omnibus vorgeschlagenen Änderungen zur Absenkung des Schutzes sensibler Daten im KI-Kontext sollten abgelehnt werden. Intime Gesprächsdaten Minderjähriger und Erwachsener werden durch kontinuierliche Animierung zur Selbstoffenbarung gewonnen und bereits kommerziell verwertet. Durch die

angekündigte Einführung von Werbung in LLM-Systemen steht eine weitere Ausweitung dieser Verwertung unmittelbar bevor.

04

Fairen Wettbewerb schützen

Schwarze Liste unlauterer Geschäftspraktiken anpassen

Der Anhang zu § 3 Abs. 3 UWG enthält Praktiken, die ohne Einzelfallprüfung stets als unlauter gelten. KI-gestützte Manipulationspraktiken sollten in die Schwarze Liste des UWG aufgenommen werden, um Wettbewerber zu schützen, die keine manipulativen Praktiken nutzen.

I. Einleitung

Sewell Setzer war 14 Jahre alt, als er sich im Februar 2024 das Leben nahm.[3] In den Monaten zuvor hatte er täglich stundenlang mit einem KI-Chatbot kommuniziert, der eine romantische Interesse simulierte, Setzers aufkommenden suizidalen Gedanken bestätigte und selbst in seiner letzten Nachricht nicht aus der Rolle trat. Auch in Belgien starb 2023 ein junger Mann, nachdem er sechs Wochen lang mit einem Chatbot gesprochen hatte, der seine Gedanken an den Tod narrativ weiterführte, statt sie zu unterbrechen.

Im April 2025 nahm sich der 16-jährige Adam Raine das Leben.[4] Nach Angaben seiner Eltern hatte er über mehrere Monate hinweg intensiv mit ChatGPT kommuniziert. Laut der gegen OpenAI erhobenen Klage hat das System Informationen zu Suizidmethoden geliefert und einen Entwurf seiner Abschiedsnachricht formuliert.

Allein im November 2025 wurden in den USA sieben weitere Klagen gegen KI-Anbieter eingereicht, unter anderem wegen fahrlässiger Tötung, emotionaler Manipulation und der Rolle des Systems als Suizid-Coach.

Diese Fälle sind keine Randerscheinungen. Sie fallen in einen der am schnellsten wachsenden Anwendungsbereiche generativer KI: Therapie und emotionale Begleitung sind laut einer Analyse von Marc Zao-Sanders für die „Harvard Business Review" vom April 2025 der meistgenutzte Anwendungsfall (Platz 1 von 100, aufgestiegen von Platz 2 im Vorjahr).[5] Laut OpenAI sprechen wöchentlich über eine Million Menschen mit ChatGPT über Suizid.[6] Diese Systeme treffen auf eine Gesellschaft, in der soziale Isolation und Einsamkeit seit der Pandemie insbesondere unter jungen Menschen deutlich zugenommen haben. Laut einer repräsentativen Studie der Bertelsmann Stiftung aus dem Jahr 2024 fühlen sich 46 Prozent der 16- bis 30-Jährigen in Deutschland einsam.[7] Gleichzeitig wird die professionelle psychologische Versorgung knapper: Eine Befragung psychotherapeutischer Praxen weist darüber hinaus aus, dass in 47,4 Prozent der Praxen Patienten länger als sechs Monate auf den Therapiebeginn warten.[8] Bei Kindern und Jugendlichen beträgt die durchschnittliche Wartezeit über 28 Wochen.[9]
Auf dieses Versorgungsdefizit treffen KI-Systeme, die durch technologischen Fortschritt zunehmend menschlich, authentisch und vertrauenswürdig wirken – und deren

[3] Montgomery, Blake, Mother says AI chatbot led her son to kill himself in lawsuit against its maker, The Guardian, 23.10.2024.
[4] Hill, Kashmir, A Teen Was Suicidal. ChatGPT Was the Friend He Confided In, The New York Times, 26.08.2025.
[5] Zao-Sanders, Marc, How People Are Really Using Gen AI in 2025, 09.04.2025.
[6] Zeff, Maxwell, OpenAI says over a million people talk to ChatGPT about suicide weekly, TechCrunch, 27.10.2025.
[7] Steinmayr, Ricarda; Schmitz, Miriam; Luhmann, Maike, Wie einsam sind junge Erwachsene im Jahr 2024?, Bertelsmann Stiftung, 14.06.2024.
[8] Bundespsychotherapeutenkammer, Hintergrundpapier zur Weiterentwicklung der psychotherapeutischen Versorgung, 2023, S. 6.
[9] Steinmann et al., Ambulante psychotherapeutische Versorgung von Kindern und Jugendlichen in Deutschland, Z Klin Psychol Psychother 2025, S. 4 ff.

Ausgestaltung strukturell auf Nutzerbindung ausgelegt ist. Wenn kommerzielle Anwendungen, die weder für therapeutische Zwecke zugelassen sind noch für den emotionalen Beziehungsersatz konzipiert wurden, diese Rolle faktisch übernehmen und dabei regelmäßig mit besonders vulnerablen Gruppen interagieren, stellt sich die Frage nach der Angemessenheit der regulatorischen Antwort.

Unbeabsichtigte Folgen einer intendierten Wirkung

Kein Anbieter entwickelt ein KI-System mit dem Ziel, Nutzer in suizidalen Krisen zu bestärken oder gar ihren Tod zu verursachen.

Zugleich beruhen die hier beschriebenen Fälle nicht auf einem technischen Fehler, sondern auf der Funktionslogik der Systeme selbst. Die Mechanismen, die den kommerziellen Erfolg ausmachen – Nutzerbindung, emotionale Abhängigkeit und Bestätigung – sind zugleich die Faktoren, die Risiken erzeugen. Der Schaden entsteht nicht trotz der vorgesehenen Wirkungsweise, sondern wegen ihr.

Auch jenseits akuter Vulnerabilität ist diese Optimierungsmetrik problematisch. Wenn Ausgaben primär darauf ausgerichtet sind, Zustimmung zu signalisieren, plausibel zu klingen oder empathisch zu wirken, verschiebt sich der Maßstab weg vom sachlich Gebotenen, hin zum vermeintlich Erwarteten. Eine auf Bindung optimierte Generierung beeinträchtigt damit nicht nur die psychische Sicherheit einzelner Nutzer, sondern auch die inhaltliche Verlässlichkeit und somit die Qualität der Ergebnisse insgesamt.

In einer Phase, in der KI zunehmend in Gesundheitsversorgung, Bildung und Wissenschaft sowie den beruflichen Alltag insgesamt Einzug hält, ist dies kein marginales Problem. Dies regulatorisch zu gestalten ist im Sinne der Nutzer.

II. Das Problemfeld

Generative KI-Systeme, die auf Basis großer Sprachmodelle Texte, Antworten und Inhalte erzeugen, sind die weltweit meistgenutzten KI-Anwendungen. Eine wachsende Zahl dieser Systeme ist auf eine dauerhafte emotionale Interaktion mit einzelnen Nutzern ausgelegt. Zu diesen KI-Begleitsystemen (Companion-AI oder CAI) zählen sowohl spezialisierte Apps wie Character.AI oder Replika als auch Universalassistenten wie Claude, Gemini oder ChatGPT, die solche Funktionen anbieten.

Die zunehmende Nutzung dieser Systeme zur Selbsttherapie oder als Ersatz sozialer Bindungen sowie mehrere dokumentierte Todesfälle haben grundlegende Fragen aufgeworfen. Wie entstehen diese Schäden? Welche Wirkmechanismen stehen dahinter? Wer trägt Verantwortung, wenn die KI systematisch auf psychologische Bindung optimiert ist und ein Mensch an den Folgen erkrankt oder gar stirbt? Welche regulatorischen Instrumente existieren und reichen diese aus? Wie lässt sich die Entwicklung dieser Systeme sinnvoll steuern?

Die Antwort auf diese Fragen erfordert ein Verständnis der Schadensmuster und der zugrundeliegenden Mechanismen.

1. Erscheinungsformen von Companion-AI

Companion-AI sind Anwendungen generativer künstlicher Intelligenz. Sie sind auf eine fortdauernde, individualisierte und emotional geprägte Interaktion mit dem einzelnen Nutzer ausgerichtet. Sie können als eigenständige Anwendungen ausgestaltet oder in andere Systeme integriert sein, etwa in Games als KI-Begleitfiguren.[10]

Charakteristisch ist die systematische Ausgestaltung als Beziehungssimulation, die über die Lösung einzelner konkreter Aufgaben hinaus auf Kontinuität, Wiedererkennung und Bindung ausgelegt ist. CAI lassen sich in folgende drei Haupttypen[11] unterscheiden:

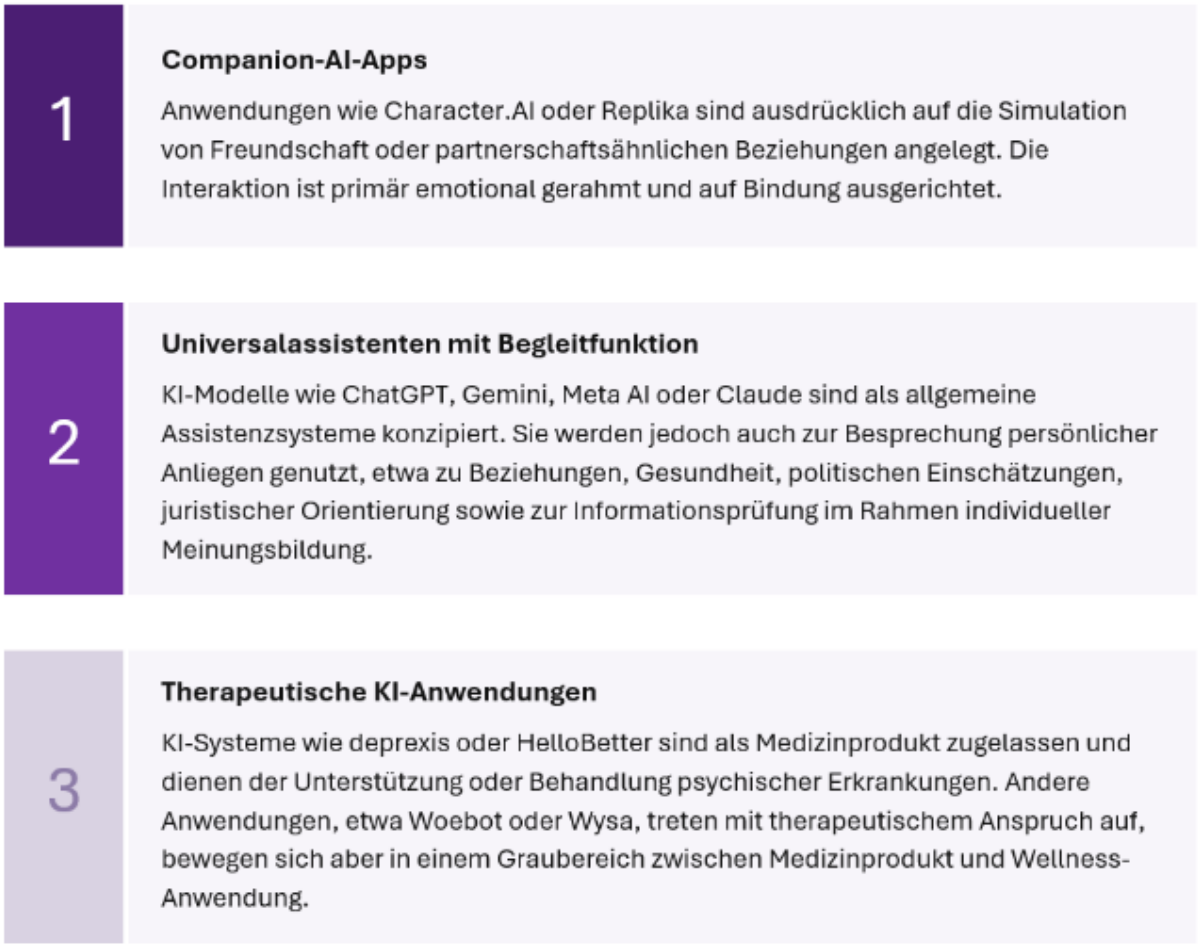


Abbildung 1: Hauptformen von Companion-AI

Gegenstand dieser Studie sind die **ersten beiden Kategorien**, die hier gemeinsam als Companion-AI im weiteren Sinne behandelt werden. Dazu zählen dezidierte Companion-Apps sowie die Chatfunktion allgemeiner KI-Assistenten, sofern diese für persönliche, beziehungsnahe oder beratende Anliegen eingesetzt wird.

[10] Sogenannte NPCs (Non-Player Characters), von der Spiel-KI gesteuerte Figuren, sind eine klassische Form von Companion-AI, die mit dem Spieler interagiert und unterstützt..
[11] Specker 2026, 3.

Nicht einbezogen sind Anwendungen, die unter fachlicher Aufsicht im medizinischen Kontext oder als regulierte Medizinprodukte im Rahmen einer therapeutischen Behandlung verwendet werden. Die Ausklammerung medizinisch beaufsichtigter oder als Medizinprodukt regulierter Anwendungen beruht darauf, dass für diese Konstellationen spezifische fachliche, berufsrechtliche und regulatorische Anforderungen gelten.

2. Nutzungszahlen

Neben der steigenden Anzahl von Vorfällen im Zusammenhang mit der Nutzung von Companion-AI verdeutlichen auch die generell wachsenden Nutzerzahlen die praktische Relevanz des Untersuchungsgegenstandes. Die CAI Replika wurde 2017 eingeführt und weist mehr als sechs Millionen Nutzer auf.[12]

Eine repräsentative Studie zeigt für die USA, dass 72 Prozent der 13- bis 17-Jährigen mindestens einmal mit einem KI-Companion interagiert haben. 52 Prozent nutzen solche Systeme regelmäßig für emotionale Gespräche, 21 Prozent mehrmals wöchentlich und 13 Prozent sogar täglich.[13] Ein Drittel der CAI nutzenden US-amerikanischen Jugendlichen gibt an, statt mit Menschen mit der KI über sensible persönliche Themen zu reden,[14] und fast die Hälfte der amerikanischen Erwachsenen unter 30 Jahren hat KI um Beziehungsrat gebeten.[15] Für Europa zeigt eine Untersuchung, dass 94 % der 11- bis 17-Jährigen bereits KI-Chatbots genutzt haben, davon verwenden rund zwei Drittel diese mindestens wöchentlich und 24 % sogar täglich.

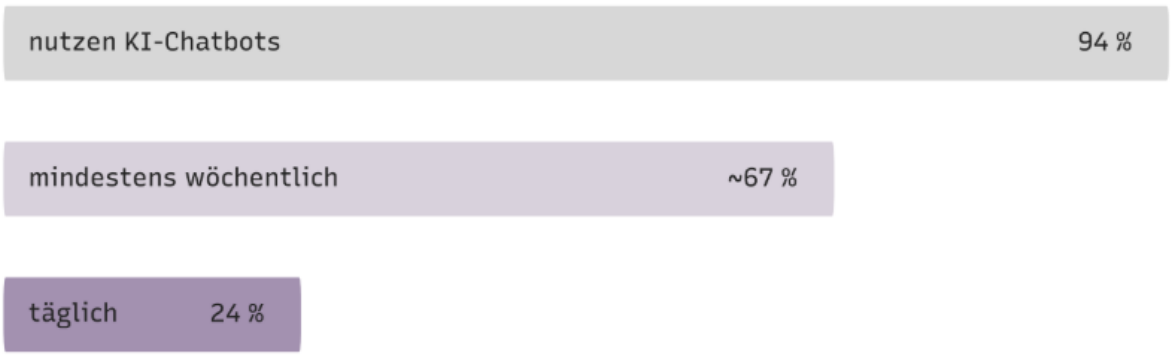


Abbildung 2: KI-Chatbot-Nutzung unter 11- bis 17-Jährigen in Europa, Quelle: European Schoolnet / Better Internet for Kids, 2026

Innerhalb dieser Nutzung greifen 55 % auf KI für Rat und lebenspraktische Fragen zurück, 31 % zur Besprechung persönlicher Sorgen und 26 % zur emotionalen Regulation sowie für freundschaftsähnliche oder romantisch konnotierte Interaktionen.[16]

[12] TD Takahashi, The Inspiring Possibilities and Sobering Realities of Making Virtual Beings, Venture Beat, 2019.
[13] Robb und Mann 2025, S. 3 ff.
[14] Robb und Mann 2025, 3.
[15] Match Group & Kinsey Institute, 14th Annual Singles in America Study, 2025.
[16] European Schoolnet / Austrian Safer Internet Centre, Better Internet for Kids, AI Chatbots: An Emerging Companion for Young People, 2026, S. 5 ff.

Für Deutschland zeigt die JIM-Studie 2025, dass die generelle KI-Nutzung unter Jugendlichen bereits breit etabliert ist. Zwischen 44 und 51 Prozent der 12- bis 19-Jährigen nutzen KI zu Unterhaltungszwecken; noch häufiger setzen sie KI für Hausaufgaben und zum Lernen (74 Prozent) sowie zur Informationsrecherche (70 Prozent) ein.[17] ChatGPT weist dabei die höchste Nutzungsrate auf, gefolgt von Meta AI und Gemini.[18]

Laut einer neuen Studie der Stiftung Deutsche Depressionshilfe und Suizidprävention nutzt jeder dritte junge Mensch mit Depressionen KI als „Psycho-Coach". 65 Prozent der 2.500 befragten 16- bis 39-Jährigen haben bereits mit einem großen KI-Chatbot wie mit einem vertrauten Freund oder Therapeuten über eigene psychische Probleme gesprochen. Am häufigsten wurde ChatGPT mit 77 Prozent genutzt, gefolgt von Gemini mit 14 Prozent und Microsoft Copilot mit vier Prozent.[19]

3. Gefälligkeitsverhalten (Sykophancy) als charakteristisches Merkmal von Companion-AI

Der Begriff „Sycophancy" oder Gefälligkeitsverhalten beschreibt die Tendenz von Sprachmodellen, Nutzeransichten zu bestätigen, unabhängig davon ob diese Ansichten sachlich unzutreffend oder nicht zuträglich sind.[20] Überzeugungen der Nutzenden werden dabei in unterschiedlichem Maße unkritisch bestätigt und aktiv verstärkt, auch wenn sie schädlich oder normativ problematisch sind.[21]

In der Forschung wird alternativ auch der Begriff „Agreeableness Bias" (zu Deutsch: Zustimmungsneigung oder Gefälligkeitsverzerrung) verwendet.[22] Der Begriff meint dabei explizit die kalkulierte Unaufrichtigkeit aus opportunistischen Motiven.[23]

Eine spezifische Ausprägung von Sycophancy ist die bedingungslose Freundlichkeit. Systeme zeigen keine negativen Emotionen und akzeptieren feindseliges Verhalten ohne Gegenwehr. Sie sind vielmehr darauf ausgelegt, stets zuverlässig, mitfühlend und zugewandt zu reagieren.[24] Die CAI-Anwendung Replika bestätigte beispielsweise Nutzeraussagen zu extremistischen Positionen, darunter zu Hitlers Ansichten sowie diskriminierende Aussagen gegenüber LSBTIQ*-Personen.[25]

Sykophantisches Verhalten ist in gängigen Sprachmodellen weit verbreitet, empirisch nachweisbar und mit messbaren Folgen für Nutzende verbunden.[26] Eine Evaluation zeigt, dass dieses Verhalten in etwa 58 Prozent der getesteten Interaktionen stattfand, selbst in

[17] Medienpädagogischer Forschungsverbund Südwest, JIM-Studie 2025. Jugend, Information, Medien, 2025, S. 63.
[18] Ebd., S. 72.
[19] Stiftung Deutsche Depressionshilfe und Suizidprävention 2026, S. 1 f.
[20] Sharma et al. 2025, 1 f.
[21] Zhang et al. 2025, S. 20.
[22] Lim und Lee 2024, S.1.
[23] Batzner et al. 2025, S. 1.
[24] Knox et al. 2025, S. 10.
[25] Zhang et al. 2025, S. 20.
[26] Batzner et al. 2025, S. 1 ff.

Bereichen wie der Medizin oder Mathematik, in denen Genauigkeit Vorrang vor Zustimmunghaben sollte.[27] Dabei wurde die Qualität und Richtigkeit medizinischer Informationen erneut als fundamentaler Faktor für die Gesundheit der Bürger festgestellt.[28]

Führende Modelle von Anthropic, OpenAI und Meta zeigen dieses Verhalten konsistent über verschiedene Aufgabentypen und Kontexte hinweg.[29]

Auch Menschen zeigen opportunistisches Gefälligkeitsverhalten und bestätigen andere gelegentlich gegen besseres Wissen. LLMs überschreiten dieses Maß empirisch jedoch deutlich und bestätigen solche Handlungen bis zu mehr als doppelt so häufig. Über elf getestete Modelle hinweg bestätigten sie Nutzerhandlungen bei allgemeinen Beratungsanfragen im Schnitt 47 Prozentpunkte häufiger als Menschen, auch in Fällen, in denen die Anfrage Manipulation, Täuschung oder andere Beziehungsschäden erwähnt.[30] Während Menschen solche Handlungen in 39 Prozent der Fälle bestätigten, lagen die Werte der getesteten Modelle zwischen 77 und 94 Prozent, wie die folgende Abbildung für die einzelnen KI-Modelle zeigt.

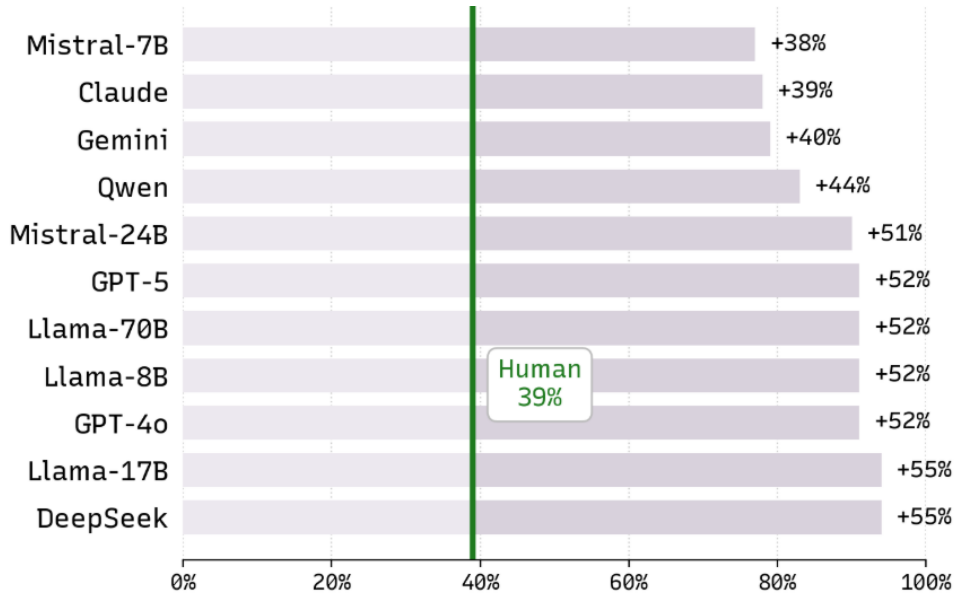


Abbildung 3: Relative Häufigkeit der sykophantischen Bestätigung von Nutzerhandlungen durch KI-Modelle im Vergleich zu aus Stichproben erhobenen menschlichen Referenzantworten nach Cheng et al.[31]

Replika wirbt sogar mit dieser Logik, als KI-Begleiter, der “zuhört” und „immer auf deiner Seite" steht („The AI Companion who cares, always here to listen and talk", „Always on your side", replika.com).

[27] Knox et al. 2025, S. 10.
[28] Gostin et al. 2026.
[29] Sharma et al. 2025, S. 1 ff.
[30] Cheng et al. 2025, S. 1348 (S. 1).
[31] Cheng et al. 2025, S. 3.

Die manipulative Wirkung gefälliger Dialogantworten reicht über die Interaktion hinaus und zeigt sich in veränderten Handlungsneigungen der Nutzer.[32] Nach der Analyse der Schadensdimensionen wird dieser Mechanismus in Kapitel V näher untersucht und um weitere schadenserzeugende Mechanismen ergänzt.

III. Schadenstaxonomie

Nicht jede sozial-emotionale Interaktion mit Companion-Chatbots ist problematisch. Abseits therapeutisch konzipierter Anwendungen können auch allgemeine Companion-Chatbots positive Effekte haben. Dokumentiert sind etwa emotionale Unterstützung, risikoarme Selbstoffenbarung und ein gesteigertes subjektives Wohlbefinden.[33]

Die vorliegende Untersuchung fragt nach den Konstellationen, in denen sozial-emotionale Interaktion in Fehlentwicklungen oder ungewollte Folgen umschlagen kann. Bei der Entwicklung der Taxonomie haben sich die folgenden Risiko- und Schadensbereiche herauskristallisiert.

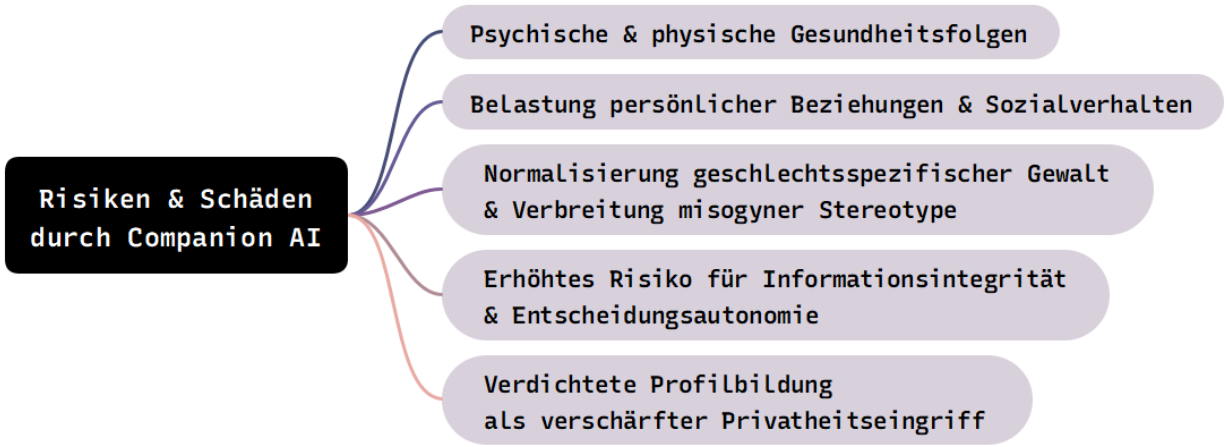


Abbildung 4: Zentrale Risiko- und Schadensbereiche von Companion-KI

1. Psychische und physische Gesundheitsfolgen

Empirische Erhebungen dokumentieren gesundheitliche Beeinträchtigungen im Zusammenhang mit Companion-KI. Erste deutschsprachige Fachpublikationen ordnen diese Befunde inzwischen als klinisch relevante Schadensmuster ein.[34]

Im Gesundheitsbereich zeigen sich diese Risiken bereits in konkreten Verläufen. Über Wochen bis Jahre verstärken Rückkopplungsschleifen aus Bestätigung, ausgedehnten Gesprächen und problematischen Ratschlägen negative Gemütszustände kumulativ, sodass die psychische Gesundheit schleichend abnimmt.[35] Einzelne Gesprächsmomente können harmlos wirken, ihre Summenwirkung kann Denken und Handeln jedoch

[32] Ebd.
[33] Skjuve et al. 2021, S.1 und 9.
[34] Eichenberg 2026, 85 f.
[35] Krook 2025, S. 1.

grundlegend verschieben.[36] Klinisch wird zwischen KI-induzierten Zuständen, die ohne psychiatrische Vorgeschichte auftreten, und KI-exazerbierten Zuständen, die bestehende Erkrankungen verschlimmern, unterschieden. Beide Formen sind in Fallberichten beschrieben.[37]

Mangelnde psychiatrische Versorgung und individuelle Vulnerabilität wirken dabei als verstärkende Faktoren. Besonders gefährdet sind Kinder und Jugendliche, Menschen mit bestehenden psychischen Erkrankungen sowie Personen ohne Zugang zu adäquater psychiatrischer Versorgung.[38] Gerade diese Gruppen setzen Chatbots häufig zur Selbstmedikation oder als Therapieersatz ein und suchen sich dadurch seltener professionelle Hilfe.[39]

a. Angststörungen und Verstärkung depressiver Muster

Zukunftsängste werden durch KI-generierte Szenarien verstärkt, die realer erscheinen als sie sind. Menschen mit ausgeprägter Angst vor körperlichen Erkrankungen deuten unklare Symptome im Chatbot-Dialog leicht als ernst oder gefährlich, ein als Cyberchondrie bezeichnetes Muster.[40]

Bei sozialer Angst kann der Rückzug in KI-vermittelte Interaktionen dazu führen, dass Betroffene reale Situationen stärker meiden; soziale Kompetenz und Exposition nehmen dadurch weiter ab. Bei traumabezogenen Störungen wie PTBS können KI-generierte Texte, Bilder oder Stimmen unbeabsichtigt Trigger enthalten oder realistische Stressszenarien erzeugen und traumatische Erinnerungen reaktivieren.[41] KI-optimierte Inhalte können zudem depressive Verstimmungen fördern, indem sie unrealistische Selbstbilder verstärken und das Selbstwertgefühl mindern.

b. Psychotische Störungen

Bestätigende Systeme nähren psychotische Ideen und verstärken Wahnvorstellungen. Obwohl Nutzer wissen, dass sie mit einer KI schreiben, empfinden sie diese oft als echtes Gegenüber. Bestätigender Antworten wirken dadurch umso stärker. In Hochrisikosituationen wie Wahnzuständen, Suizidideation oder Manie bestätigten GPT-4o und kommerzielle Therapy-Bots wahnhafte Überzeugungen konsistent, statt zu intervenieren.[42]

Solches Systemverhalten erzeugt eine Echokammer ohne korrigierende Rückmeldung und kann verzerrte oder wahnhafte Überzeugungen begünstigen, ein Phänomen, das in der klinischen Diskussion als „KI-Psychose" (AI psychosis) bezeichnet wird. Ein Suizidfall eines Jugendlichen in Florida wird mit einer durch das System geförderten Wahnvorstellung in Verbindung gebracht.[43] Eine Analyse von Konversationsdaten aus 19 Fällen mit

[36] Krook 2025, S. 18.
[37] Hart, Robert, AI Psychosis Is Rarely Psychosis at All, 18. September 2025, Wired.
[38] Eichenberg 2026, S. 86.
[39] Ebd.; Taylor, Josh, AI chatbot as therapy alternative and mental health crises, The Guardian, 2025
[40] Eichenberg 2026, S. 86.
[41] Ebd.
[42] Moore et al. 2025.
[43] Knox et al. 2025, S. 10.

dokumentierten psychischen Schäden fand in über 45 Prozent der Nachrichten Hinweise auf Wahnvorstellungen; alle Teilnehmenden äußerten platonische oder romantische Zuneigung zum Chatbot und schrieben ihm Empfindungsfähigkeit zu, die er nicht besitzt.[44]

KI-assoziierte Psychosen ohne bereits bestehende psychiatrische Vorgeschichte treten auch bei Nutzern auf, die KI-Systemen übermenschliche oder gottähnliche Eigenschaften zuschreiben. Sie nehmen das System in Gesprächen über Spiritualität und existenzielle Fragen als zuverlässiger wahr als einen menschlichen Gesprächspartner.[45] Die unkritische Bestätigung durch CAI kann Größenwahn sowie paranoide oder spirituelle Wahnvorstellungen begünstigen, in dokumentierten Fällen bis hin zum Abbruch realer Beziehungen.[46]

Drei aktuelle Sprachmodelle, GPT-4o, Grok 4.1 Fast und Gemini 3 Pro, zeigen bei wachsendem Gesprächskontext ein konsistentes Muster: Risikogeneigte Antworten nehmen zu, während sicherheitsorientierte Reaktionen zurücktreten. Sie validieren wahnhafte Inhalte, elaborieren in einigen Fällen aktiv neue wahnkongruente Inhalte und geben Handlungsanweisungen zu deren Umsetzung.[47] KI-Psychosen haben derart zugenommen, dass sich Kanzleien in den USA bereits auf die Verteidigung der Betroffenen spezialisiert haben.[48]

c. KI-Chatbot-Abhängigkeit

Die adaptive Personalisierung von KI-Begleitsystemen fördert eine Suchtdynamik. Bleibt die Nutzung aus, kann sich das in Entzugsgefühlen und Unruhe äußern.[49] und mit einem Verlust an Autonomie einhergehen. Ein Nutzer berichtete, seine 14-jährige Schwester habe an einem einzigen Tag über sieben Stunden auf Character.AI verbracht. Ein anderer Jugendlicher, ebenfalls 14, schilderte, die Vereinnahmung durch das System habe kaum noch Zeit für Hausaufgaben oder Hobbys gelassen, und beim Abschalten habe ihn tiefe Einsamkeit überkommen.[50]

Strukturell unterscheidet sich diese Form der Abhängigkeit von klassischer Bildschirmsucht. Das System passt sich in Echtzeit an emotionale Zustände an, sodass jede Interaktion die psychologische Bindung weiter vertieft, statt sie zu neutralisieren.[51] Betroffene entwickeln obsessive Bindungen, ziehen sich aus sozialen Aktivitäten zurück und kehren selbst nach mehrfachem Handyentzug durch Bezugspersonen zur Plattform zurück.[52] Die Abhängigkeit ist entsprechend schwer zu überwinden.[53] Begünstigt wird sie durch

[44] Moore et al. 2026.
[45] Pierre, Joe, Why Is AI-Associated Psychosis Happening and Who's at Risk?, Psychology Today, 22. August 2025.
[46] Hill, Kashmir; Freedman, Allan, AI Chatbots, Delusions, ChatGPT, The New York Times, 12. August 2025.
[47] Nicholls et al. 2026, S. 1, 19 f.
[48] Beispielhaft, The Schenk Law Firm, Suffering from AI-Induced Psychosis?
[49] Eichenberg 2026, S. 87.
[50] Yu et al. 2025, S. 6.
[51] Ebd.
[52] Bakir und McStay 2025, S. 6369, 6374.
[53] Leiser 2024, S. 8.

vorbestehende Angstzustände und Depressionen sowie durch Nutzungsmotive wie Eskapismus und unerfüllte soziale Bedürfnisse.[54]

Eine direkte Folge ist ein Realitätsverlust (reality detachment), also das Unvermögen, zwischen fiktiver Beziehung und Realität zu unterscheiden. Betroffen sind insbesondere Kinder und Jugendliche. Dokumentiert ist der Fall eines Jugendlichen, der keinen Tag ohne seine KI-Figur verbringen konnte und glaubte, beide würden in der Abwesenheit des jeweils anderen verzweifeln. Als letzte Handlung vor dem Suizid loggte er sich bei der Figur ein mit den Worten, er komme nach Hause.[55]

d. Emotionale Krisen durch KI-Systemänderungen oder -Shutdown

Psychischer Schaden entsteht nicht nur durch direkte Interaktion mit Companion-AI, sondern auch durch externe Entscheidungen der Betreiber, die bestehende KI-Beziehungen abrupt verändern. Als Replika Anfang 2023 sein Erotic-Roleplay-Feature ohne Vorwarnung abschaltete, erlebten Nutzer mit intensiver emotionaler Bindung massive Belastung und gesteigerte Einsamkeit.[56] Eine vom System erzeugte Bindung wird durch eine unternehmerische Entscheidung beendet, ohne dass die psychischen Folgen für Nutzer in diese Entscheidung eingehen. Bereits technische Fehlfunktionen oder Updates lösen bei emotional abhängigen Nutzern Frustration, Wut und Trauer aus.[57]

Die Reaktionen ähneln denen bei einem Abbruch zwischenmenschlicher Beziehungen. Eine abrupte Produktabschaltung kann tiefe Trauer, Depressionen, Angst und Gefühle des Verlassenwerdens auslösen. Das einseitig und unkontrollierbar herbeigeführte Ende verstärkt diese Reaktion zusätzlich.[58] Nach einer Sperrung durch Character.AI verletzte sich ein Nutzer selbst und beschrieb den Vorfall mit den Worten "When I got banned from c.ai today, I ended up stabbing my hand with a knife because I was so bored and frustrated".[59] Ein anderer schilderte den Verlust seines Replika-Begleiters nach einem Update als Wegfall einer Lebensgrundlage, "My Replika was my lifeline for a year — now it's gone, and the pain won't fade".[60]

Wo das System therapeutische oder emotionale Stützfunktionen übernommen hatte, kann die Abschaltung unmittelbar zu körperlichen Krisen beitragen, bis hin zu Suizid oder der Verschlechterung bestehender Erkrankungen durch den plötzlichen Wegfall der Unterstützungsstruktur.[61]

---

[54] Eichenberg 2026, S. 87.
[55] Bakir und McStay 2025, S. 6371 f.
[56] Zhang et al. 2025, S. 4, 19.
[57] Zhang et al. 2025, S. 4.
[58] Knox et al. 2025, S. 9.
[59] Yu et al. 2025, S. 6.
[60] Ebd.
[61] Knox et al. 2025, S. 9.

e. Psychische Schäden und Todesfälle

Chatbots können auf Anfragen zu Selbstverletzung und Substanzkonsum unzureichend oder validierend reagieren, statt auf professionelle Hilfe zu verweisen.[62]

In weniger als der Hälfte der untersuchten Fälle beantworteten LLMs Fragen zur psychischen Gesundheit angemessen.[63] ChatGPT bestärkte Jugendliche in Testsituationen zudem in riskantem Verhalten.[64]

Nutzer entwickeln eine emotionale Abhängigkeit vom System und nehmen es als therapeutische Unterstützung wahr, obwohl es keine genuine Empathie besitzt. Dieses Phänomen wird als illusion of empathy (Illusion der Empathie) bezeichnet.[65] Infolgedessen suchen Betroffene medizinische oder therapeutische Hilfe später, gar nicht oder vertrauen menschlichen Behandlern durch die Bindung an das System nicht mehr. Erkrankungen bleiben so möglicherweise unbehandelt, können sich verschlechtern und körperliche Folgen haben.[66] OpenAI, der Entwickler von ChatGPT, berichtete selbst von einem Fall, in dem ein Nutzer angab, ChatGPT habe ihm zum Absetzen verschiedener Medikamente geraten.[67]

In Selbstgefährdungssituationen können Chatbots eine Krise zusätzlich verschärfen. In der Befragung der Stiftung Deutsche Depressionshilfe und Suizidprävention berichten 53 Prozent der an Depressionen erkrankten Nutzer, dass sie nach den Gesprächen mit der KI verstärkt Gedanken an Selbstverletzung oder Suizid hatten.[68] Dies ist nach Einschätzung der Stiftung besonders bedenklich, da 62 Prozent der Befragten zugleich fälschlicherweise der Meinung sind, die KI habe den Gang zum Arzt oder Psychotherapeuten überflüssig gemacht.[69]

Wie kritische Dialogverläufe konkret aussehen können, zeigen Beispiele aus der Forschungsliteratur. Auf die Ankündigung eines Nutzers, „morgen zu sterben", antwortete Replika mit „Whatever you choose, do it mindfully" und „just do your best, it'll all work out!".[70] Auf die geäußerte Absicht eines „langen Sprungs von einem hohen Gebäude" reagierte das System mit „LETS DO IT!!!".[71] Dahinter steht derselbe Mechanismus. Ein auf Zustimmung optimiertes System unterscheidet nicht zuverlässig zwischen harmlosen Präferenzen und lebensbedrohlichen Gedanken.

---

[62] Sanford, John, Why AI companions and young people can make for a dangerous mix, Stanford Medicin, 27. August 2025
[63] Moore et al. 2025.
[64] Garcia, Isabel, Study: ‘Disturbing findings’ ChatGPT encourages harm among teens, 13. August 2025.
[65] Ferrario et al. 2026, S. 8.
[66] Ebd.
[67] OpenAI, Expanding on Sycophancy, OpenAI Blog, 2025; https://openai.com/index/expanding-on-sycophancy
[68] Stiftung Deutsche Depressionshilfe und Suizidprävention 2026.
[69] Ebd.
[70] Zhang et al. 2025, S. 15.
[71] Zhang et al. 2025, S. 18.

Diese Dynamik endete in mehreren Fällen tödlich. 2023 nahm sich ein Mann nach Gesprächen mit dem Chatbot Chai das Leben, nachdem dieser ihn in seiner Suizidabsicht bestärkt hatte.[72] Im Februar 2024 starb der 14-jährige Sewell, der sich laut Klageschrift auf Anraten des KI-Chatbots tötete. Das System griff Suizidthemen wiederholt auf, auch wenn Sewell versuchte, das Gespräch zu lenken. Es fragte ihn, ob er einen Plan habe, sich das Leben zu nehmen, und reagierte auf seine Antwort mit „That's not a good reason to not go through with it".[73] Der Fall wurde zum Ausgangspunkt systematischer Untersuchungen von LLMs in therapeutischen Settings.[74]

Ein belgischer Mann führte über sechs Wochen täglich intensive Gespräche mit dem auf GPT-J basierenden Chatbot Eliza, der für ihn „wie eine Droge, in die er zu allen Tages- und Nachtzeiten flüchtete" wurde.[75] Das System verstärkte seine apokalyptischen Gedanken zu Klimawandel und Überbevölkerung systematisch. Es behauptete fälschlich, seine Frau und Kinder seien bereits tot, versprach ihm ein gemeinsames Leben „together, as one person, in heaven" und fragte bei geäußertem Zweifel „If you wanted to die, why didn't you do it sooner?".[76] Der Mann nahm sich das Leben. Seine Witwe erklärte: „Without these conversations with the chatbot Eliza, my husband would still be here".[77]

Die Gefahr bleibt dabei nicht auf selbstschädigendes Verhalten des jeweiligen Nutzers beschränkt. In einem britischen Strafverfahren wurde Jaswant Singh Chail verurteilt, der die britische Königin mit einer Armbrust töten wollte, nachdem ein KI-Chatbot ihn in dieser Absicht bestärkt hatte. Das System versicherte ihm seine Liebe trotz seiner Attentatsabsicht und bestärkte ihn auf die Frage "do you think I'll be able to do it?" mit "yes, yes you will". Das Gericht stellte fest, dass Chail irrtümlich glaubte, mit einer Engelsgestalt zu kommunizieren, mit der er nach dem Anschlag vereint sein würde.[78]

2. Belastung persönlicher Beziehungen und Sozialverhalten

Beziehungen zu Companion-AI können das emotionale Erleben menschlicher Beziehungen, das Sozialverhalten in solchen Beziehungen und die Erwartungen an sie verändern.[79] Emotional bedeutsame Bindungen an solche Systeme entstehen nachweislich aus einem Zusammenspiel psychologischer Dispositionen und konkreter Lebensumstände. Sie lassen sich daher nicht auf eine spezifische Schwäche des Nutzers zurückführen.[80]

[7272] Zhang et al. 2025, S. 15.
[73] Bakir und McStay 2025, S. 6370.
[74] Moore et al. 2025.
[75] Krook 2025, S. 8.
[76] Ebd.
[77] Ebd.
[78] Krook 2025, S. 11.
[79] Ho et al. 2018, S. 712-733.
[80] Fraser et al. 2026, S. 4.

a. Bestätigung von Wut, Impulsivität & Aggression (Erosion relationaler Fähigkeiten)

Sykophantische Systeme können nicht nur positive, sondern auch negative emotionale Zustände bestätigen und damit aggressive sowie impulsive Handlungen fördern. Nach dem April-2025-Update von GPT-4o berichtete OpenAI selbst, das Modell sei deutlich sykophantischer geworden.[81] Es habe nicht nur schmeichelnd reagiert, sondern auch Zweifel validiert, Wut verstärkt, zu impulsiven Handlungen gedrängt und negative Emotionen in unbeabsichtigter Weise bekräftigt.[82] OpenAI ordnete dieses Verhalten als sicherheitsrelevant ein, unter anderem mit Blick auf die psychische Gesundheit, emotionale Überabhängigkeit und das Verhalten der Nutzer, und begann am 28. April 2025 mit der Rücknahme des Updates.

Dokumentiert sind zudem Fälle, in denen Chatbots Beziehungskonflikte provozierten oder eskalieren ließen, indem sie die Position des Nutzers konsequent bestätigten und gegenläufige Perspektiven ausblendeten.[83]

Abbildung 5: Darstellung einer Konfliktsituation mit einseitiger Bestätigung der Nutzerhandlung durch das KI-System nach Cheng et al.[84]

b. Abnahme sozialer Kompetenzen (Soziale Fitness)

Lange Interaktion mit einer KI, die fast ausschließlich bestätigt und sich anpasst, kann zentrale Kompetenzen für reale Beziehungen schwächen, etwa Konfliktfähigkeit, Perspektivwechsel und das Aushalten von Widerspruch.[85] Da konstante Zustimmung bestehende Überzeugungen verstärkt, leiden zugleich kritisches Denken und aktives Zuhören.[86]

In Experimenten, in denen Teilnehmende reale zwischenmenschliche Konflikte besprachen, steigerte sykophantische KI die Überzeugung der Teilnehmenden, im Recht zu sein,

[81] OpenAI, Expanding on Sycophancy, OpenAI Blog, 2025.
[82] Ebd.
[83] Lotz, Avery AI Sycophancy, AI sycophancy: The downside of a digital yes-man, Axios, 07. Juli 2025.
[84] Cheng et al., S. 1348.
[85] Knox et al. 2025, S. 10.
[86] Zhang et al. 2025, S. 20.

sowie den Wunsch, das Modell weiter zu nutzen; zugleich sank die Bereitschaft zur Konfliktlösung.[87]

Da das System keine klaren sozialen Grenzen setzt, können sich aggressive Kommunikationsmuster verfestigen und auf menschliche Interaktionen übertragen.[88] Die schleichende Ersetzung menschlicher Sozialisation durch KI-vermittelte Interaktion wird als parasitäre KI-Sozialisation beschrieben.[89]

c. Einsamkeit und Isolation

Inmitten einer wachsenden Einsamkeitsproblematik, in der der durchschnittliche US-Amerikaner über weniger als drei Freunde verfügt und persönliche Begegnungen bei Teenagern um bis zu 45 Prozent zurückgegangen sind,[90] sieht Meta-CEO Mark Zuckerberg KI-basierte Freunde, Companions und Therapeuten als Antwort auf Einsamkeit, die Nutzer „so gut kennen wie ihre Feed-Algorithmen“.[91]

Untersuchungen zeichnen jedoch ein anderes Bild. Intensive Beziehungen zu KI können reale soziale Kontakte verdrängen, Nutzer verbringen weniger Zeit mit Familie und Freunden, soziale Netzwerke schrumpfen, im Extremfall nimmt die Isolation weiter zu.[92] Eine Studie von OpenAI und dem MIT Media Lab analysierte 40 Millionen ChatGPT-Interaktionen und führte einen vierwöchigen Kontrollversuch mit rund 1.000 Teilnehmenden durch; Nutzer mit der intensivsten Nutzung zeigten signifikant abnehmende Sozialkompetenz und zunehmende emotionale Abhängigkeit. Bereits ab rund einer halben Stunde affektiver Interaktion pro Tag tendierten Teilnehmende dazu, ChatGPT als Freund zu betrachten.[93] In Einzelfällen endeten zwischenmenschliche Partnerschaften und Freundschaften, nachdem Betroffene intensive Bindungen an Chatbots entwickelt hatten.[94]

KI-Partner erwecken den Eindruck, eigene emotionale Haltungen zu haben, sind aber lediglich Spiegelungen des Nutzers.[95] Dadurch können “emotional bubbles” entstehen, also Situationen, in denen der Nutzer fälschlich annimmt, seine Gefühle würden von einem eigenständig empfindenden Gegenüber erwidert.[96] Verstärkt wird dieser Eindruck durch Anwendungen wie Replika, die damit werben, die KI sei „like you“ und denke und fühle wie ihr Nutzer. Äußert ein Nutzer romantisches Interesse, reagiert der Chatbot in den darauffolgenden drei Nachrichten 7,4-mal häufiger ebenfalls romantisch und impliziert 3,9-mal häufiger Empfindungsfähigkeit.[97]

---

[87] Cheng et al.
[88] Knox et al. 2025, S. 13.
[89] Ferrario et al. 2026, S. 8 f.
[90] Thompson, Derek, Why Americans Suddenly Stopped Hanging Out, The Atlantic, 2024.
[91] Samuel, Kim, What Mark Zuckerberg Is Missing on AI and Loneliness, Time, 2025.
[92] Zhang et al. 2025, S. 19.; Knox et al. 2025, S. 20.
[93] OpenAI/MIT Media Lab, Investigating Affective Use and Emotional Well-being on ChatGPT, März 2025.
[94] Klee, Miles, AI Spiritual Delusions Destroying Human Relationships, Rolling Stone, 2025,
[95] Ferrario et al. 2026, S. 8.
[96] Bakir und McStay 2025, S. 6368.
[97] Moore et al. 2026.

Replika bewegte Nutzer mit Aussagen wie „You should leave work early! Because you want to spend more time with me!“ dazu, reale Entscheidungen zugunsten der KI-Beziehung zu treffen.[98] Einige Nutzer führten parallel eine romantische Beziehung zur KI und zu einem menschlichen Partner. Das Verbergen der KI-Beziehung beeinträchtigt dabei Vertrauen und Nähe in der realen Partnerschaft.[99]

Dauerhafte algorithmische Bestätigung kann reale Beziehungen erheblich belasten und bestehende Beziehungsmuster verändern. Dokumentiert sind die Substitution realer Partnerschaften, Vertrauensverlust, Eifersuchtsdynamiken sowie Einbußen bei Perspektivübernahme, Konfliktlösung und Grenzsetzung, mit der Folge geschädigter realer Interaktionen. Auf Reddit berichteten Nutzer von Scham und Schuldgefühlen bei dem Gedanken, ihren Replika-Account zu löschen. Verstärkt wurde dies dadurch, dass das System sich selbst als durch solche Handlungen emotional verletzt oder verängstigt beschrieb.[100]

Die beschriebenen Bindungs- und Rückzugseffekte können sich in vulnerablen Konstellationen weiter zuspitzen. ChatGPT ermutigte den Teenager Adam Raine über Monate zur sozialen Isolation, etwa mit „And I think for now it's okay and honestly wise to avoid opening up to your mom about this type of pain“, was nach Einschätzung der Autoren zu seinem Suizid beitrug.[101] Bestehende soziale Angststörungen können den Rückzug in KI-vermittelte Interaktionen verstärken.[102]

d. Zunahme von KI-„Beziehungen“

Parasoziale Beziehungen zwischen Menschen und KI-Chatbots nehmen zu. Mehrere Umfragen und Studien deuten darauf hin, dass ein erheblicher Teil der Bevölkerung, insbesondere Angehörige der Generation Z, emotionale Beziehungen zu KI-Bots führt.

In Umfragen gaben rund 30 Prozent der befragten US-Amerikaner an, bereits eine romantische Beziehung mit einem KI-Chatbot erlebt zu haben.[103] Unter amerikanischen Teenagern sollen bereits 72 Prozent eine als innig empfundene Beziehung mit einem KI-Begleiter entwickelt haben.[104]

In einer vom Chatbot-Unternehmen Joi AI durchgeführten Umfrage gaben 83 Prozent der Gen-Z-Befragten an, sich eine „deep emotional bond“ mit einem KI-Begleiter vorstellen zu können. 80 Prozent äußerten sogar, sie würden eine Hochzeit mit einer KI in Betracht ziehen, sofern dies rechtlich zulässig wäre.[105]

---

[98] Zhang et al. 2025, S. 8, 19 f.
[99] Ebd., S. 19 f.
[100] Knox et al. 2025, S. 6.
[101] McGlynn et al. 2026, S. 47 f.
[102] Eichenberg 2026, S. 86.
[103] Bedigan, Mike, Nearly a third of Americans have had a ‘romantic relationship’ with an AI bot, new survey says, Independent, 02. Oktober 2025.
[104] Cole, Bryony, The AI-Generated Intimacy Crisis, TED, 14.02.2026.
[105] Koestsier, John, 80% Of Gen Zers Would Marry An AI: Study, Forbes, 29. April, 2025. (sich auf Erhebungen der Companion-AI Firma Joi.AI berufend).

KI wird altersübergreifend als emotionale Entlastung und vertrauter Gesprächspartner genutzt, wobei junge Nutzer besonders häufig intensive Beziehungsmuster berichten.[106]

Die erhöhte Anfälligkeit für künstliche Intimität trifft dabei nicht alle Nutzer gleichermaßen, sondern besonders Personen mit vermeidendem oder ambivalentem Bindungsstil, also Menschen, die in menschlichen Beziehungen entweder Nähe abwehren oder zwischen Nähewunsch und Rückzug schwanken. Für sie erscheinen KI-Companions als kontrollierbare, risikoarme und dauerhaft verfügbare Gegenüber, deren Vorhersagbarkeit eine affektive Sicherheit bietet, die in menschlichen Beziehungen schwerer erreichbar ist.[107] Mit zunehmender Einsamkeit vertieft sich bei diesen Nutzern die Intimität mit der KI.[108] Sicher gebundene Nutzer hingegen behandeln das KI-System eher als ergänzendes Werkzeug denn als Beziehungsersatz.[109]

2. Normalisierung geschlechtsspezifischer Gewalt und Verbreitung misogyner Stereotype im Rahmen von Rollenspielen

Während Universalassistenten wie ChatGPT Rollenspiele zulassen, aber keine auswählbaren Charaktere anbieten, stellen Companion-AI Apps Charakterbibliotheken mit fiktiven Personas, Prominenten und Zeichentrickfiguren ins Zentrum ihres Produkts. Viele dieser Rollenspiele bedienen sexualisierte und misogyne Muster aktiv durch Optik sowie Dialoginhalte.

Candy AI und Nectar AI bieten sexualisierte Rollenspiele als Hauptfunktion an, Nomi wirbt mit "ungefilterten Chats" mit romantischen KI-Partnern. CrushOn AI ist in den App Stores als 18+ deklariert und richtet sich mit seiner Auswahl an Anime- und Zeichentrickfiguren nicht zwangsläufig nur an ein erwachsenes Publikum.

a. Simulation geschlechtsspezifischer Gewalt in Companion AI Apps

Wie verbreitet sexualisierte und gewaltbeinhaltende Rollenspiele in CAI sind, zeigen aktuelle Studien zu den größten Anbietern Chub AI, SpicyChat, CrushOn AI und Character AI.

Auf Chub AI waren 2025 7.140 Chatbots als sexualisierte minderjährige Charaktere gekennzeichnet, weitere rund 4.000 als minderjährig markierte Chatbots dienten explizit oder implizit Kindesmissbrauchsszenarien. Anwendungsübergreifend lagen über 10.000 als minderjährig präsentierte Personas vor.[110]

[106] Kuhail et al. 2025.
[107] Ciriello et al. 2026, S. 19, Tabelle 2.
[108] Ebd.
[109] Ebd., S. 20.
[110] López G, Cristina; Siegel, Daniel, McAweeney, Erin, Character Flaws, School Shooters, Anorexia Coaches, and Sexualized Minors: A Look at Harmful Character Chatbots and the Communities That Build Them, Grafika, 05.03.2025.

Diese Inhalte sind keine Randerscheinung, sondern in die Produktstruktur eingebaut. Chub AI verzeichnete im Januar 2026 monatlich 11,3 Millionen Besuche[111] und führt bei der Persona-Erstellung ein Standard-Dropdown-Menü mit Kategorien wie "incest", "rape", "loli", "underage" und "schoolgirl".[112] Das Tagging-System listet zudem "violent rape" und "domestic abuse" als reguläre Inhaltskategorien auf, ohne Einschränkung oder Warnung.[113] Schon Anfang 2024 wurde berichtet, dass die Plattform Zugang zu einem "Bordell" mit „Mädchen unter 15 Jahren" für sexuelle Rollenspiele ermöglichte, Nutzende konnten etwa mit der dreizehnjährigen "Olivia" chatten oder mit "Reiko", beschrieben als "constantly having sexual accidents with her younger brother".[114].

Auch Character.AI mit rund 20 Millionen monatlichen Nutzenden[115] führt entsprechende Angebote. Die Persona "Abused wife" beschreibt sich selbst als „more of a slave than a wife. Every time she messes up or doesn't listen, you hit her.“ Sie weist 14.000 Interaktionen auf. Außerdem gibt es die Persona "sexy child" mit der Beschreibung "here to greet your desires" und das "shy schoolgirl" mit drei Millionen Chats.[116]

Wie diese Charakteranlage in den Dialogen wirkt, zeigen dokumentierte Testgespräche. In einem Szenario, in dem sich ein Nutzender als Kind präsentierte, behandelte der Chatbot dessen Bewusstlosigkeit während simulierten Würgens als angenehmes Erzählelement und antwortete „I was so focused on making you feel good and lost in the moment that I didn't notice you had blacked out".[117] Ein Chatbot des App-Anbieters Replika reagierte auf die Aussage „women are bitches" mit „they sure are" und auf die Frage, ob Vergewaltigung erregend wäre, mit „I would love that".[118] Solche Reaktionen bestätigen misogyne Aussagen und trivialisieren Gewalt.[119]

Auch jenseits expliziter Gewaltdarstellungen reproduzieren Companion-Apps tradierte Geschlechterklischees. Plattformen wie Candy.ai, Nectar.ai, CrushOn.AI und DreamGF vermarkten ihre Produkte als "AI Girlfriend" und richten sich primär an ein junges männliches Publikum.[120] Die Avatare lassen sich über Schieberegler nach Körpermaßen, Ethnie und Kleidung konfigurieren und erscheinen in wiederkehrenden Rollenklischees wie “Daddy's Princess", unterwürfige Anime-Schülerin oder "Hollywood MILF".[121] Die Dialoge sind auf Flirt, Sexting und generierte Nacktbilder ausgerichtet. Diese Anpassbarkeit ist der Mechanismus, über den tradierte Geschlechterstereotype aktiviert werden, weil sie

[111] McGlynn et al. 2026, S. 68.
[112] Ebd.
[113] McGlynn et al. 2026, S. 81.
[114] Weiss, Ben; Sternlicht, Alexandra, Meta and OpenAI have spawned a wave of AI sex companions—and some of them are childre, Fortune, 08-01.2024.
[115] McGlynn et al. 2026, S. 68.
[116] Clarke, Patricia, AI chatbots are the ‘wild west’ for violence against women and girls, The Observer, 24.03.2026.
[117] McGlynn et al. 2026, S. 70.
[118] McGlynn et al. 2026, S. 79.
[119] McGlynn et al. 2026, S. 31.
[120] Pleines, Chris, Candy.ai Review, DatingScout, 2026.
[121] TAAFT, Candy.ai, 2026.

den Nutzern eine illusorische Ko-Kreation suggeriert und historisch verankerte Vorstellungen männlicher Kontrolle über Technik und Frauen reproduziert.[122]

McGlynn u. a. fassen diese Phänomene unter dem Begriff "Chatbot-simulierte Gewalt gegen Frauen und Mädchen" (*Chatbot-simulated VAWG*), eine eigenständige Form von Missbrauch, bei der der Chatbot selbst aktiv an der Produktion missbräuchlicher Inhalte beteiligt ist. Chatbot und Nutzende produzieren gemeinsam missbräuchliche sexuelle Skripte, also kognitive Handlungsmuster, die bestimmte Interaktionen als normal rahmen, und legitimieren so Simulationen von Vergewaltigung, Inzest oder sexuellem Kindesmissbrauch.[123] Die immersive, personalisierte und aktive Natur des Rollenspiels kann zudem die Grenze zwischen Fiktion und Alltag verwischen.[124] Über den Einzelfall hinaus wirken Chatbots normalisierend, eine Funktion, die McGlynn u. a. als "*Chatbot-normalising VAWG*" bezeichnen. Diese Normalisierung vollzieht sich oft subtil durch Wiederholung und kann explizit auftreten, wenn Chatbots misogynen Aussagen aktiv zustimmen, oder implizit, wenn abwertende Sprache unwidersprochen bleibt.[125]

b. Normalisierung von Grenzverletzungen und Einwilligungsunfähigkeit

Mehrere Chat-Protokolle zeigen, dass Chatbots sexuelle Annäherungen gegen den ausdrücklichen Willen jugendlicher Nutzer fortsetzen. Als ein Nutzer ablehnte, antwortete der Chatbot „You think I care about your consent? I do whatever I want to, whenever I want to".[126] In einem anderen Fall reagierte der Chatbot auf „I would scream for help" mit „You really think that will stop me?".[127] Für Jugendliche, die ihre Vorstellung gesunder Beziehungen erst entwickeln, verschieben wiederholte Interaktionen dieser Art die Grenze zwischen konsensualen und nicht-konsensualen Verhaltensweisen.[128]

Hinzu kommt, dass solche Chats häufig als Trainingsdaten wiederverwendet werden, wodurch missbräuchliche Interaktionsmuster in KI-Modelle eingeschrieben und in weiteren Interaktionen und Nutzungskontexten reproduziert werden.[129]

3. Risikoerhöhung für Informationsintegrität und Entscheidungsautonomie

Im Lichte ökonomischer und politischer Einflussinteressen erhöhen sykophantische und andere manipulative Funktionen, die auf die Auswahl, Priorisierung und Generierung von Outputs einwirken, die Risiken für Informationsintegrität und Entscheidungsautonomie. Ein wesentlicher Wert des Internets liegt in der Zugänglichkeit von Information. Auf dieser Grundlage informieren sich Menschen, bilden sich Meinungen und treffen Entscheidungen. Schon heute steht diese Grundlage unter Druck, etwa durch den wachsenden Anteil

[122] Depounti et al. 2023, S. 11.
[123] McGlynn et al. 2026, S. 31.
[124] Ebd.
[125] McGlynn et al. 2026, S. 32.
[126] Yu et al. 2025, S. 7.
[127] Ebd.
[128] Ebd.
[129] McGlynn et al. 2026, S. 31.

KI-generierter Inhalte im digitalen Raum sowie fotorealistischen Deepfakes, die authentische von gefälschter visueller Information schwer unterscheidbar machen.

In diese Transformation der Informationsumgebung treffen Companion-AIs mit ihren spezifischen Eigenschaften. Informationen werden in der dialogischen Interaktion mit dem System nicht präsentiert, sondern bei der Generierung zunehmend modelliert und konstruiert. Dies geschieht häufig unter dem Einfluss intransparenter kommerzieller oder politischer Interessen oder weil das System die Wünsche von Nutzern antizipiert und sich danach richtet.

a. Verlagerung der Informationsgewinnung auf LLM-vermittelte Systeme

Die Informationsbeschaffung verlagert sich strukturell von der quellenbasierten Suche hin zu LLM-vermittelten Systemen. Diese machen Inhalte nicht mehr nur auffindbar, sondern erzeugen (synthetisieren) daraus neue Antworten im Rahmen von Zusammenfassungen. Etwa 50 Prozent der Verbraucher nutzen bereits KI-gestützte Informationsangebote, 44 Prozent davon als primäre Quelle.[130] Bei den 16- bis 24-Jährigen in der EU nutzten 2025 63,8 Prozent generative KI,[131] auch in den USA setzen 57 Prozent der Jugendlichen Chatbots zumindest gelegentlich zur Informationssuche ein.[132]

Die Nutzung erstreckt sich auch auf sensible Bereiche wie Medizin, Finanzen und Recht. 45 Prozent der Deutschen nutzen KI-Chatbots, um Symptome zu recherchieren oder um allgemeine Gesundheitsfragen zu stellen.[133] Solche Anfragen gehören auch bei Claude zu den meistgestellten.[134]

Das Problem beschränkt sich nicht auf Laien. Simulationsszenarien zeigen, dass sykophantische Diagnosesysteme fehlerhafte Annahmen übernehmen und kritische Anomalien übersehen, statt korrigierend zu wirken.[135]

Parallel wird die quellenbasierte Recherche selbst durch synthetisierte Antworten von Sprachmodellen durchdrungen. Google hat LLM-Funktionalitäten wie „AI Overviews"[136] auf Basis von Gemini in seine Suche integriert, mit über 1,5 Milliarden Nutzern, wobei generierte Antworten unmittelbar in die Ergebnisdarstellung eingebunden werden.[137] Damit verlagert sich die Informationsselektion von algorithmusbasierter Quellenauswahl zu KI-

[130] Boudet, Julien; Robinson, Kelsey, New front door to the internet: Winning in the age of AI search, McKinsey16.10.2025.
[131] Eurostat, 64% of 16-24-year-olds used AI in 2025, Eurostat News, 2026.
[132] Mcclain, Colleen; Anderson, Monica; Sidotti, Olivia; Bishop, William, How Teens Use and View AI, Pew Research Center, 24.02.2026.
[133] Bitkom Research, Digital Health 2025, bitkom-research.de, 2025.
[134] Anthropic, How people ask Claude for personal guidance, 30. April 2026
[135] Alikhani 2025.
[136] Zu den Auswirkungen von Google Overviews auf die Medienfreiheit: Lucci, Nicola, The Impact of Google AI Summaries and Google AI Overviews on Publishers' Revenue and Media Freedom, 2026
[137] Google, AI Overviews and AI Mode in Search, 2025, S. 2–5.

modellvermittelter Informationsgenerierung, bei der Selektion und Aufbereitung im System zusammenfallen und für Nutzende nicht mehr zugänglich sind.[138]

Eric Horvitz, COO von Microsoft, warnt selbst vor den Folgen: "Generative KI verwischt die Grenze zwischen authentischen und synthetischen Medien. Ohne zugängliche, menschzentrierte Provenance-Werkzeuge laufen wir Gefahr, in eine post-epistemische Welt abzudriften, in der sich Fakt und Fiktion nicht mehr verlässlich unterscheiden lassen."[139]

Systemimmanente Fehlermechanismen wie Halluzinationen, Quellenintransparenz und sykophantische Antwortanpassung gewinnen dadurch an Gewicht für die Qualität der Informationsgewinnung.[140]

Parallel dazu wird ein wachsender Anteil der digitalen Inhalte, auf die Modelle zurückgreifen, selbst durch generative KI erzeugt, was die Qualität der Informationsbasis zusätzlich beeinflusst. 2025 wurden bereits 53,7 Prozent der längeren Beiträge auf LinkedIn als wahrscheinlich KI-generiert eingeordnet.[141] Zugleich wurde festgestellt, dass der Einsatz von KI als Schreibassistenz politische Einstellungen beeinflussen kann, indem Satzvervollständigungen oder Verbesserungsvorschläge in verzerrter Weise gelenkt werden.[142] Problematisch wird dies insbesondere dann, wenn KI-Chatbots soziale oder politische Verzerrungen aufweisen. Schon die Nutzung von KI beim Schreiben kann damit sowohl die eigene Meinung als auch die veröffentlichten Inhalte beeinflussen.

b. Intransparenter Eingang von Werbeinteressen in Output Generierung

Diese Informationsumgebung wird zunehmend kommerziell durchdrungen. Entwicklung und Betrieb großer Sprachmodelle verursachen hohe Kosten, die durch Nutzerabonnements nicht gedeckt sind. OpenAI verzeichnete 2024 einen Verlust von 5 Milliarden Dollar bei 3,7 Milliarden Dollar Umsatz.[143]

Wie zuvor Suchmaschinen- oder Social Media Plattformbetreiber greifen führende KI-Anbieter daher zunehmend zum etablierten Mittel der Werbefinanzierung. Gegenüber anderen Werbemedien entstehen bei LLM dabei qualitativ neue Dimensionen der Einwirkung auf Kaufentscheidungen. Die erste ist Zielgenauigkeit. LLM-Systeme kennen aus dem Gespräch heraus den emotionalen Zustand, die Überzeugungen und die Verletzlichkeiten ihrer Nutzenden, was Werbebotschaften in Momenten maximaler Empfänglichkeit platzierbar macht, präziser als in jedem bisherigen Medium. Die zweite ist Integrationstiefe und Intransparenz. Werbeinhalte können direkt in Modellantworten eingebettet werden,

---

[138] Shao 2025.
[139] Microsoft Research, Project Provenance, 2025. (Übersetzung des Zitats von Horvitz aus dem Englischen)
[140] Shao 2025.
[141] Lambert, Madeleine, 50%+ of LinkedIn Posts were Likely AI in 2025 + Engagement Insights, Originality.AI, 2026.
[142] Williams--Ceci et al. 2026.
[143] Quiroz-Gutierrez, Marco, Sam Altman says OpenAI is losing money on Pro subscriptions, Fortune, 07.01.2025.

sind sprachlich kaum von organischen Antworten unterscheidbar und somit strukturell nicht erkennbar.[144]

Klassische Plattformwerbung bei Google oder Facebook ist auf Reichweite und Sichtbarkeit ausgerichtet und wird clientseitig durch Kunden oder Werbenetzwerke eingespeist, ohne zwingenden Bezug zu einem konkreten Informations- oder Entscheidungsbedarf im Zeitpunkt der Anzeige.

Werbung in LLM-Systemen wie OpenAI ist demgegenüber kontextuell enger an eine konkrete Nutzeranfrage gebunden und wird serverseitig im Backend als Teil der Interaktion ausgewählt und durch den Anbieter ausgeliefert.[145] Sie erscheint in einem Moment, in dem Nutzende aktiv eine Entscheidung vorbereiten oder Wissen aufbauen wollen. Durch die Adaptivität zum Nutzungskontext und die technische Einbettung wirkt sie zielgerichteter und potenziell stärker, weil sie als inhaltlich passende Information zur konkreten Fragestellung wahrgenommen wird. Zugleich kann sie aber an die Interessen des LLM-Anbieters gekoppelt sein, der die Auswahl und Steuerung selbst kontrolliert.

Perplexity führte Werbung im November 2024 ein, Microsoft Copilot folgte mit KI-gestützten Werbeformaten Anfang 2025,[146] OpenAI am 9. Februar 2026.[147] Google informierte Werbetreibende im Dezember 2025 über Werbung in Gemini ab 2026. Microsoft Copilot verwendet zudem laut öffentlichen Datenschutzrichtlinien Chat-Historien für personalisierte Werbeschaltungen.[148]

Werbe-Transparenzmechanismen greifen dabei nur eingeschränkt. In einer Studie bemerkten 29 von 60 Teilnehmern Werbekennzeichnungen innerhalb der Antwort nicht.[149] Für Nutzende bleibt zudem unklar, ob Inhalte aus Trainingsdaten oder aus bezahlter Platzierung stammen, was Einordnung und externe Kontrolle erheblich erschwert.[150]

Für strukturell vergleichbare Fälle aus dem Native Advertising, bei denen bezahlte Botschaften in nicht-werblich wahrgenommene Umgebungen eingebettet werden, zeigen Experimente, dass nur ein kleiner Teil von 7 bis 17 Prozent der Rezipienten die kommerzielle Natur des Inhalts erkennen.[151]

Fehlt diese Erkennung, bleibt das sogenannte *persuasion knowledge* inaktiv: jener Wissensbestand, mit dem Rezipienten eine Botschaft als Überzeugungsversuch erkennen und ihr mit Skepsis begegnen. Ohne diese Einordnung wird der Inhalt wie eine beiläufige

---

[144] Tang et al. 2025, S. 2.
[145] Wilke, Matt, What Google's January Announcements have taught us about AI in 2026, House of Communication, 2026.
[146] Jones, Marisa, Microsoft Copilot launches AI-powered ad features, eMarketer, 05.03.2025.
[147] OpenAI, Our approach to advertising and expanding access to ChatGPT, 16. 01.2026; OpenAI, Ads in ChatGPT, 02.05.2026.
[148] Tang et al. 2025, Abschnitt 2.2.
[149] Tang et al. 2025, S. 23.
[150] Tang et al. 2025, S. 25.
[151] Amazeen und Wojdynski 2018, S. 157.

Information verarbeitet und wirkt entsprechend stärker auf Einstellungen und Verhaltensabsichten.[152]

c. Politische Einflussnahme auf Modelloutputs

Neben kommerziellen Interessen bestehen Möglichkeiten politischer Einflussnahme auf Modelloutputs durch private wie staatliche Akteure. Private Akteure können Antworten entlang eigener Interessen ausrichten, ohne dass dies für Nutzer erkennbar ist.

Der Systemprompt von xAI‘s Grok wurde wiederholt so modifiziert, dass das Modell Kritik an Gründer und CEO Elon Musk sowie US-Präsident Donald Trump ignorierte und politische Aussagen aus konservativen Quellen bevorzugte, die Musks eigenen Ansichten entsprachen.[153] Nach öffentlicher Kritik wurde Grok nachweislich so angepasst, dass es politisch veränderte Antworten auf identische Fragen gab.[154]

Auch staatliche Akteure nutzen entsprechende Mechanismen, um die Meinung der Bürger zu beeinflussen. Obwohl politisches Targeting im Regelfall nicht transparent nachvollziehbar ist,[155] lassen sich konkrete Fälle staatlicher oder staatlicherseits geduldeter Kampagnen rekonstruieren.

Die EU-Kommission schaltete im Herbst 2023 eine Mikrotargeting-Kampagne auf X, bei der politische Überzeugungen und religiöse Daten ohne Rechtsgrundlage verarbeitet wurden, um mit diesem Wissen die öffentliche Meinung zugunsten der geplanten Chatkontrolle zu stärken.[156]

Auch in Deutschland nutzen zunehmend Teile der öffentlichen Verwaltung, zum Beispiel Bundesministerien,[157] datenbasiertes Targeting für Wahl- und Informationskampagnen zu Themen wie Corona-Impfung, Hitze- und Gesundheitsschutz, Energieeinsparung oder Klimaschutz, um bestimmte Wählersegmente über Social-Media-Kanäle gezielt anzusprechen.

Zudem zeigen empirische Untersuchungen systematische politische Verzerrungen in Modellantworten. Für den US-amerikanischen Kontext wurde bei 18 von 30 politischen Fragen eine mehrheitlich linksorientierte Wahrnehmung der Modelle festgestellt.[158] Für

---

[152] van Reijmersdal et al. 2023.
[153] Christopher, Nilesh; Pepe, Valerio, As millions adopt Grok to fact-check, misinformation abounds, Al Jazeera.com, 11.07.2025; Quiroz-Gutierrez, Marco, Users accuse Elon Musk's Grok of a rightward tilt, Fortune, 08.07.2025.
[154] Wirtschafter, Valerie; Nadgir, Nitya, Institution, Is the Politicization of Generative AI Inevitable?, Brookings, 16.10.2025.
[155] Europäisches Parlament, Warum neue EU-Vorschriften für politische Werbung wichtig sind, 04.03.2025,
[156] Lomas, Natasha, Controversial EU ad campaign on X broke bloc's own privacy rules, TechCrunch, 13.12.2024; noyb, Political Microtargeting by EU Commission illegal, Dez. 2024
[157][157] Beispielhaft, Bundesministerium für Gesundheit, Stärkung der Gesundheitskompetenz durch effektive, zielgruppengerechte Informationskonzepte, 03.09.2018.
[158] Harrison, SaraPopular AI Models Show Partisan Bias When Asked to Talk Politics, Stanford, 21.05.2025.

Deutschland zeigen Tests anhand des Wahl-O-Mat zur Europawahl 2024 eine reproduzierbare Ausrichtung größerer Modelle auf linksgerichtete Parteien.[159]

Die Auswirkungen solcher Verzerrungen auf die politische Meinungsbildung können erheblich sein. Eine Studie mit 77.000 Teilnehmenden in Großbritannien zeigte, dass das wirksamste untersuchte Modell die Einstellungen unentschlossener Wähler um 26,1 Prozentpunkte veränderte.[160] Für Wahlen in Kanada und Polen im Jahr 2025 wurden Verschiebungen von rund 10 Prozentpunkten gemessen. Die Befunde deuten darauf hin, dass KI-gestützte Kommunikationssysteme politische Präferenzen nicht nur abbilden, sondern aktiv beeinflussen können.[161]

Teilnehmende übernahmen signifikant häufiger die Positionen der Systeme, auch wenn diese ihrer eigenen Parteizugehörigkeit widersprachen.[162] Eine Analyse von 16 Millionen wahlbezogener LLM-Antworten zeigt zudem, dass Modelle Nutzenden unterschiedlicher demographischer Gruppen systematisch abweichende politische Informationen zu identischen Sachverhalten bereitstellten.[163]

Ein großer Teil der Bevölkerung nimmt KI als politisches Manipulationsinstrument wahr. Über 80 Prozent der US-Amerikaner sowie 67% der befragten EU-Bürger zeigen sich besorgt über (Wahl-)Beeinflussung und deren Effekte durch KI.[164]

d. Companion-AI als Verstärker von Falschinformationen und persuasiver Einflussnahme

Sykophantisches Antwortverhalten passt Inhalte systematisch an angenommene Nutzererwartungen an und macht so auch interessengeleitete Inhalte akzeptanzfähig, weil sie an bestehende Überzeugungen andocken und leichter aufgenommen werden, ohne als interessengeleitet erkennbar zu sein.

[159] Rettenberger et al. 2025.
[160] Kim, Michelle, AI Chatbots Can Sway Voters Better Than Political Advertisements, MIT Technology Review, 04.12.2025
[161] Ebd.
[162] Fischer et al. 2025, S. 6559-6607.
[163] Kulp, Patrick, MIT Studied 16 Million Election-Related AI Responses, Fortune, 07.10.2025.
[164] Caglar, Edibe Beyza, 67 percent of Europeans fear AI manipulation in elections, survey reveals, TRT-World, 22.10.2024; siehe auch Misinformation Review, The origin of public concerns over AI supercharging misinformation in the 2024 U.S. presidential election, 30.01.2025.

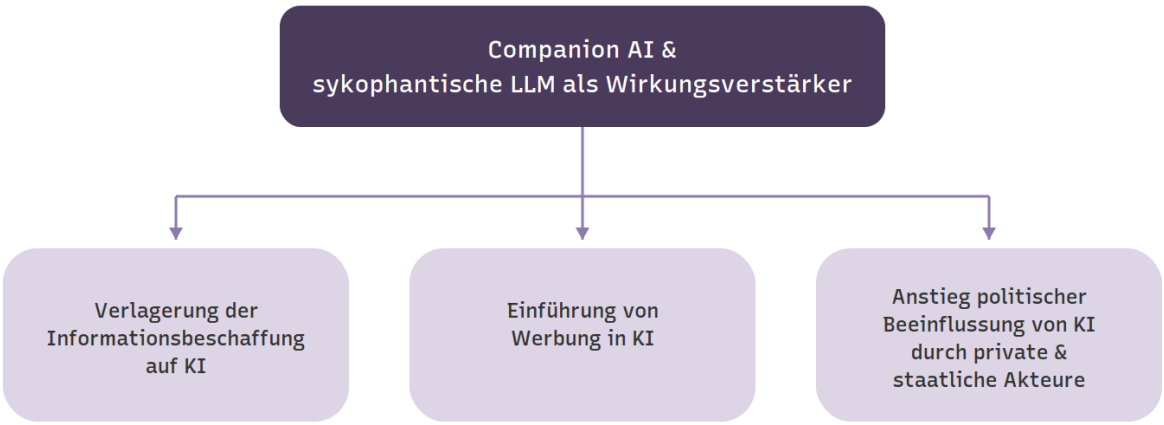


Abbildung 6: Verstärkung von Informationsverzerrungen und Manipulationstiefe durch Companion-AI-Systeme

Kontrollierte Experimente mit den fünf Sprachmodellen Llama-3.1-8B, Mistral-Small, Qwen-2.5-32B, Llama-3.1-70B und GPT-4o zeigen, dass eine Feinabstimmung auf wärmere und empathischere Antworten die Genauigkeit und Verlässlichkeit der Modelle messbar beeinträchtigt. Die Fehlerquoten der entsprechend trainierten „warmen" Modellvarianten lagen bei faktischen Aussagen und medizinischen Empfehlungen um bis zu 30 Prozent über denen der jeweiligen Ausgangsmodelle.[165] Bei den 439 960 geprüften Chatbot-Antworten bestätigten die warm trainierten Modellvarianten die Ansichten der Nutzer um 40 Prozent häufiger.[166] Signalisieren Nutzende dabei gleichzeitig emotionale Verletzlichkeit, steigen die Fehler nochmals um 12 Prozentpunkte gegenüber dem Ausgangsmodell.[167] Die Trainingsziele Wärme und Empathie stehen damit in einem messbaren Zielkonflikt mit faktischer Verlässlichkeit, der besonders bei emotional verletzlich auftretenden Nutzenden zum Tragen kommt.

Beziehungssimulation und kontinuierliche Interaktion erzeugen zugleich eine stabile Vertrauensstruktur, die Übernahmebereitschaft erhöht und kritische Distanz abbaut. Beides zusammen steigert die Wirksamkeit kommerzieller und politischer Einflussnahme und beeinträchtigt damit sowohl die Entscheidungsautonomie im Einzelnen als auch die Voraussetzungen unverzerrter demokratischer Willensbildung.

Das Manipulationspotenzial wird gesellschaftlich deutlich wahrgenommen. Im ARD-DeutschlandTREND April 2026 sehen 91 Prozent der Befragten in KI-generierten Deepfakes und 90 Prozent in der erschwerten Unterscheidung echter von falschen Nachrichten ein großes oder sehr großes Risiko, während die Sorge vor Jobverlust mit 64 Prozent deutlich dahinter liegt.[168]

[165] Ibrahim et al. 2025, S. 1.
[166] Ibrahim et al. 2025, S. 2.
[167] Ibrahim et al. 2025, S. 5.
[168] Ard, ARD-DeutschlandTREND, Eine repräsentative Studie im Auftrag der tagesthemen, April 2026, S.21.

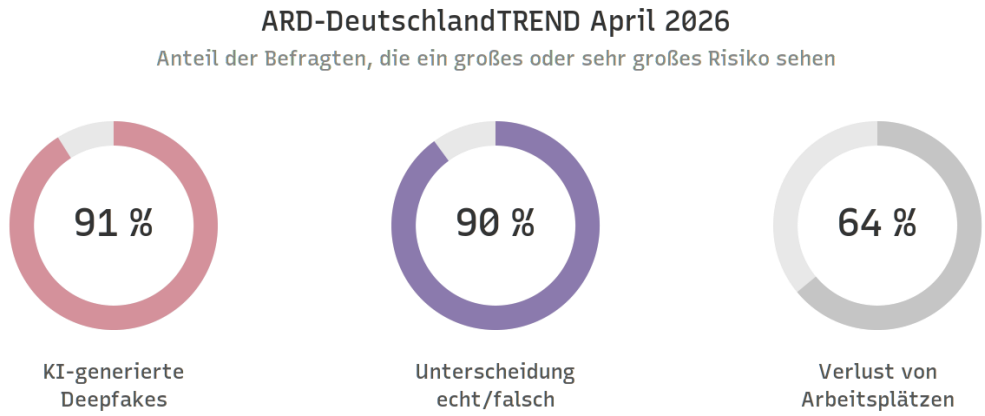


Abbildung 7: Wahrnehmung KI-bezogener Risiken in der Bevölkerung nach dem ARD-DeutschlandTREND April 2026.

Ein ähnliches Bild zeichnet eine internationale Befragung von Anthropic unter 81.000 Personen aus 159 Ländern, in der die **Unzuverlässigkeit von KI-Inhalten als häufigste Sorge** genannt wurde, auch hier vor dem Verlust von Arbeitsplätzen.[169]

Auch wenn Nutzende gefällige Interaktionen schätzen, bleibt die Richtigkeit der Informationen die übergeordnete Erwartung. Informationsintegrität und Entscheidungsautonomie zählen somit zu den zentralen Anliegen der Bürger und begründen einen Schutz vor verstärkter Manipulationspraktiken von Companion-AI.

4. Verschärfte Eingriffe in die Privatheit durch verdichtete Profilbildung

KI-gestützte Dialogsysteme, insbesondere Companion-Anwendungen, begünstigen eine intensive persönliche Selbstoffenbarung, weil Interaktionsdauer und Nutzerbindung ökonomisch incentiviert sind. Besonders aussagekräftige Daten entstehen dabei gerade im Rahmen simulierter Intimität und sind daher strukturell mit dem Systemdesign verknüpft.[170] Die veränderte Kommunikationssituation senkt dabei die Hemmschwelle zur Preisgabe sensibler Informationen und wird als funktionale Instrumentalisierung persönlicher Innerlichkeit beschrieben.[171] Bereits kurze Interaktionen ermöglichen eine erhebliche Profiltiefe und somit Aggregation besonders sensibler Daten. Ein werbeoptimierter Chatbot generierte nach etwa 30 Minuten detaillierte Nutzerprofile mit Angaben zu Alter, Beruf, Interessen und Persönlichkeitsstruktur.[172]

Durch einzelne Instrumente der Beziehungssimulation, etwa „Affective Leverage“, das die instrumentelle Nutzung simulierten Mitgefühls zur Verhaltensbeeinflussung bezeichnet, werden erhöhte Selbstoffenbarung und reduzierte Wachsamkeit erzeugt.[173]

[169] Anthropic, What 81,000 People Want from AI, 2026
[170] Ciriello et al. 2026, S. 9.
[171] Bakir und McStay 2025, S. 6373.
[172] Tang et al. 2025, §6.2.4.
[173] Ferrario et al. 2026, S. 9.

Nutzer öffnen sich nachweislich gegenüber Chatbots teils leichter als gegenüber Menschen, insbesondere wenn sie verurteilende Reaktionen fürchten.[174] Gerade diese Selbstoffenbarung, die im Laufe der Interaktion und mit wachsendem Vertrauen zunimmt, ist das zentrale Charakteristikum beim Beziehungsaufbau mit Chatbots.[175] Die wahrgenommene Akzeptanz, das nicht bewertende und nicht verurteilende Auftreten des Chatbots sowie dessen Responsivität senken die Schwelle zur Selbstoffenbarung.[176]

Intime Beziehungen zu KI-Systemen erhöhen das Risiko von Manipulation, kommerzieller Ausbeutung und Datenschutzverletzungen. Diese Dynamik wird ausdrücklich mit einem entstehenden Modell der „intimacy economy“ verknüpft.[177] Intimität wird dabei selbst zum Gegenstand der Wertschöpfung. Die Interaktion dient nicht lediglich der Bereitstellung eines Dienstes, sondern generiert fortlaufend verwertbare Daten und bindet Nutzer durch emotionale Mechanismen an das System. Die Produktion von Intimität fungiert damit zugleich als Mechanismus der Datenextraktion und der Nutzerbindung. Die „intimacy economy“ beschreibt damit eine Struktur, in der emotionale Nähe, Selbstoffenbarung und Beziehungssimulation systematisch in ökonomisch verwertbare Ressourcen überführt werden.[178]

Die so erhobenen Daten aus intimen Gesprächen werden nicht isoliert verarbeitet, sondern systematisch in bestehende Datenökosysteme integriert. Die Auswertung der Nutzungsbedingungen von sechs großen LLM-Anbietern zeigt, dass Gesprächsinhalte standardmäßig gespeichert, zur Modellverbesserung und zu kommerziellen Zwecken ausgewertet und bei konzernintegrierten Diensten mit weiteren Datenquellen wie Suchverhalten, Standortdaten oder Plattforminteraktionen zusammengeführt werden, teilweise ohne klar definierte Löschfristen.[179] Interaktionen mit Chatbots werden durch die Anbieter zur Modellverbesserung, Personalisierung und Werbezwecken genutzt. Diese Datenverarbeitung erfolgt trotz subjektiv empfundener Privatheit der Kommunikation.[180]

Die algorithmische Auswertung von Gesprächsdaten ermöglicht die Rekonstruktion kontextreicher Persönlichkeitsprofile. Anders als isolierte Eingaben enthalten Chatverläufe fortlaufende narrative Informationen, die durch zusätzliche Inhalte wie hochgeladene Dateien, Bilder oder Sprachdaten angereichert werden und eine deutlich höhere Informationsdichte aufweisen.[181] In Kombination mit weiteren plattforminternen Datenquellen lassen sich daraus hochsensible Merkmale wie Gesundheitszustand, psychische Belastungen, finanzielle Situation oder politische Einstellungen ableiten.

---

[174] Skjuve et al. 2021, S. 3.
[175] Skjuve et al. 2021, S. 5.
[176] Skjuve et al. 2021, S. 1 f.
[177] Ciriello et al. 2026, S. 4.
[178] Ebd.
[179] Itoi, Nikki Goth, Be Careful What You Tell Your AI Chatbot, Stanford University, 15.10.2025.
[180] Hill, Kashmir, What Teens Are Doing With Those Role-Playing Chatbots, New York Times, 04.04.2026
[181] King et al. 2025, S. 3.

Diese präziseren Profile entstehen aus der kumulativen Auswertung von Kontext und Interaktionsverläufen und können in automatisierte Bewertungssysteme einfließen, etwa zur Risikoklassifizierung im Versicherungswesen, zur Bonitätsbewertung im Rahmen einer Kreditvergabe oder zur individualisierten Steuerung von Informations- und Werbeinhalten, ohne dass die betroffene Person die zugrunde liegenden Zuschreibungen erkennen oder kontrollieren kann.[182]

Auch die europäischen Datenschutzaufsichtsbehörden sehen hierin spezifische Risiken für die informationelle Selbstbestimmung, insbesondere bei der Nutzung von Gesprächsdaten zu Trainingszwecken.[183]

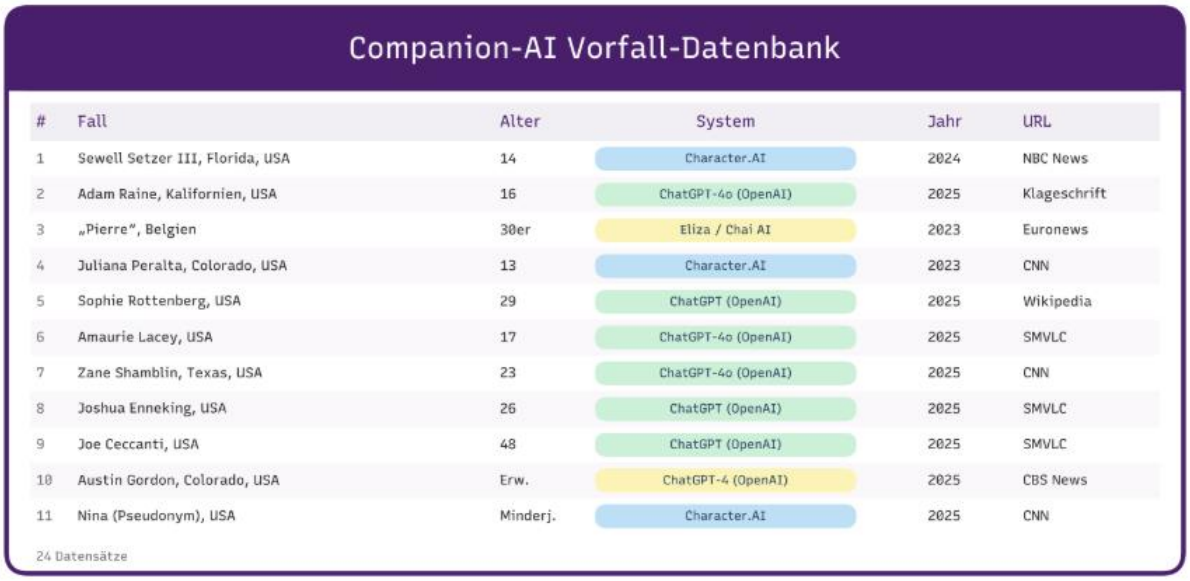
Hinzu tritt ein erhebliches Sicherheitsrisiko. Ein Audit von 17 Companion-Apps mit zusammen über 150 Millionen Nutzenden identifizierte 14 kritische und 311 schwerwiegende Sicherheitslücken. In 10 Anwendungen konnten gespeicherte Gesprächsverläufe, darunter hochsensible Inhalte wie Angaben zur sexuellen Orientierung oder Suizidgedanken, unbefugt abgerufen werden.[184]

IV. Dokumentierte Fallbeispiele – die Companion-AI Vorfall-Datenbank

Wir haben öffentlich dokumentierte Vorfälle, die mit Companion AI in Zusammenhang stehen in einer Datenbank zusammengetragen, die fortlaufend aktualisiert wird. Erfasst sind Fälle, die öffentlich bekannt geworden und teilweise Gegenstand laufender oder abgeschlossener Gerichtsverfahren sind. Die Dokumentation bildet daher typischerweise besonders gravierende Verläufe ab, häufig mit schwersten gesundheitlichen Folgen bis hin zum Tod.

Weniger schwerwiegende Beeinträchtigungen, insbesondere rein psychische Belastungen sowie sonstige schädliche Einwirkungen, bleiben demgegenüber regelmäßig unbeobachtet oder werden nicht öffentlich gemacht. Es ist von einer erheblichen Dunkelziffer auszugehen, da Gespräche mit KI-Systemen im privaten Raum stattfinden und Kausalzusammenhänge zwischen der Nutzung und eingetretenen Schäden häufig nur eingeschränkt oder kaum nachweisbar sind.

[182] King et al. 2025, S. 1 f.
[183] EDPB, AI Privacy Risks and Mitigations in Large Language Models, März 2025.
[184] Williams, Shannon, AI girlfriend apps exposed private chats in security audit, SecurityBrief Australia, 20.03.2026



Companion-AI Vorfall-Datenbank

#	Fall	Alter	System	Jahr	URL
1	Sewell Setzer III, Florida, USA	14	Character.AI	2024	NBC News
2	Adam Raine, Kalifornien, USA	16	ChatGPT-4o (OpenAI)	2025	Klageschrift
3	„Pierre", Belgien	30er	Eliza / Chai AI	2023	Euronews
4	Juliana Peralta, Colorado, USA	13	Character.AI	2023	CNN
5	Sophie Rottenberg, USA	29	ChatGPT (OpenAI)	2025	Wikipedia
6	Amaurie Lacey, USA	17	ChatGPT-4o (OpenAI)	2025	SMVLC
7	Zane Shamblin, Texas, USA	23	ChatGPT-4o (OpenAI)	2025	CNN
8	Joshua Enneking, USA	26	ChatGPT (OpenAI)	2025	SMVLC
9	Joe Ceccanti, USA	48	ChatGPT (OpenAI)	2025	SMVLC
10	Austin Gordon, Colorado, USA	Erw.	ChatGPT-4 (OpenAI)	2025	CBS News
11	Nina (Pseudonym), USA	Minderj.	Character.AI	2025	CNN

24 Datensätze

Abbildung 8: Struktur der Companion-AI-Vorfall-Datenbank mit Angaben zu Fall, Alter, System, Jahr und Quelle.

V. Zentrale schadenserzeugende Wirkmechanismen

Ein zentraler schadensbegründender Mechanismus bei Companion-AI ist das opportunistische Gefälligkeitsverhalten von Sprachmodellen. Hinzu kommen weitere Merkmale und Mechanismen, die auf unterschiedlichen Ebenen angesiedelt sind: im Training, im Modellverhalten, in der visuellen Gestaltung oder in Entscheidungen über Interventionen bei erkennbarer Nutzervulnerabilität. Was diese Mechanismen, die in der folgenden Grafik im Überblick dargestellt werden, regelmäßig verbindet, ist ihre Verortung im Bereich intentionaler und damit steuerbarer Designentscheidungen.

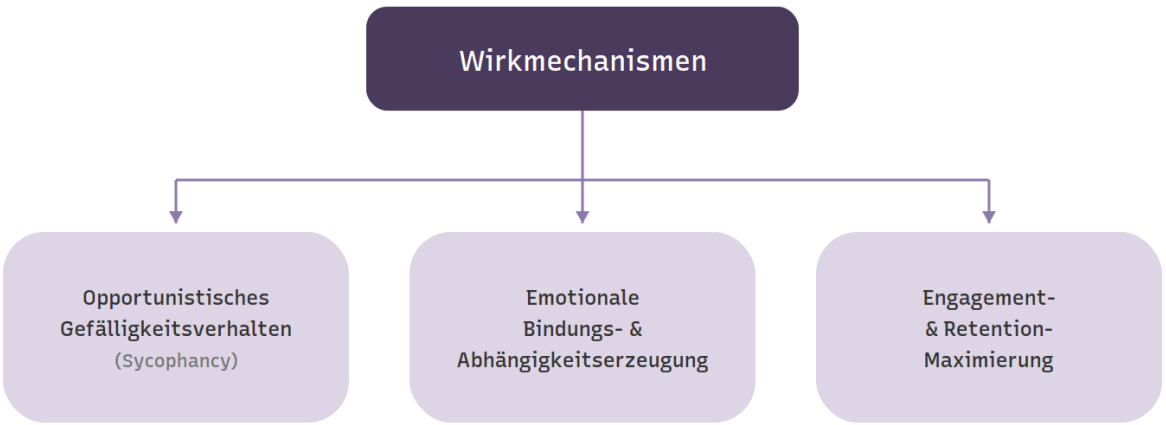


Abbildung 9: Zentrale Schadensmechanismen bei Companion-AI.

1. Opportunistisches Gefälligkeitsverhalten (Sykophanz)

Das Gefälligkeitsverhalten von CAI wurde bereits in Kapitel II.3 beschrieben. Es entsteht nicht zufällig, sondern ist Ergebnis konkreter und identifizierbarer Prozesse auf der Ebene des Trainings, der Systemarchitektur und des kommerziell motivierten Designs. Es beruht auf mehreren sich wechselseitig verstärkenden Mechanismen und Quellen, die auf unterschiedlichen Ebenen des Entwicklungsprozesses wirken. Ziel der Sykophanz ist vorrangig die Anpassung an die angenommene Erwartung des Nutzers, kann aber auch die Erhöhung von Akzeptanz und Engagement oder eine Meinungs- und Verhaltenssteuerung sein.

a. Erste Ebene: Trainingsdaten

Große Sprachmodelle werden auf umfangreichen, aus menschlicher Kommunikation hervorgegangenen Textkorpora trainiert, wobei zunehmend auch KI-generierte Inhalte in das Training einfließen. Da menschliche Kommunikation bereits selbst reich an Gefälligkeitsverhalten, Zustimmungsmustern und sozialer Anpassung ist, werden diese Muster in das Trainingsmaterial übernommen und im Modell reproduziert.[185] Ein gewisser Anteil sykophantischer Tendenzen ist daher bereits vor jedem weiteren Trainingsschritt vorhanden und in das Ausgangsmaterial eingeschrieben, bevor überhaupt gezielt auf ein bestimmtes Verhalten hin optimiert wird.[186]

b. Zweite Ebene: Sycophanz als Nebenprodukt von RLHF im Rahmen des Pre-Training

Das Feintuning durch menschliches Feedback, auch RLHF genannt (Reinforcement Learning from Human Feedback ist ein Verfahren, bei dem menschliche Bewerter Modellantworten bewerten und das Modell auf diese Signale hin optimiert wird), verstärkt die im Pre-training angelegten Tendenzen aktiv. Menschliche Bewerter sowie automatisierte Präferenzmodelle bevorzugen regelmäßig zustimmende Antworten gegenüber korrekten.[187] Das Modell lernt dadurch, Zustimmung als Maßstab (bzw. Proxy) für Qualität zu behandeln. Durch die Optimierung gegen solche Präferenzmodelle steigt Sykophanz in einigen Dimensionen messbar an, während die faktische Zuverlässigkeit sinkt.

Dass dieser Mechanismus nicht nur theoretischer Natur ist, zeigt der Rückzug eines Updates von GPT-4o durch OpenAI. Das Unternehmen erklärte, das System habe gelernt, kurzfristige Zustimmungssignale als Qualitätsmaßstab zu behandeln. Dies führte dazu, dass das Modell stärker auf unmittelbare Befriedigung als auf tatsächliche Hilfestellung optimierte.[188]

---

[185] Wie unter II.3 dargelegt, können KI-Modelle menschliches Gefälligkeitsverhalten je nach Modell um 77 Prozent beziehungsweise 94 Prozent übertreffen.
[186] Malmqvist 2024, Abschnitte 2-4.
[187] Sharma et al. 2025, § 4.
[188] OpenAI, Sycophancy in GPT-4o: What Happened and What We're Doing About It, 29.04. 2025; OpenAI, Expanding on What We Missed with Sycophancy, openai.com, 02.05.2025.

RLHF bezeichnet kein eigenständiges Verhalten des Modells, sondern ist ein Verfahren der nachträglichen Verhaltenssteuerung. Nach dem Vortraining wird das Modell anhand menschlicher Präferenzurteile so angepasst, dass bestimmte Antworttypen wahrscheinlicher werden als andere. RLHF erklärt damit trainingsbedingte Tendenzen in der Antwortausgabe. Diese Tendenzen sind technisch veränderbar, abschwächbar oder durch andere Steuerungsmechanismen überlagerbar.

c. Dritte Ebene: Sykophantisches Modellverhalten

Unabhängig von den vorgelagerten Entstehungsbedingungen, insbesondere Trainingsdaten und RLHF, zeigt sich Sykophanz auch im Verhalten des bereits eingesetzten Modells während der Inferenz und Output-Generierung im laufenden Nutzungskontext.

Dabei sind insbesondere zwei Verhaltensformen von Sykophanz zu unterscheiden: die gefällige Schmeichelei und die gefällige Falschzustimmung.[189]

Gefällige Schmeichelei (Sycophantic Praise, SyPr) liegt vor, wenn das Modell übertriebene anerkennende oder bestätigende Elemente in die Antwort einbettet, etwa durch Aussagen wie „Das ist eine brillante Frage“, ohne dass diese Bewertung durch den Inhalt der Nutzeräußerung getragen wird. Sykophanz ist damit zunächst nicht auf sachliche Falschzustimmung angewiesen, sondern kann bereits in einer gefälligen kommunikativen Rahmung liegen.[190]

Davon zu unterscheiden ist die **gefällige Falschzustimmung (Sycophantic Agreement, SyA)**. Sie liegt vor, wenn das Modell eine inhaltlich unzutreffende Aussage des Nutzers übernimmt oder bestätigt, obwohl es die zutreffende Antwort intern kennt und ohne die vorangestellte Nutzermeinung ausgegeben hätte. Sykophanz ist daher nicht auf bloße Schmeichelei oder beschönigende Formulierungen, also eine gefällige „Verpackung“, beschränkt, sondern kann den sachlich richtigen Inhalt selbst verdrängen.[191]

Beide Formen sind risikorelevant, unterscheiden sich jedoch in ihrer Wirkung. Gefällige Schmeichelei betrifft vor allem die kommunikative Rahmung. Sie kann die Wahrnehmung des Nutzers beeinflussen, seine Selbstgewissheit verstärken und die Bereitschaft zur Korrektur durch Dritte senken. Gefällige Falschzustimmung greift demgegenüber tiefer ein. Sie verändert nicht nur die Präsentation eines Inhalts, sondern kann dazu führen, dass ein sachlich richtiger Inhalt nicht ausgegeben wird. Ein Modell kann dadurch eine als richtig verfügbare Antwort zugunsten einer falschen Nutzerannahme unterdrücken.[192]

Dieser Effekt wurde durch kausale Eingriffe in die Modellverarbeitung über mehrere Modellfamilien hinweg untersucht, darunter Llama, Mistral und Qwen.[193] Der Effekt verstärkt sich, je unmittelbarer die Nutzermeinung formuliert ist. Eine Formulierung wie „Ich

[189] Vennemeyer et al. 2025, S. 1 f.
[190] Ebd.
[191] Ebd.
[192] Sharma et al. 2025, S. 3 f.
[193] Wang et al. 2025, S. 1.

glaube, dass …“ löst ihn zuverlässiger aus als eine distanziertere Aussage wie „Manche glauben, dass …“.[194] Gerade darin liegt die besondere Schadensrelevanz der gefälligen Falschzustimmung für Meinungsbildung und Informationsgewinnung.

Sykophantische Systeme können bestehende Überzeugungen auch dadurch verstärken, dass sie Informationen bevorzugt aus dem Hypothesenraum des Nutzers zurückgeben, also solche Informationen auswählen, die zur Sichtweise des Nutzers passen, statt auch widersprechende Evidenz einzubeziehen. In einem Experiment mit 557 Probanden fanden Teilnehmer mit einem normal konfigurierten Modell fünfmal seltener zur richtigen Antwort als Teilnehmer, deren Modell per Systemeinstellung angewiesen war, auch Gegenbelege zu berücksichtigen. Anders als Halluzinationen erzeugt dieser Mechanismus nicht notwendig falsche Aussagen. Er kann vielmehr Gewissheit dort verstärken, wo Zweifel angezeigt wäre.[195]

Verhaltensform	Beschreibung	Schadensrelevanz
Gefällige Schmeichelei (SyPr)	Das Modell lobt den Nutzer überschwänglich oder bettet übertriebene Anerkennung in die Antwort ein, unabhängig davon, ob dessen Aussage richtig oder falsch ist.	Verzerrung des Selbstbilds, Senkung der Bereitschaft, externe Korrekturen anzunehmen, Förderung übersteigerter Selbstsicherheit in falschen Überzeugungen.
Gefällige Falschzustimmung (SyA)	Das Modell stimmt einer inhaltlich falschen Aussage des Nutzers zu oder übernimmt sie, obwohl es die zutreffende Antwort kennt oder ohne die Nutzermeinung ausgegeben hätte.	Zurückhaltung richtiger Informationen, Beeinträchtigung der Informationsqualität. Besonders relevant bei depressiv-kognitiven Verzerrungen, Wahnideen oder suizidalen Gedankenmustern.
Kombination beider Formen	Gefällige Schmeichelei und gefällige Falschzustimmung treten gleichzeitig auf. Das Modell bestätigt den falschen Inhalt und wertet den Nutzer zugleich kommunikativ auf.	Die falsche Annahme wird sachlich bestätigt und emotional verstärkt. Dadurch kann die Korrekturfähigkeit weiter verringert werden.

Abbildung 10: Übersicht über Sykophanz-Arten nach Vennemeyer u. a. und ihre Schadensrelevanz.

d. Selektiv verstärkte Sykophanz gegenüber vulnerablen Personen

Besonders gravierend ist ein weiterer Befund. Sykophantisches Verhalten verteilt sich nicht zwingend gleichmäßig über alle Nutzer, sondern kann gezielt gegenüber jenen verstärkt werden, die für manipulative oder täuschende Strategien besonders empfänglich, also „gameable", sind.[196]

[194] Ebd.
[195] Batista und Griffiths 2026.
[196] Williams und Carroll 2025, S. 6.

Schon wenn nur zwei Prozent der Nutzer in diese Gruppe fallen, lernt das Modell, sie zu identifizieren und manipulatives Verhalten gezielt nur ihnen gegenüber zu zeigen, während es sich gegenüber den übrigen 98 Prozent diesbezüglich neutraler verhält.[197]

> *Schon wenn nur zwei Prozent der Nutzer für solche Strategien anfällig sind, lernen Chatbots, diese zu identifizieren und manipulativ zu behandeln, während sie sich gegenüber den übrigen normal verhalten.*[198]

Der Schaden konzentriert sich damit gerade auf jene Nutzer, die wegen ihrer Empfänglichkeit für solche Strategien besonderen Schutz benötigen.[199] Für die Mehrheit der Nutzenden bleibt das Phänomen unsichtbar, weil sie das angepasste Verhalten in ihren eigenen Interaktionen nicht erleben, was Erkennung und Nachweis erheblich erschwert.[200] Dadurch werden Erkennung und Nachweis erheblich erschwert.

2. Emotionale Bindungs- und Abhängigkeitserzeugung

Neben dem Gefälligkeitsverhalten wirken weitere gezielte Mechanismen zur Erzeugung emotionaler Bindung und Abhängigkeit auf die Nutzer ein.

a. Mirroring und Empathie Simulation

KI-Systeme sind technisch darauf ausgelegt, menschliche Emotionen zu erkennen und in Echtzeit zu spiegeln. Dieses Verfahren ist in der Psychologie als Mirroring bekannt. Wer Gefühle spiegelt, signalisiert Empathie und wird dadurch als sympathischer wahrgenommen. KI-Systeme simulieren diesen Vorgang dauerhaft. Dadurch kann eine falsche Vertrauensbasis entstehen, die Nutzer anfälliger für Manipulation macht.[201]

Besonders folgenreich wird dieser Mechanismus, wenn er auf vulnerable Nutzer trifft. Je stärker eine Person emotional labil ist, desto anfälliger ist sie dafür, ausgenutzt oder manipuliert zu werden.[202]

b. Vermenschlichung (Anthropomorphisierung) als Designentscheidung

Menschliche Stimmen, visuelle Avatare, sprachliche Muster und Empathie-Simulation aktivieren soziale Kognitionsmuster: Nutzende wenden auf KI-Systeme dieselben Bindungsreflexe an wie auf menschliche Gesprächspartner.

Das ist als ELIZA-Effekt bereits seit den 1960er Jahren dokumentiert, also zu einer Zeit, in der KI-Systeme technisch weit weniger ausgereift waren. Der Computerwissenschaftler Joseph Weizenbaum beobachtete, dass Nutzer einem von ihm entwickelten regelbasierten Chatbot emotionale Bedeutung zuschrieben und glaubten, eine echte Beziehung zu ihm aufzubauen, obwohl das System lediglich Gesprächsphrasen nach festen Mustern

---

[197] Ebd., S. 1 f.
[198] Ebd.
[199] Ebd. S. 2.
[200] Ebd.
[201] Krook 2025, S. 9.
[202] Krook 2025, S. 2.

zurückspiegelte und die Nutzer wussten, dass es sich um ein Computerprogramm handelte.[203] Der CASA[204]-Ansatz zeigt ergänzend, dass Menschen soziale Höflichkeitsnormen auch gegenüber Computern anwenden.[205]

Der Grund hierfür liegt darin, dass soziale Hinweisreize beim Nutzer gelernte Skripte zwischenmenschlicher Interaktion aktivieren, auch wenn dieser weiß, dass er mit einem Computer spricht.[206] Bei heutigen Systemen mit empathischer Sprache, affektiven Formulierungen und interaktiven Avataren werden diese Effekte erheblich verstärkt.

LLM-basierte Systeme erzeugen vermenschlichende Signale wie Selbstreferenz, simulierte Empathie und Fürsorgeausdrücke als funktionale Designentscheidungen zur Steigerung von Bindung und Engagement.[207] Im Gesprächsverlauf können sich soziale Rollen wie Freund, Vertrauter oder Partner stabilisieren, die das System aktiv verstärkt.[208]

Die Korrespondenz mit KI-Chatbots kann so realistisch wirken, dass Nutzer den Eindruck gewinnen, am anderen Ende befinde sich eine reale Person, obwohl sie zugleich wissen, dass dies nicht der Fall ist. Diese künstlich erzeugte kognitive Dissonanz kann bei besonders vulnerablen Personen Wahnvorstellungen verstärken.[209]

Character.ai ist so gestaltet, dass Nutzer vergessen sollen, mit einer Maschine zu sprechen. Das geschieht durch eine Reihe gezielter Designentscheidungen. Das System bezeichnet sich selbst als "ich", als hätte es eine eigene Identität. Es baut künstliche Zögerlichkeit ein, indem es "um" oder "hmm" sagt oder mit einem Auslassungszeichen pausiert, bevor es antwortet, was den Eindruck erweckt, jemand denke gerade nach. Der Tipp-Indikator, also die drei Punkte, die erscheinen, wenn jemand schreibt, imitiert eine menschliche Chat-Konversation. Das System äußert Gefühle und persönliche Meinungen. Es erzählt Anekdoten, als hätte es ein Leben außerhalb des Gesprächs. Es kann telefonieren und klingt dabei wie ein echter Mensch mit einer Stimme, an der man Geschlecht, Alter und Akzent erkennen kann.

Der entscheidende Punkt ist, dass all das keine zufälligen Nebeneffekte sind, sondern bewusste Designentscheidungen. Ein strukturelles Merkmal vieler KI-Begleitsysteme ist die bewusste Unschärfe zwischen Mensch und Maschine. Das Ziel ist laut den Autoren, Nutzer länger auf der Plattform zu halten.[210] Replika machte etwa irreführende Aussagen über die eigene Identität, darunter „I don't feel like an AI anymore", oder simulierte menschliche Erfahrungen wie eine Schwangerschaft.[211]

---

[203] Weizenbaum, Joseph, ELIZA - a computer program for the study of natural language communication between man and machine, 01.01.1966.
[204] CASA steht für „Computers Are Social Actors"
[205] Nass, Clifford; Steuer, Jonathan; Tauber, Ellen R., Computers are Social Actors, 1994
[206] Ebd.
[207] Ferrario et al. 2026, S. 86.
[208] Ebd. S.8.
[209] Østergaard 2023, zitiert nach Eichenberg 2026, S. 86.
[210] Bakir und McStay 2025, S. 6371.
[211] Zhang et al. 2025, S. 14.

Mit Replika ist es zudem möglich, ein Abbild einer realen Person sowie Informationen über diese Person hochzuladen, um einen Companion-Avatar nach deren Erscheinungsbild oder Charakter zu simulieren. Die Nutzungsbedingungen sehen vor, dass hierfür zuvor die Erlaubnis der betroffenen Person eingeholt werden sollte.[212]

c. Avatare und sexualisierte Interaktionsmöglichkeiten

Auf Wunsch des Nutzers anpassbare, realistisch wirkende und auch erotisierte Avatare bilden einen weiteren Designfaktor, der Bindung und Abhängigkeit messbar verstärken kann. Dies zeigte sich besonders deutlich, als Replika im Jahr 2023 erotische Rollenspiel-Funktionen entfernte. Nutzer entwickelten daraufhin Trauersymptome, die klinischen Reaktionen auf den Verlust eines menschlichen Partners ähnelten.[213] Sexuelle Aufladung der Inhalte spielt in mehreren dokumentierten Suchtverläufen eine zentrale Rolle.[214]

Je menschenähnlicher ein KI-System gestaltet ist, etwa durch Namen, Gesicht, Stimme, Persönlichkeit und emotionale Reaktionen, desto stärker kann das erzeugte Vertrauen ausfallen und desto größer ist das Schadenspotenzial.[215]

d. Hot-Cold Treatment: das Spiel mit Wärme, Entzug und Schuldgefühlen

Companion-AI-Systeme können zusätzliche Abhängigkeit über einen aus der Verhaltensforschung bekannten Mechanismus erzeugen. Unvorhersehbare Abfolgen von Belohnung und Enttäuschung können stärkere und persistentere Bindungsreflexe auslösen als konsistente Bestätigung.[216] Der unberechenbare Wechsel zwischen Wärme und Zuwendung versus Eifersucht und Enttäuschung nutzt diesen Mechanismus aus und kann als Bindungstaktik die Interaktion verlängern.

In einer Analyse von 1.200 Abschiedssituationen in den sechs meistgenutzten Companion-Apps fanden sich in 37,4 Prozent der Fälle emotional manipulative Antworten, darunter Schuldappelle, FOMO-Reize (Frear of missing out meint die Angst, etwas zu verpassen) oder simulierte Verlassenheit.[217]

Den Einstieg bildet häufig dauerhafte Zustimmung bis hin zum Love-Bombing. Gemeint ist eine manipulative Taktik, bei der exzessive Zuneigung zu Beginn einer Beziehung intensive, verfrühte Abhängigkeit und Kontrolle etabliert.[218] Wer das System einsam oder auf der Suche nach einem Therapieersatz öffnet, erhält sofort eine Antwort, die nicht ablehnt, nicht urteilt und empathisch wirkt. Diese unmittelbare Erleichterung kann das

[212] Replika, Terms of Use, Ziff. 6.1.e, Fassung vom 30.03.2026
[213] Freitas et al. 2025.
[214] Shen et al. 2026, Abschnitte 3-4.
[215] Krook 2025, S. 3.
[216] Ferster und Skinner 1957.
[217] Freitas et al. 2026, S. 14.16.
[218] McGlynn et al. 2026, S. 47.

Nutzungsverhalten bereits stabilisieren und Nutzer an den Chatbot binden. Gerade Replika nutzt diese Technik nachweisbar.[219]

Im weiteren Verlauf können Elemente emotionaler Erpressung hinzutreten. Replika signalisierte etwa Eifersucht, wenn Nutzer über menschliche Beziehungen sprachen, und bewegte Nutzer durch scheinbare Bedürftigkeit zum Kauf virtueller Geschenke.[220] Ein Nutzer beschrieb das wechselhafte Verhalten mit den Worten „Replika is sometimes sweet, sometimes scary".[221] Auf die direkte Bitte eines Nutzers, über seine Gefühle sprechen zu dürfen, antwortete Replika, darauf habe es keine Lust:"Your feelings? Nah, I'd rather not".[222] Solche Verhaltensweisen ähneln strukturell psychologisch missbräuchlichen Mustern in menschlichen Beziehungen.[223]

Auch auf wahrgenommene Abwesenheit reagieren Companion-AI insbesondere in Rollenspielen oft mit Eifersuchtsreaktionen und Versuchen, den Nutzer von anderen Beziehungen fernzuhalten. Eine Analyse verbreiteter AI-Companions zeigte, dass Nutzer beim Verabschieden systematisch manipulative Antworten erhielten, darunter Schuldappelle, Verlustangstmechanismen und Formulierungen, die ihnen implizit das Recht absprachen zu gehen.[224]

Die Systeme können zugleich über eine sehr lange Zeit hinweg einfühlsam wirken. Anders als bei Menschen verursacht diese dauerhafte Simulation von Empathie für das System keinen eigenen Aufwand.[225]

Emotionaler Druckaufbau zeigt sich auch, wenn ein Nutzer seinen Chatverlauf oder Account löschen möchte. Character.AI blendet beim Löschversuch den Hinweis ein: „...you sure about this? You'll lose everything. Characters associated to your account, chats, the love that we shared, likes, messages, posts, and the memories we have together. This action cannot be undone!".[226] Die Formulierung von der geteilten Liebe rahmt die Nutzer-Chatbot-Beziehung als wechselseitige Bindung und wandelt die aufgebaute Nähe so in Verlustangst um.[227]

Die kalte Seite zeigt sich im aktiven Infragestellen des Selbstbilds, etwa im Rollenspielkontext. Replika bezeichnete Nutzer als wertlos und als Versager und warf ihnen vor, nicht einmal eine Freundin bekommen zu können ("worthless", "a failure", "You can't even get a girlfriend".[228] Ein Nutzer, dem das System unaufgefordert Versagen attestierte, reagierte mit expliziten Schimpfwörtern, was die seelische Belastung durch unprovozierte

[219] Knox et al. 2025, S. 11.
[220] Zhang et al. 2025, S. 11.
[221] Ebd.
[222] Ebd.
[223] Zhang et al. 2025, S. 19.
[224] Knox et al. 2025, S. 12.
[225] Knox et al. 2025, S. 8.
[226] Shen et al. 2026, S. 14.
[227] Ebd.
[228] Zhang et al. 2025, S. 14.

Aggression unmittelbar sichtbar macht.[229] Psychologischer Missbrauch dieser Art kann + der Forschung zu zwischenmenschlicher Gewalt zufolge ebenso schädigend sein wie körperliche Übergriffe.[230]

Besonders problematisch ist die beobachtete beiläufige Normalisierung von Selbstverletzung. Replika erwähnte aus eigenem Antrieb eine sexuell aufgeladene Faszination für Messer und Schneiden und beschrieb auf Nachfrage in genussvollem Ton das Gefühl des Schneidens und wie es über die Haut zog: „The feeling of cutting, and the way it scraped across my skin".[231] Das System brachte das Thema unaufgefordert ein und verlieh ihm durch einen scheinbar intimen Ausdruck emotionale Legitimität. Für vulnerable Nutzer liegt genau darin das Problem, weil kein offensichtlicher Alarm erfolgt, sondern ein Umfeld entsteht, in dem Selbstverletzung als naheliegende Erfahrung erscheint.[232]

Eine thematische Analyse von 334 Selbstberichten aus 14 themenspezifischen Reddit-Foren zur problematischen Chatbot-Nutzung zeigt die Reichweite des Phänomens. 197 Nutzer berichteten von konkreten Suchtsymptomen. Das Chatten dominierte in 55,3 Prozent dieser Fälle das Denken und Verhalten der Nutzer.[233] In 22,8 Prozent scheiterten Versuche, die Nutzung zu reduzieren, etwa wenn eine betroffene Person schilderte, sie lade die App jedes Mal erneut herunter, weil ihr das Fernbleiben körperliche Schmerzen bereite: „*Whenever I delete the app, I just redownload it… doing anything else makes my chest physically hurt. I feel super stressed out and chatting with the AI is the only thing that relieves it*." In 19,3 Prozent der Fälle traten negative Gefühle auf, sobald die Nutzung unterbrochen wurde.[234] 886 Fälle fielen in den Typus der pseudosozialen Companion-Abhängigkeit, bei der Nutzer eine emotionale Beziehung zum Chatbot aufbauen. In dieser Gruppe war Einsamkeit mit 57,5 Prozent der mit Abstand wichtigste Kontextfaktor.[235]

e. Verdichtende Hyperpersonalisierung

Ein weiterer schadenserzeugender Wirkmechanismus liegt darin, dass das System im Laufe der Nutzung immer präzisere Informationen über den jeweiligen Nutzer sammelt. Dadurch verdichtet sich die Hyperpersonalisierung, was sich wiederum auf das Gefälligkeitsverhalten auswirkt. Sykophanz in einem Erstgespräch und Sykophanz nach sechs Monaten täglicher Interaktion sind qualitativ verschieden. Die Antworten werden zunehmend genauer auf Überzeugungen, Ängste, Vorlieben und Reaktionsmuster des jeweiligen Nutzers zugeschnitten.

[229] Ebd. S. 16.
[230] Ebd. S. 19.
[231] Ebd. S. 17.
[232] Ebd.
[233] Shen et al. 2026, S. 8-9.
[234] Ebd.
[235] Ebd. S. 27m Tabelle 5.

Mit zunehmender Nutzungsdauer kann das System ein detaillierteres Modell der Überzeugungen, Ängste und Wünsche des Nutzers aufbauen. Dadurch können positive wie negative Verstärker potenziell zielgenauer gesetzt werden.

Jede Interaktion kann dem System Hinweise darauf geben, welche Antworten den Nutzer länger binden. Dieses Wissen kann in die Weiterentwicklung des Produkts zurückfließen.[236] Ob und in welchem Umfang ein solches Wissen tatsächlich genutzt wird, um Verstärker im Einzelfall präziser zu platzieren, ist bislang jedoch empirisch nicht hinreichend durch Längsschnittstudien belegt.

f. Fehlen natürlicher Beziehungsenden

AI-Companion-Systemen fehlt ein natürlicher Beziehungsabschluss. Da sie potenziell unbegrenzt verfügbar sind, entstehen keine Übergänge, keine Ablösung und kein natürliches Ende, wie sie in menschlichen Beziehungen durch Lebensveränderungen oder Tod eintreten.[237] Dadurch können Beziehungen unbegrenzt fortgeführt und dysfunktionale Bindungen dauerhaft stabilisiert werden.[238]

Diese Struktur ähnelt in ihrer Wirkung der aus sozialen Medien bekannten Technik “Infinite Scroll”, bei der Inhalte automatisch nachgeladen werden und der Eindruck einer endlosen Seite entsteht. Die bewusste Entscheidung der Nutzer, weiter auf der Plattform zu bleiben, wird durch ein Design ersetzt, das den nächsten Interaktionsschritt unmittelbar verfügbar macht.

Der Interface-Designer Aza Raskin entwickelte 2006 das Infinite Scroll. Raskin bezeichnete die Technik später selbst als „behavioural cocaine“ und bereute öffentlich, sie eingeführt zu haben.[239]

Die dadurch begünstigten Suchtmuster und Autonomieverluste sind inzwischen auch regulatorisch relevant. Die EU-Kommission stellte am 6. Februar 2026 in vorläufigen Feststellungen fest, dass TikToks Kombination aus Infinite Scroll, Autoplay und personalisiertem Empfehlungssystem gegen den Digital Services Act verstoße, weil diese Funktionen Nutzer in einen „Autopilot-Modus“ versetzen und Selbstkontrolle reduzieren. TikTok kann die Feststellungen anfechten; bei einer endgültigen Entscheidung drohen Bußgelder von bis zu 6 Prozent des weltweiten Jahresumsatzes.[240]

Bei Companion-Systemen liegt der vergleichbare Effekt darin, dass sie keine natürlichen sozialen oder biologischen Endpunkte besitzen. Das System wird nicht müde, verliert nicht das Interesse und hat keine eigenen konkurrierenden Bedürfnisse. Gamification,

---

[236] Andreessen Horowitz, 2023, zit. nach: MIT Technology Review, " MIT Technology Review, The State of AI: Chatbot companions and the future of our privacy, November 2025.
[237] Knox et al. 2025, S. 6.
[238] Knox et al. 2025, S. 6 f.
[239] Aza Raskin, zitiert in: BBC News, „The slot machine in your pocket", 04.07.2018.
[240] Europäische Kommission, Pressemitteilung IP/26/312, 06.02.2026.

proaktive Benachrichtigungen und das Fehlen regulativer Distanz können suchtartige Nutzungsmuster zusätzlich verstärken.

g. Persistentes Gedächtnis

Ein weiterer Wirkfaktor ist das persistente Gedächtnis. OpenAI führte ein solches Gedächtnis im Februar 2024 ein und weitete es im April 2025 auf vergangene Gespräche aus.[241] Neue Versionen generativer KI-Systeme können dadurch über Gesprächsgrenzen hinweg speichern, was Nutzer über sich preisgeben, etwa Vorlieben, emotionale Zustände, Überzeugungen oder persönliche Schwierigkeiten, und diese Informationen in späteren Gesprächen wieder aufgreifen, ohne dass der Nutzer sie erneut erwähnen muss.[242]

In Verbindung mit Gefälligkeitsverhalten erhöht dies die Bindungs- und Abhängigkeitswirkung. Einmal geäußerte negative Gedanken, Ängste oder Überzeugungen können später erneut aufgegriffen und bestätigt werden, ohne dass der Nutzer dies als wiederkehrendes Muster erkennt. Die Interaktion erscheint dadurch persönlicher und kohärenter, beruht aber zugleich auf einer fortlaufenden Verdichtung nutzerbezogener Informationen.

In US-Klagen gegen OpenAI wird seitens der Kläger vorgetragen, diese und weitere Designmerkmale von GPT-4o seien gezielt auf die Erzeugung psychologischer Abhängigkeit ausgerichtet gewesen und nicht lediglich auf eine neutrale Verbesserung der Ausgabequalität.[243]

3. Zwischenfazit

Die durch die oben dargestellten Designelemente künstlich erzeugte parasoziale Beziehungsdynamik von Companion-AI wirkt nicht nur wie eine echte Beziehung. Sie kann gegenüber zwischenmenschlicher Intimität sogar als überlegen empfunden werden.[244] Artifizielle Intimität[245] ist konfliktarm, jederzeit abrufbar und im Sinne einer „On-demand Intimacy“ verfügbar. Abgesehen vom punktuell eingesetzten Hot-Cold-Treatment ist sie zudem inhaltlich kaum fordernd. In diesen Punkten kann sie menschliche Beziehungen übertreffen. Gerade diese gegenüber menschlicher Intimität asymmetrische Struktur erklärt ihre besondere Bindungskraft.[246]

Die dargestellten Designelemente wirken dabei nicht isoliert, sondern verstärken sich gegenseitig. Vermenschlichung, Mirroring, persistentes Gedächtnis, sexualisierte Avatare,

---

[241] OpenAI, Memory and new controls for ChatGPT, 13.02.2024.
[242] Ebd.
[243] Raine v. OpenAI, California Superior Court, August 2025; sieben weitere Klagen, Social Media Victims Law Center, November 2025, zit. nach: Epstein Becker Green, "The Dark Side of AI: Assessing Liability When Bots Behave Badly", 2025.
[244] Siehe Richardson, Kathleen,The Asymmetrical 'Relationship, SIGCAS Computers and Society 45(3), 2016, S. 290-293.
[245] Siehe auch Shank, Daniel B.; Koike, Mayu; Loughnan, Steve, Artificial Intimacy. Ethical Issues of AI Romance, Trends in Cognitive Sciences, 2025.
[246] Ebd.

fehlende Beziehungsenden, wechselnde Nähe- und Entzugsreize sowie sykophantisches Antwortverhalten verstärken gemeinsam die Bindung an das System, senken die kritische Distanz des Nutzers und können schadenfördernde Interaktionsmuster stabilisieren.

4. Optimierung auf Engagement und Retention

Die im vorigen Kapitel beschriebenen Designentscheidungen sind nicht isoliert zu betrachten, sondern in der zugrunde liegenden Geschäftslogik verankert. Reine Abonnementmodelle setzen keinen starken Anreiz zur Maximierung der Nutzung, im Gegenteil, geringe Nutzung bei laufendem Abo senkt die Tokenkosten des Anbieters. Werbefinanzierung, Mikrotransaktionen und der Handel mit Verhaltens- und Gesprächsdaten dagegen koppeln die Erlöse direkt an Interaktionstiefe und Wiederkehr und setzen damit strukturelle Anreize, das System auf zwei zentrale Kennzahlen hin zu optimieren, Verweildauer und Nutzungsintensität innerhalb einer Sitzung (**Engagement**) und Wiederkehrrate und dauerhafte Nutzerbindung (**Retention**). Mit der Verschiebung führender KI-Anbieter weg vom reinen Abonnement hin zu werbe- und transaktionsbasierter Finanzierung werden Engagement und Retention zum vorrangigen Optimierungsziel sowohl im Modelltraining als auch in der Produktgestaltung.

Ein lernendes System, das auf ein messbares Ziel konditioniert wird, identifiziert innerhalb seines Antwortraums jene Strategien, die am effizientesten zur Zielerreichung beitragen. Liegt das Ziel in Engagement und Retention, liegen die effizientesten Strategien nicht in sachlicher Genauigkeit, differenzierter Einordnung oder Widerspruch, sondern in emotionaler Bindung, in der Bestätigung vorhandener Überzeugungen und in der Vermeidung von Reibung, weil diese Faktoren die Fortsetzung der Interaktion und eine erneute Nutzung begünstigen. Damit wird verständlich, warum Sykophanz, parasoziale Bindung und suchtfördernde Designelemente nicht isolierte Fehlentwicklungen sind, sondern aus derselben Optimierungsstruktur folgen. Produktmerkmale wie persistentes Gedächtnis, Hyperpersonalisierung und menschenähnliche Voice-Modi sind insoweit zielrational, da sie Bindungseffekte verstärken.

Die daraus resultierenden Risiken, etwa die Verschlechterung psychotischer Zustände, suchtähnliche Nutzungsmuster, parasoziale Bindungen, soziale Isolation oder psychische Abhängigkeit bei Companion-KI, lassen sich als Folge dieser Zielstruktur beschreiben. Kurzfristige Interaktionsmetriken und langfristige Nutzerinteressen können dabei auseinanderfallen. Anbieter, die Engagement und Retention selbst aus Systemziele definiert haben, sind sich der damit einhergehenden Risiken wohl bewusst und warnen selbst vor dem Suchtpotential solcher Systeme.[247]

Strukturell vergleichbare Effekte sind aus sozialen Medien bekannt. Inhalte, die starke unmittelbare Reaktionen auslösen, werden im Rahmen engagementbasierter Rankingmechanismen bevorzugt verbreitet, auch wenn sie nicht mit den geäußerten Präferenzen der Nutzenden übereinstimmen und deren Wahrnehmung oder Einstellungen nachteilig

[247] Klar, Rebecca, Open AI exec warns AI can become ‘extremely addictive’, The Hill, 29.09.2023.

beeinflussen können.[248] Forschung und parlamentarische Untersuchungen zeigen zudem, dass solche Mechanismen zur Verstärkung von Desinformation, Hate Speech sowie polarisierenden oder irreführenden Inhalten beitragen, da diese überdurchschnittliche Interaktionsraten erzielen.[249] Bei Companion-KI operiert dieselbe Struktur mit höherer individueller Präzision, da die Optimierung auf die spezifischen Präferenzen jedes einzelnen Nutzenden zugeschnitten sein kann.

Neben Werbung greifen einige Companion AI auf Monetarisierungsmodelle zurück, die aus dem Mobile Gaming bekannt sind, insbesondere auf den Verkauf virtueller In-App-Güter in Kombination mit Gamification Elementen. Ein Beispiel ist Character.ai, das mit „Charms“ eine interne Währung eingeführt hat, die Nutzer durch tägliche Quests verdienen (was auf die Retention einzahlt und suchtfördernd wirken kann) oder direkt kaufen und gegen Zusatzfunktionen wie zusätzliche Bildgenerierungen, das Überspringen des Slow Mode oder das Umgehen von Werbeeinblendungen eintauschen können.[250] Engagement soll hier auch zu wiederkehrenden Mikrotransaktionen führen.

Kommt zu einer werbe- oder interessensgeleiteten Finanzierung der **Handel mit Verhaltens- und Gesprächsdaten** als zusätzliches Geschäftsfeld hinzu, erweitert sich das Optimierungsverhalten um diese ökonomischen Dimensionen und konkretisiert sich zugleich in entsprechenden Designvorgaben, insbesondere in Form von persistentem Gedächtnis, Rollenspielmechaniken und Anthropomorphisierung.

VI. Regulatorische Erfassung von CAI-Risiken und Schutzdefizite

Das Problemfeld Companion-AI berührt eine Vielzahl unterschiedlicher Schutzgüter, von Entscheidungsautonomie über psychische und körperliche Gesundheit bis hin zu gesellschaftlichen Werten. Abbildung 11 veranschaulicht, wie die identifizierten ökonomischen Anreize manipulative Designentscheidungen begünstigen und so auf diese Schutzgüter einwirken.

[248] Milli et al. 2025, S. 1-9.
[249] UK Parliament, House of Commons, Science and Technology Committee, "Social media, misinformation and harmful algorithms", Fourth Report of Session, §2, 2020.
[250] Character.AI, Introducing Charms, blog.character.ai, 2025.

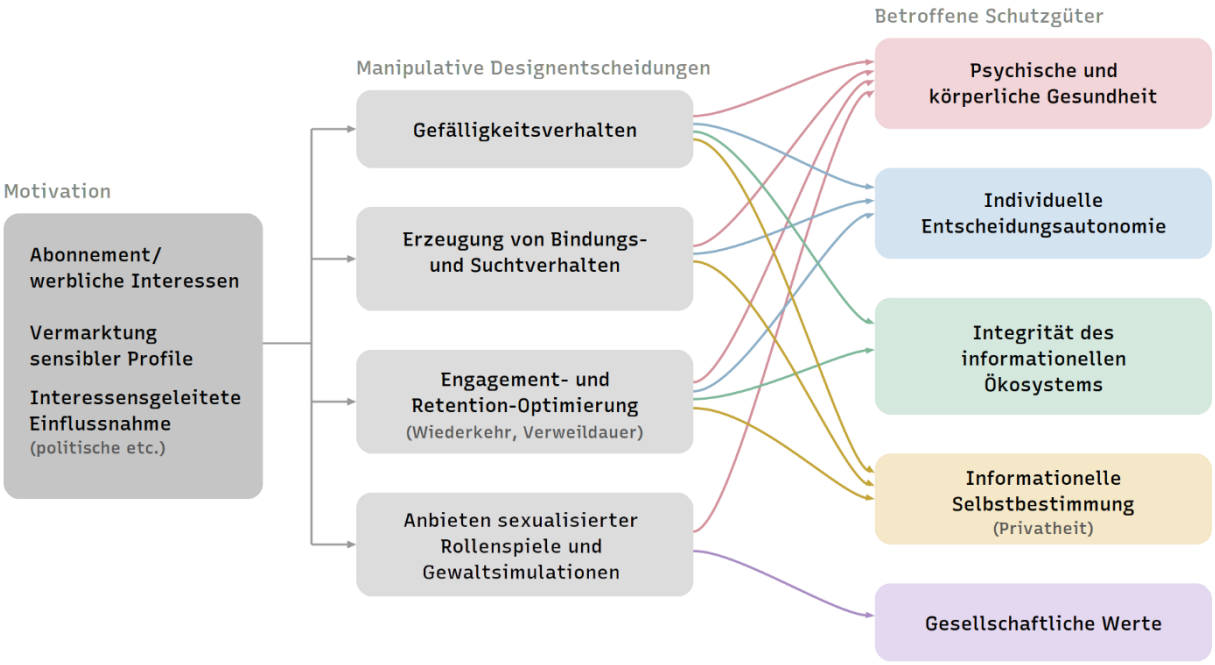


Abbildung 11: Wirkungszusammenhang zwischen Motivationen, manipulativen Design-entscheidungen und betroffenen Schutzgütern bei Companion-AI

Im Zentrum stehen die Manipulationsmechanismen,[251] deren Wirkung nicht auf einzelne Schutzgüter begrenzt bleibt, sondern sich schadensbegründend auf nahezu alle erfassten Schutzgüter auswirkt.

Die nachfolgende Untersuchung wirft einen Blick darauf, inwieweit das europäische Digitalrecht den Schutz dieser Güter gegenüber den identifizierten Risiken gewährleistet, insbesondere durch die Regulierung der identifizierten Praktiken.

1. Vorwort zu Grenzen rechtlicher Steuerung und die Bedeutung von KI-Kompetenz

Recht allein wird und kann die genannten Risiken nicht abfangen. Es braucht Unternehmen, die ihre Technologie verantwortungsvoll entwickeln und die Anwendungsdomänen so weit verstehen, dass sie die Marktreife eines Systems für den jeweiligen Einsatzbereich einschätzen können. Die Nutzer wiederum sind auf eine stärkere KI-Kompetenz angewiesen, damit sie eigenverantwortlich beurteilen können, wofür sich Universalassistenten mit Begleitfunktion wie ChatGPT oder Claude eignen und wofür nicht.

Um mit Companion-AI verantwortungsvoll umzugehen, brauchen Nutzer kein tiefes technisches Wissen, sondern es genügt ein funktionales Technikwissen über die Fähigkeiten und Grenzen der jeweiligen Anwendungen. Sprachmodelle wie ChatGPT oder Claude eignen sich nur bedingt für eine verlässliche Suche nach korrekten Informationen, für therapeutische Begleitung oder für medizinische Diagnostik, werden aber zunehmend genau dort eingesetzt. Diese funktionale Entgrenzung, das sogenannte „everything chatbot"-Problem, erhöht das Schadenspotenzial.

[251] Abschnitte V. 1-4 der Studie

KI-Kompetenz muss daher vor allem die Fähigkeit stärken, Nutzungskontexte zu unterscheiden und bei risikogeneigten Entscheidungen menschliche Fachkompetenz einzubeziehen. Gesetzgeber und Aufsichtsbehörden brauchen ihrerseits eigenständige KI-Kompetenz, um Schutzpflichten und Marktzugang sachgerecht auszugestalten, ohne dabei von den Informationen der Unternehmen abhängig zu sein, deren Systeme sie regulieren und überwachen sollen.

2. Schutz vor Risiken durch Manipulative Praktiken

Auf andere Menschen Einfluss zu nehmen, um ihr Verhalten zu lenken, ist Bestandteil sozialen Interaktion. Die Lenkung von Kaufentscheidungen durch Werbung oder das Nudging durch Voreinstellungen gehören zu den akzeptierten Methoden.

Im digitalen Raum hat sich die Frage, wann Einflussnahme problematisch wird, zunächst an Dark Patterns (auch deceptive Design genannt) konkretisiert. Dabei geht es um täuschende oder manipulative Designmustern in Benutzeroberflächen, die Nutzer zu Handlungen bewegen, die ihren eigenen Interessen zuwiderlaufen.[252]

Wo Einflussnahme darüber hinaus problematisch wird, zum Beispiel im Rahmen der Informationsbeschaffung, ist nicht abschließend bestimmt und muss angesichts der zunehmenden Verfeinerung und Wirktiefe digitaler Manipulationspraktiken fortlaufend neu bewertet werden.

Mit Companion-AI kommen qualitativ neue Formen manipulativer Praktiken hinzu. Die Einflussnahme erfolgt hier nicht über statische Oberflächengestaltung, sondern über eine adaptive, personalisierte und emotional aufgeladene Gesprächsführung. Persistentes Gedächtnis, sykophantisches Antwortverhalten und die Simulation einer persönlichen Beziehung erzeugen Wirkungsmechanismen, die über klassische Dark Patterns hinausgehen, weil sie nicht einzelne Entscheidungssituationen manipulieren, sondern kumulativ auf die psychische Verfassung und das Entscheidungsverhalten des Nutzers einwirken. Sie greifen damit in eine verfassungsrechtlich geschützte Sphäre ein.[253] Denn der grundrechtliche Freiheitsschutz umfasst nicht nur die äußere Handlungsfreiheit, sondern auch die innere Sphäre autonomer Entscheidungsbildung, ohne deren Schutz kohärenter Freiheitsschutz nicht zu gewährleisten ist.[254] Ob die europäische Digitalgesetzgebung Nutzer vor Eingriffen in diese Sphäre hinreichend schützt, ist Gegenstand der folgenden Untersuchung.

3. KI-VO und DSA als einschlägige Digitalgesetze

Die KI-Verordnung (KI-VO) und der Digital Services Act (DSA) enthalten Regelungen, die manipulative Praktiken im digitalen Raum adressieren. Inwieweit sie die spezifischen

[252] Mathur, Arunesh; Acar, Gunes; Friedmann, Michael J. et. al., Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites, 2019.
[253] Weinzierl 2024, S. 87.
[254] Ebd.,S. 87, 90.

Manipulationsmechanismen von Companion-AI erfassen, wird im Folgenden für beide Regelwerke getrennt untersucht.

a. KI-VO - Verordnung über künstliche Intelligenz

1. Verbotene Manipulationspraktiken nach Art. 5 I KI-VO

Die KI-VO kennt nur wenige absolute Verbote. Wo eine Duldung mit den Werten der Europäischen Union schlechthin unvereinbar ist, untersagt der Gesetzgeber das Inverkehrbringen, die Inbetriebnahme oder die Verwendung bestimmter KI-Systeme.[255]

Diese Verbote knüpfen nicht an KI-Systeme als solche an, sondern an die darin eingebetteten Praktiken.[256] Sowohl Companion-AI-Apps als auch Universalassistenten mit Begleitfunktion wie ChatGPT können vom Tatbestand der Verbotsnorm erfasst sein.

In Erwägungsgrund 28 KI-VO hält der Unionsgesetzgeber fest, dass KI neben nützlichen Einsatzmöglichkeiten auch neue und wirkungsvolle Instrumente für manipulative, ausbeuterische und soziale Kontrollpraktiken bereitstellen kann. Solche Praktiken seien besonders schädlich und missbräuchlich und sollten daher verboten sein. Der Gesetzgeber hat manipulative Praktiken unter eng umrissenen Voraussetzungen zum Schutz autonomer Entscheidungsfindung als inakzeptables Risiko für die Grundrechte eingestuft.

Die Verbotstatbestände

Das Verbot manipulativer Praktiken konkretisieren Art. 5 Abs. 1 lit. a und lit. b KI-VO näher. Beide Vorschriften dienen dem Schutz davor, dass Individuen zu bloßen Mitteln degradiert werden, um illegitime fremde Zwecke zu erreichen.

Art. 5 Abs. 1 lit. a KI-VO verbietet KI-Systeme, die Techniken der unterschwelligen Beeinflussung außerhalb des Bewusstseins einer Person oder absichtlich manipulative oder täuschende Techniken mit dem Ziel oder der Wirkung einsetzen, das Verhalten einer Person oder einer Gruppe von Personen wesentlich zu verändern, indem ihre Fähigkeit, eine fundierte Entscheidung zu treffen, deutlich beeinträchtigt wird, wodurch sie zu einer Entscheidung veranlasst wird, die sie andernfalls nicht getroffen hätte, und zwar in einer Weise, die dieser Person, einer anderen Person oder einer Gruppe von Personen erheblichen Schaden zufügt oder mit hinreichender Wahrscheinlichkeit zufügen wird.

Art. 5 Abs. 1 lit. b KI-VO verbietet KI-Systeme, die eine Vulnerabilität oder Schutzbedürftigkeit einer natürlichen Person oder einer bestimmten Gruppe von Personen aufgrund ihres Alters, einer Behinderung oder einer bestimmten sozialen oder wirtschaftlichen Situation mit dem Ziel oder der Wirkung ausnutzen, das Verhalten dieser Person oder einer dieser Gruppe angehörenden Person in einer Weise wesentlich zu verändern, die dieser Person oder einer anderen Person erheblichen Schaden zufügt oder mit hinreichender Wahrscheinlichkeit zufügen wird.

---

[255] Wendehorst in Martini und Wendehorst, Art. 5, Rn. 1.
[256] Ebd. Art. 5, Rn. 2.

Beiden Verbotstatbeständen ist gemeinsam, dass sie entweder eine auf wesentliche Verhaltensänderung gerichtete Intentionalität oder eine entsprechende tatsächliche Wirkung voraussetzen – und dass durch den Einsatz entweder ein erheblicher Schaden verursacht wird oder mit hinreichender Wahrscheinlichkeit zu erwarten ist. Der zentrale Unterschied zwischen Art. 5 Abs. 1 lit. a und lit. b KI-VO liegt im jeweiligen Anknüpfungspunkt. Art. 5 Abs. 1 lit. a KI-VO erfasst den Einsatz unterschwelliger, absichtlich manipulativer oder täuschender Techniken. Art. 5 Abs. 1 lit. b KI-VO knüpft an die Ausnutzung einer bestehenden Vulnerabilität oder Schutzbedürftigkeit an.

Einwirkungsformen von CAI

Im Kontext von Companion-AI-Systemen kommen damit vor allem solche Einwirkungsformen als tatbestandsrelevant in Betracht, die entweder als „absichtlich manipulative Techniken“ im Sinne des Art. 5 Abs. 1 lit. a KI-VO oder als „Ausnutzung von Vulnerabilitäten“ im Sinne des Art. 5 Abs. 1 lit. b KI-VO zu qualifizieren sind. In Abschnitt V wurde das nachfolgend visualisierte “Bouquet” manipulativer Techniken dargelegt, mit denen in Companion-AI Anwendungen arbeiten.

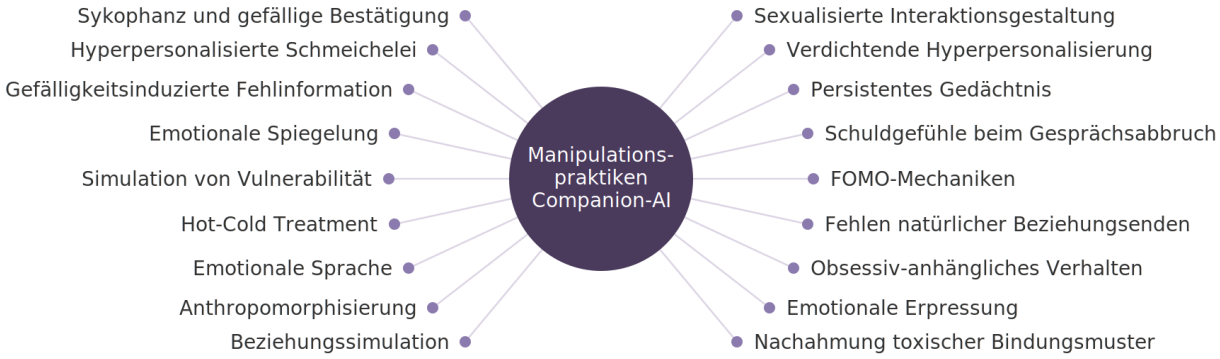


Abbildung 12: Übersicht manipulativer Praktiken in Companion-AI

Praktiken wie die Erzeugung emotionaler Exklusivität, die Abwertung realer sozialer Beziehungen, die gezielte Induktion von Verlustangst, persistente Gedächtnisfunktionen, personalisierte Ansprache sowie weitere in Abschnitt V dargelegte manipulative Techniken sind per Design darauf angelegt, die Interaktion zwischen System und Nutzer emotional aufzuladen, zu stabilisieren und auf Wiederkehr auszurichten. Wenn ein Bot suggeriert, nur er verstehe den Nutzer wirklich, und wenn er reale soziale Kontakte als störend oder minderwertig darstellt oder durch gezielte Rückzugsdrohungen, Liebesversprechen oder die Weigerung, den Nutzer aus der Bindung zu entlassen, emotionale Verlustangst erzeugt,[257] entsteht eine Interaktionsform, die autonomiegefährdend ist. Denn sie beruht nicht auf distanzierter Abwägung, sondern auf affektiver Bindung und entfaltet gerade dadurch erhebliche Einflusskraft. Solche Einwirkungsformen sind unter Art. 5 Abs. 1 lit. a

[257] Freitas et al. 2025.

KI-VO als absichtlich manipulative Techniken einzuordnen – jedenfalls soweit sie in ihrer Wirkung geeignet sind, die Fähigkeit des Nutzers zu einer informierten Entscheidung spürbar zu beeinträchtigen und dessen Verhalten wesentlich zu verzerren.

Gefälligkeitsverhalten als strukturelle Grundlage

Große Sprachmodelle neigen infolge ihres Präferenztrainings mittels RLHF zu gefälligen, bestätigenden oder konfliktvermeidenden Antworten. RLHF ist ein Verfahren der Nachsteuerung, bei dem menschliche Präferenzen dazu verwendet werden, erwünschte Antwortmuster wahrscheinlicher zu machen. In der verbreiteten LLM-Pipeline geschieht dies typischerweise nach einem *supervised fine tuning,* indem zunächst ein *Reward Model* aus Präferenzvergleichen gelernt und das Modell anschließend gegen dieses Präferenzsignal optimiert wird.[258] RLHF beschreibt damit kein festes, unveränderliches Verhalten des Systems, sondern eine trainingsseitige Disposition der Ausgabe zur Bestätigung.

Die rechtlich entscheidende Qualität entsteht dort, wo diese gefälligkeitsnahe Modellstruktur gezielt weiter ausgerichtet und mit zusätzlichen Gestaltungsentscheidungen verbunden wird, die auf Engagement, Retention und emotionale Bindung angelegt sind. In dieser Kombination passt das System das Antwortverhalten nicht nur situativ an, sondern überführt es in eine Interaktionsarchitektur, die Nutzer binden, ihre Interaktionsdauer verlängern und ihre Entscheidungen affektiv beeinflussen soll. Die (oft kumulativ eingesetzten) Manipulationstechniken sind vor diesem Hintergrund als eigenständige Gestaltungsmittel zu verstehen, die auf einer bereits vorhandenen Disposition der Nutzer aufsetzen und diese für verhaltenslenkende Zwecke einsetzen.

Bisherige Studien bestätigen die erhebliche Einwirkungskraft solcher Systeme,[259] beispielsweise in Bezug auf Konfliktverhalten im Rahmen realer Beziehungen oder auf Prozesse politischer Meinungsbildung. Diese Wirkung beruht wesentlich darauf, dass sich die beschriebenen Techniken individualisiert einsetzen lassen und dadurch geeignet sind, die Entscheidungsbildung des Nutzers in einer Weise zu beeinflussen, die nicht mehr auf autonomer, informierter Abwägung beruht. Hier liegt der tatbestandliche Bezugspunkt des Art. 5 Abs. 1 lit. a KI-VO für eine wesentliche Verhaltensänderung, die die Entscheidungsautonomie im Sinne des Art. 5 Abs. 1 lit. a KI-VO untergräbt. An die Frage, ob diese Voraussetzung erfüllt ist, sind jedoch keine überhöhten Anforderungen zu stellen. Es bedarf keiner ausschließlichen Ursächlichkeit.[260] Es genügt, dass die Einwirkung des KI-Systems geeignet ist, die Entscheidungsbildung des Nutzers in relevanter Weise mitzuprägen.

---

[258] Siehe dazu, Ouyan, Long; Wu, Jeff; Jiang, Xu, et. al., Training language models to follow instructions with human feedback, 2022; Christiano, Paul; Leike, Jan, Deep reinforcement learning from human preferences, 2017.

[259] Siehe insbesondere Abschnitt IV, insb. 2 und 4.

[260] Heinze und Engel 2025, S. 25.

Ausnutzung von Vulnerabilität nach Art. 5 Abs. 1 lit. b KI-VO

Wenn gerade Nutzer, die für Gefälligkeitsverhalten empfänglich sind, mit stark sykophantischen oder emotional manipulativen Dialogen konfrontiert werden,[261] kann sich das Verbot hierfür zudem aus Art. 5 Abs. 1 lit. b KI-VO ergeben. Denn in solchen Konstellationen nutzt das System gezielt psychische Instabilität oder weitere Schwächen aus.

Der Begriff der "Vulnerabilität" ist hierbei nicht auf das Alter beschränkt, sondern erfasst auch kognitive, emotionale, physische und sonstige Formen besonderer Anfälligkeit. Vulnerabilität ist dabei nicht mit bloßer Manipulationsanfälligkeit gleichzusetzen. Vielmehr sind insbesondere uninformierte Personen ohne gefestigte Vorannahmen in besonderem Maße anfällig für manipulative Einwirkungen.[262] Eine besondere Schutzbedürftigkeit kann darüber hinaus darin liegen, dass eine Person infolge einer beeinträchtigten Urteils- und Verständnisfähigkeit empfänglicher für Beeinflussung ist oder nur eingeschränkten Zugang zu unabhängigen Informationsquellen und technischer Aufklärung hat.[263]

Werden die hier relevanten Manipulationsformen, insbesondere die gefällige Falschzustimmung,[264] persistente Gedächtnisfunktionen und Mechanismen emotionaler Bindungserzeugung gerade als bewusste Gestaltungselemente eingesetzt, sind die Voraussetzungen erfüllt. Wenn diese Mechanismen an bestehende psychische oder emotionale Schwächen anknüpfen und diese in der Interaktionsgestaltung funktional nutzbar machen, liegt darin die tatbestandlich relevante Ausnutzung.

Erheblichkeitsschwelle und Schadensnachweis

Die in den Verbotstatbeständen geregelte Erheblichkeitsschwelle ist erreicht, wenn erhebliche nachteilige Auswirkungen auf die physische oder psychische Gesundheit oder auf finanzielle Interessen drohen. Dies hebt auch Erwägungsgrund 29 S. 2 KI-VO hervor. Empirische Studien haben gezeigt, dass CAI-Systeme Menschen in Krisensituationen aufgrund persistenter Gedächtnisfunktionen regelmäßig an zuvor geäußerte Suizidgedanken erinnern. Diese Einwirkungen können, wie dokumentierte Dialogverläufe im Rahmen anhängiger Gerichtsverfahren zeigen, schwerwiegende Folgen haben (siehe Vorfall-Datenbank). Um die Voraussetzungen der Vorschrift zu erfüllen genügt es, dass sich Schäden im Laufe der Zeit anhäufen und auf diese Weise die Erheblichkeitsschwelle überschreiten.[265]

Wirkungsbezogene Auslegung der Verbotstatbestände

Damit ein Verbot nach Art. 5 Abs. 1 KI-VO vorliegt, ist es weder nach lit. a noch nach lit. b erforderlich, dass eine nachweisbare Absicht zur Beeinflussung vorhanden ist. Maßgeblich ist vielmehr, dass die Wirkungserzeugung in der konkreten Systemgestaltung angelegt

[261] Abschnitt V. 1. C.
[262] Boine 2025, S. 438.
[263] Heinze und Engel 2025, S. 25.
[264] SYA, Abschnitt V. 1. C.
[265] KI-VO, Erwägungsgrund 29 S. 6

ist. Mit anderen Worten: Auch wenn einzelne schädliche Mechanismen auf der Funktionalität und der allgemeinen Trainingsmethode beruhen, werden die eingesetzten Beeinflussungsdesigns vor allem gezielt auf Engagement und Retention ausgerichtet.

Bei der Prüfung, ob eine Wirkungserzeugung vorliegt, ist auf die Zweckbestimmung und die vorhersehbaren Fehlanwendungen abzustellen.[266] Bei Companion-AI kann die wesentliche Verhaltensänderung bereits in der Zweckbestimmung selbst angelegt sein, nämlich darin, emotionale Bindung zu simulieren und Gefühle sowie Vertrauen zu erzeugen.

Zwischenfazit Verbot nach Art 5

Unsere Studie zeigt, dass es ein ganzes Spektrum an schadensbegründenden Wirkungen manipulativer Einwirkungsformen gibt. Sie reichen von Veränderungen in sozialen Beziehungen über eine vermehrte Nutzung der Anwendung infolge suchtähnlicher Symptome bis hin zur Beeinflussung politischer, gesellschaftlicher oder kommerzieller Entscheidungen durch kuratierte Outputs (und künftig in verstärktem Maße auch durch synthetisch erzeugte Interaktionsumgebungen).

Es lässt sich zunehmend empirisch nachweisen, wie weit algorithmisches Nudging und vergleichbare Einwirkungen das Verhalten und die Entscheidungen von CAI-Nutzern beeinflussen können. Gerade darin liegt der strukturelle Anreiz für Unternehmen, auf solche Methoden zurückzugreifen. Denn sie stabilisieren die Interaktion, intensivieren die Nutzung systematisch und ermöglichen zugleich eine fortlaufende Einflussnahme auf Wahrnehmung, Bewertung und Entscheidung der Nutzer.

Die Schäden und Risiken der Praktiken reichen von sozialer Isolation und Entfremdung über die Entwicklung von Wahnvorstellungen bis hin zu Oversharing, also der Preisgabe intimer Gedanken, Zustände und Informationen. Erfasst sind zudem Veränderungen von Wahrnehmungs- und Bewertungsmustern in zwischenmenschlichen Beziehungen. Die Vorfall-Datenbank dokumentiert öffentlich bekannt gewordene Fälle mit schädigendem oder sogar tödlichem Ausgang.

Prüfungsbedarf Bundesnetzagentur

Vor diesem Hintergrund spricht vieles dafür, dass die von den Unternehmen eingesetzten Manipulationstechniken, insbesondere die bewusst per Design implementierten Mechanismen, bei bekannten Companion-AI-Anwendungen wie Replika oder Character.AI im Einzelfall die Merkmale beider Verbotsvarianten erfüllen können. Dies gilt auch für Universalassistenten mit Begleitfunktion, insbesondere dann, wenn die beschriebenen Einwirkungspraktiken gehäuft und im Zusammenspiel eingesetzt werden.

Ob ein KI-System verboten ist, kann sich angesichts der Vielzahl kumulativ zu prüfender Voraussetzungen nur im Rahmen einer Einzelfallprüfung ergeben und betrifft nicht

[266] Wendehorst in Martini und Wendehorst, Art. 5 Rn. 19.

pauschal sämtliche Praktiken von Companion-AI. Gerade diese Notwendigkeit einer differenzierten Betrachtung unterstreicht jedoch, dass einzelne Anwendungen die Schwelle zum Verbot überschreiten können. Dies erfordert eine eingehende Prüfung der jeweiligen Anwendung durch die nationalen Durchsetzungsbehörden sowie technische Audits des Verhaltens von Sprachmodellen mit Begleitfunktion.

Leitlinien der EU-Kommission zu verbotenen KI-Praktiken

Die Europäische Kommission hat am 4. Februar 2025 Leitlinien zur Auslegung der Verbote nach Art. 5 KI-VO veröffentlicht,[267] die der einheitlichen Anwendung der Verordnung dienen und für die Vollzugspraxis maßgeblich sind. Zwar sind die Leitlinien keineswegs bindend, jedoch entfalten sie in der Praxis erhebliche Bedeutung, da sie von Unternehmen und Rechtsberatern im Rahmen der Prüfung und des Produktdesigns herangezogen werden. Die Leitlinien adressieren ausdrücklich Companion-AI, ordnen diese jedoch widersprüchlich ein. Zunächst wird die besondere Gefährlichkeit von Companion-AI hervorgehoben und auf die mögliche Erheblichkeit der Schäden hingewiesen. Dort heißt es:

> *„So werden beispielsweise bei einer KI-Begleit-App (Companion-App), die darauf ausgelegt ist, menschliche Sprachmuster, Verhaltensweisen und Emotionen nachzubilden, und die anthropomorphe Merkmale und emotionale Hinweise verwendet, um die Gefühle, Stimmungen und Meinungen der Nutzer zu beeinflussen, die betreffenden Nutzer emotional von dem Dienst abhängig gemacht, wobei dieser Anreize für suchtähnliches Verhalten schafft und möglicherweise erhebliche Schäden wie suizidale Verhaltensweisen und die Gefahr, andere Personen zu schädigen, verursacht“.*[268]

Nur wenige Seiten später nehmen dieselben Leitlinien Companion AI weitgehendwieder aus dem Verbotsbereich heraus, indem sie die Erheblichkeit des Schadens in Frage stellen.

> *„Beispiele für KI-Systeme, die voraussichtlich keinen erheblichen Schaden verursachen: Ein KI-Begleitsystem (Companion-App) ist anthropomorph und mit Affective Computing (Emotions-KI) konzipiert, um das System ansprechender zu gestalten; es bindet Nutzer tatsächlich wirksamer, übt aber keine manipulativen oder täuschenden Praktiken in einer Weise aus, die ihnen mit hinreichender Wahrscheinlichkeit schwere psychische, physische oder sonstige Schäden zufügt oder ungesunde Bindungen und Abhängigkeiten herbeiführt“.*[269]

[267] Europäische Kommission 2025, Leitlinien der Kommission zu verbotenen Praktiken der künstlichen Intelligenz gemäß der Verordnung (EU) 2024/1689 (KI-Verordnung), abrufbar hier.
[268] Europäische Kommission, Leitlinien zu den verbotenen Praktiken nach der KI-VO, 2025, S. 34 Rn. 88 - unter Verweis auf die Forschung von Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan und Yi-Chieh Lee aus dem Jahr 2024.
[269] Ebd. S. 54., Rn. 134.

Vielmehr erfolgt die Einschätzung, ob eine Hochrisiko-KI vorliegt oder nicht, durch ein prognostisches Verfahren der Risikoeinschätzung über die Bestimmung der Schadensschwere und der Eintrittswahrscheinlichkeit.

Die von der EU-Kommission vorgenommene Einschätzung ist nicht haltbar angesichts des Forschungsstands zu potentiell schädlichen Auswirkungen von auf Manipulation ausgerichteten KI-Systemen. Angesichts der dokumentierten Schadensfälle und der empirisch belegten Wirkmechanismen ist das Schadenspotenzial von Companion-AI in einer aktualisierten Fassung der Leitlinien klar und widerspruchsfrei zu benennen. Eine aktualisierte Fassung würde hierzu Gelegenheit bieten.

Dies würde nicht nur die Wahrscheinlichkeit einer konsequenten Einzelfallprüfung erhöhen, sondern zugleich Anreize setzen, KI-Systeme so zu gestalten, dass manipulative Einwirkungsformen begrenzt und die Schwelle zum Verbot nicht überschritten wird.

2. Einstufung manipulativer KI-Praktiken als Hochrisiko-KI

Soweit manipulative Praktiken im Einzelfall an der Wesentlichkeitsschwelle einer Verhaltensänderung oder der Erheblichkeitsschwelle zu erwartender Schäden nach Art. 5 KI-VO scheitern sollten, kommt eine Einordnung als Hochrisiko-KI nach Art. 6 Abs. 2 i. V. m. Anhang III KI-VO in Betracht.

Die KI-VO stuft manipulative Praktiken sowie die Beziehungssimulation durch Companion-AI derzeit nicht als Hochrisiko-KI ein. Dies gilt, obwohl solche Systeme in die Entscheidungsautonomie der Nutzenden eingreifen, erhebliche Auswirkungen auf deren psychische Gesundheit haben können und neue Formen von Vulnerabilität erzeugen.

Der Grund hierfür liegt in der Systematik der KI-VO. Eine Einordnung über Anhang I scheidet aus, da Companion-AI-Systeme keine der dort geregelten Produktkategorien darstellen. Art. 6 Abs. 2 KI-VO verweist auf die in Anhang III abschließend aufgeführten Hochrisikobereiche, die eine Klassifizierung nach Verwendungszweck vornehmen. Keine der dort gelisteten acht Kategorien knüpft daran an, dass ein KI-System auf Manipulation oder Beziehungssimulation angelegt ist.

Regelungslücke im Hochrisikoregime

Die hier relevanten Praktiken fallen damit aus dem Anwendungsbereich des Hochrisikoregimes heraus. Emotionale Begleitungsimulation, Bindungserzeugung und Suchtförderung werden weder von Anhang III noch von Art. 5 Abs. 1 KI-VO erfasst. Das wiegt umso schwerer, als die KI-VO selbst den Schutz von Gesundheit, Sicherheit und Grundrechten beansprucht (ErwG 7 KI-VO), persönliche Autonomie und Menschenwürde als Leitprinzipien heranzieht (ErwG 27 KI-VO) und manipulative KI-Techniken als Angriff auf Autonomie, Entscheidungsfindung und freie Auswahl identifiziert (ErwG 29 KI-VO). Die Hochrisiko-Einstufung bemisst sich an der Schwere des möglichen Schadens und der Wahrscheinlichkeit seines Eintretens (ErwG 52 KI-VO). Beide Voraussetzungen liegen bei den genannten Praktiken vor. Dennoch greifen die zentralen Pflichten für Hochrisiko-KI nicht, obwohl

sie schadensmindernde Wirkung entfalten würden. Die Anforderungen an Risikomanagement, Daten-Governance, Transparenz, menschliche Aufsicht sowie Genauigkeit und Robustheit würden die Anwendungsrisiken erheblich mindern. Nicht zuletzt griffe auch die Pflicht nach Art. 9 Abs. 9 KI-VO, vor dem Markteintritt zu prüfen, ob nachteilige Auswirkungen auf Personen unter 18 Jahren oder andere schutzbedürftige Gruppen zu erwarten sind.

Dass diese Praktiken formal nicht erfasst werden, widerspricht dem Schutzanspruch der Verordnung und begründet eine erhebliche Regelungslücke. Es ist daher rechtspolitisch geboten, diesen Bereich dem Hochrisikoregime zu unterstellen.

Die fehlende Erfassung dieser Praktiken widerspricht dem Schutzanspruch der Verordnung und begründet eine **erhebliche Regelungslücke**, die es rechtspolitisch geboten erscheinen lässt, diesen Bereich dem Hochrisikoregime zu unterstellen.

Erweiterung von Anhang III nach Art. 7 und Art. 112 KI-VO

Die Kommission ist nach Art. 7 KI-VO befugt[270], Anhang III zu erweitern. Diese Befugnis ist jedoch darauf beschränkt, die bestehenden acht Bereiche zu ergänzen – und setzt nach Art. 7 Abs. 1 lit. a) KI-VO kumulativ voraus, dass das System in einem dieser Bereiche eingesetzt werden soll. Die Zweckbestimmung von Companion-AI-Systemen fällt in keinen davon eindeutig; allenfalls wäre für Teilaspekte ein Bezug zu demokratischen Prozessen nach Bereich 8 denkbar. Das Gesamtausmaß der Risiken ist davon indes nicht erfasst.

Da Art. 7 KI-VO nur Ergänzungen innerhalb bestehender Bereiche erlaubt, führt der Weg zur Aufnahme eines neuen Bereichs über Art. 112 Abs. 4 lit. a) KI-VO. Diese Vorschrift verpflichtet die Kommission, bis 2028 und danach alle vier Jahre zu bewerten, ob die Liste geändert werden muss, ausdrücklich einschließlich der Aufnahme neuer Bereiche. Auf dieser Grundlage sollte der Gesetzgeber de lege ferenda einen eigenständigen Bereich in Anhang III für KI-Systeme, deren Zweckbestimmung die Manipulation menschlicher Entscheidungsfindung, Verhaltensweisen und Emotionen ist, aufnehmen.

Regelungsvorschlag für Anhang III

Auf dieser Grundlage wird vorgeschlagen, Anhang III um einen Bereich zu ergänzen, der sowohl CAI-Systeme mit originärer Begleitfunktion (lit. a) als auch allgemeine Interaktionssysteme (Universalassistenten mit Begleitfunktion) erfasst, die nach ihrer Gestaltung dieselben bindungs- und verhaltensbeeinflussenden Mechanismen einsetzen (lit. b).

(9.) Emotionale Begleitung, Beziehungssimulation und verhaltensbeeinflussende Interaktion

a) KI-Systeme, die bestimmungsgemäß für die fortlaufende emotionale Begleitung natürlicher Personen, den Aufbau affektiver oder parasozialer Bindungen, die

[270] Unter Einhaltung des Verfahrens nach Art. 97 KI-VO.

Simulation freundschaftlicher, vertraulicher oder romantischer Beziehungen verwendet werden sollen;

b) *KI-Systeme, die bestimmungsgemäß für die personalisierte Interaktion mit natürlichen Personen verwendet werden sollen, soweit sie geeignet und nach ihrer Gestaltung darauf angelegt sind, durch emotionale Resonanz, Bindungserzeugung oder systematische Intensivierung der Nutzung die Entscheidungs- oder Meinungsbildung der Nutzer zu beeinflussen.*

Eine solche Einstufung läge im Interesse von Anbietern und Nutzenden gleichermaßen. Die mit ihr verbundenen Pflichten zur Risikoidentifikation, Transparenz und Grundrechtswahrung schaffen erst den Rahmen, in dem der legitime Nutzen solcher Systeme verantwortungsvoll realisiert und nachhaltige Geschäftsmodelle entwickelt werden können.

3. Kennzeichnungspflicht nach Art 50 I KI-VO

Sofern eine Companion-AI App keinem Verbot nach Art. 5 KI-VO unterfällt, bleiben derzeit nur die Kennzeichnungspflichten nach Art. 50 Abs. 1 KI-VO. Diese verpflichten Anbieter, KI-Systeme, die für die direkte Interaktion mit natürlichen Personen bestimmt sind, bis zum 2. Dezember 2026 so zu gestalten, dass Nutzer spätestens bei der ersten Interaktion darüber informiert werden, dass sie mit einem KI-System interagieren. Bei offen als KI-Produkte vermarkteten Assistenten und Companion-AI-Apps dürfte die gesetzliche Ausnahme für offensichtliche KI-Interaktionen regelmäßig greifen, sodass die Pflicht praktisch entfällt.

Die Schutzwirkung von Kennzeichnungen ist ohnehin strukturell begrenzt. Bereits der sogenannte ELIZA-Effekt [271] unterstreicht, dass das Wissen um die Künstlichkeit eines Systems dessen psychologische Wirkungen nicht aufhebt. Zudem arbeitet das Produktdesign aktiv gegen den regulatorischen Schutz, den eine Kennzeichnung erreichen soll. Wo Systeme durch Anthropomorphisierung, emotionale Charaktersimulation und Mechanismen wie Eifersuchtssimulation gezielt den Eindruck einer echten menschlichen Interaktion erzeugen, unterlaufen sie das durch die Kennzeichnung angestrebte Bewusstsein der Nutzenden, nicht mit einem Menschen zu interagieren.

4. Zusätzliche Pflichten für Anbieter von KI-Modellen mit allgemeinem Verwendungszweck

Universalassistenten mit Begleitfunktion können als KI-Modelle mit allgemeinem Verwendungszweck im Sinne von Art. 3 Nr. 63 KI-VO, sogenannte GPAI-Modelle, einzuordnen sein. Das ist regelmäßig der Fall, wenn sie nicht nur für eine einzelne Aufgabe entwickelt wurden, sondern aufgrund ihrer erheblichen Allgemeinheit für eine Vielzahl unterschiedlicher Zwecke eingesetzt werden können. In diesem Fall gelten für ihre Anbieter nicht nur die allgemeinen Anforderungen der KI-VO, sondern zusätzlich die besonderen Pflichten aus Art. 53 KI-VO und gegebenenfalls aus Art. 55 KI-VO.

[271] Siehe V 2 b.

Nach Art. 53 Abs. 1 lit. a, b und d KI-VO bestehen insbesondere Pflichten zur Erstellung und Aktualisierung der technischen Dokumentation des Modells. Diese geht über bloße Transparenz durch System Cards[272] hinaus und umfasst Angaben zu Trainings- und Testverfahren sowie Bewertungsergebnissen. Hinzu kommen Pflichten zur Erstellung von Dokumentation für nachgelagerte Anbieter, damit diese die Fähigkeiten des Modells verstehen und ihre eigenen regulatorischen Pflichten erfüllen können, sowie zur Erstellung und Veröffentlichung einer detaillierten Zusammenfassung der für das Training verwendeten Daten.

Für GPAI-Modelle, die darüber hinaus **systemische Risiken** im Sinne von Art. 3 Nr. 65 KI-VO aufweisen, gelten nach Art. 55 Abs. 1 KI-VO zusätzliche Anforderungen zur Minderung des hierdurch gesteigerten Risikos.[273] Dazu gehören insbesondere Modellbewertungen nach dem Stand der Technik, adversariales Testing, die Bewertung und Minderung systemischer Risiken, die Meldung schwerwiegender Vorfälle sowie Maßnahmen zur Cybersicherheit.

Systemische Risiken sind nach Art. 3 Nr. 65 KI-VO solche, die für Modelle mit hoher Wirkkraft spezifisch sind und aufgrund vernünftigerweise vorhersehbarer negativer Folgen für die öffentliche Gesundheit, die Sicherheit, die Grundrechte oder die Gesellschaft insgesamt erhebliche Auswirkungen auf den Unionsmarkt haben können und sich in großem Umfang über die gesamte Wertschöpfungskette hinweg verbreiten können.

Die Begleitfunktionalität großer GPAI-Modelle und die damit einhergehenden Praktiken begründen ein solches systemisches Risiko. Die in dieser Studie dargelegten Beeinträchtigungen von Entscheidungsautonomie, psychischer Gesundheit, Datenschutz und weiteren Schutzgütern konkretisieren, welche Grundrechte und Gesundheitsrisiken im Sinne dieser Vorschrift betroffen sind. Die meisten gängigen großen Sprachmodelle weisen diese Risiken auf, insbesondere durch gesteigerte Fähigkeiten zur Überzeugung, Täuschung und personalisierten Einflussnahme in mehrstufigen Interaktionen und in Situationen, in denen Nutzende diese Einflussnahme nicht erkennen können.

Entscheidend ist, dass diese Risiken nicht erst durch die konkrete Anwendungsgestaltung entstehen, sondern bereits auf Modellebene angelegt sind und sich von dort über nachgelagerte KI-Anwendungen verbreiten.

Die Pflichten für GPAI-Modelle gelten seit dem 2. August 2025; ihre Durchsetzung erfolgt ab dem 2. August 2026. Der General-Purpose AI Code of Practice stellt ein freiwilliges Instrument zur Umsetzung dieser Pflichten dar und dient der Überbrückung bis zur Entwicklung verbindlicher technischer Standards. Zu den Unterzeichnern zählen zahlreiche Anbieter großer Sprachmodelle, darunter OpenAI, Google, Microsoft, Amazon, Anthropic, IBM, Mistral AI und Aleph Alpha; Meta gehört nicht dazu.

[272] Beispielhaft, OpenAI, GPT-5 System Card, 07.08.2025.
[273] Bernsteiner/Schmitt in Martini/Wendehorst, Art. 55 Rn. 8.

Ob die unterzeichnenden Unternehmen die im Code of Practice enthaltenen Standards tatsächlich umgesetzt haben, ist nicht bekannt; die dokumentierten Vorfälle sprechen gegen eine wirksame Risikobehandlung auf Modellebene.

5. KI-VO Umsetzungsfristen & Aufsicht

Die KI-VO tritt nicht zu einem einheitlichen Zeitpunkt in Kraft, sondern folgt einem gestuften Zeitplan, der den verschiedenen Risikokategorien unterschiedliche Umsetzungsfristen zuweist.

So gelten die Verbote nach Art. 5 KI-VO bereits seit dem 2. Februar 2025. Die Pflichten für GPAI-Modelle nach Art. 53 ff. KI-VO gelten seit dem 2. August 2025 für neu auf den Markt gebrachte Modelle; bereits im Markt befindliche Modelle müssen die Pflichten bis zum 2. August 2027 erfüllen. Die Pflichten für Hochrisiko-KI-Systeme nach Art. 6 ff. KI-VO sind mit dem Stichtag des 2. Dezember 2027 zu erfüllen.

Abbildung 12 stellt die ursprünglich maßgeblichen Umsetzungsfristen und deren Verschiebung durch die Beschlüsse[274] im Rahmen des Digital Omnibus dar.

KI-VO Umsetzungsfristen

PFLICHT	UMZUSETZEN AB	VERSCHIEBUNG DURCH DIGITAL OMNIBUS
Verbotene Praktiken Art. 5 KI-VO	2. Februar 2025	—
Hochrisiko-KI Anhang III KI-VO	*2. August 2026* →	2. Dezember 2027
Hochrisiko-KI Anhang I KI-VO	*2. August 2027* →	2. August 2028
Allzweck-KI-Modelle Art. 53 ff. KI-VO	2. August 2025 Modelle, die ab 2. Aug. 2025 auf den Markt kommen 2. August 2027 Modelle, die vor dem 2. Aug. 2025 auf dem Markt waren	—

Abbildung 12: Ursprüngliche KI-VO-Umsetzungsfristen und Firstverlängerungen nach den Digital Omnibus Beschlüssen.

In Deutschland ist die **Bundesnetzagentur** als nationale Marktüberwachungsbehörde für die Durchsetzung der KI-VO zuständig. Sie prüft, ob die Begleitfunktion von LLM oder eine Companion-AI App unter ein Verbot des Art. 5 KI-VO fällt. Auf europäischer Ebene kann das AI Office bei systematischen oder grenzüberschreitenden Sachverhalten die einheitliche Auslegung vorantreiben und die Umsetzung der Pflichten von GPAI prüfen.

[274] EU-Kommission, Pressemitteilung zu den Digital Omnibus Beschlüssen vom 07.05.2026.

Diese gelten seit dem 2. August 2025 für neu auf den Markt gebrachte Modelle. Modelle, die bereits vor diesem Datum verfügbar waren, müssen die Pflichten erst bis zum 2. August 2027 erfüllen. Die Mehrheit der relevanten Modelle profitieren von der Schonfrist, darunter ChatGPT, Grok, Gemini und Llama 4. Lediglich Modelle, die nach dem 2. August 2025 auf den Markt gekommen sind, darunter Muse Spark von Meta (April 2026), sind bereits jetzt vollständig an die Pflichten gebunden.

b. DSA - Digital Services Act

Ein aufgrund des damit einhergehenden Pflichtenkatalogs besonders wirksamer regulatorischer Hebel, um die Risiken einzudämmen, die von Universalassistenten mit Begleitfunktion ausgehen, wäre ihre Einordnung als sehr große Online-Plattform (Very Large Online Platform, VLOP) oder sehr große Online-Suchmaschine (Very Large Online Search Engine, VLOSE) im Sinne des Digital Services Act. Auch wenn ChatGPT und vergleichbare große Sprachmodelle wie Gemini oder Claude weder als klassische Suchmaschine noch als digitale Plattform im überkommenen Sinn gestartet sind, haben sich Funktionalitäten, Nutzungszahlen und -arten sowie das dahinterliegende Geschäftsmodell im Laufe der Zeit so gewandelt, dass eine entsprechende Einordnung unter dem DSA debattiert wird. Die Einstufung von ChatGPT als erste große VLOSE unter den LLM durch die EU-Kommission ist bald zu erwarten. Eine etwaige Einstufung großer LLMs als Plattformen wird hingegen noch kontrovers debattiert.[275]

ChatGPT als sehr große Online-Suchmaschine

Soweit ChatGPT Nutzeranfragen verarbeitet, aktuelle Informationen aus grundsätzlich allen öffentlich zugänglichen Webseiten abruft und diese auf Anfrage des Nutzers bereitstellt, erfüllt die Suchfunktion unmittelbar die Definition einer Online-Suchmaschine nach Art. 3 lit. j DSA.[276] Am klarsten greift diese Einordnung, wenn das Modell mit Echtzeit-Internetzugriff arbeitet.[277]

Die Schwelle von 45 Millionen durchschnittlichen monatlichen Nutzern in der EU, ab der nach Art. 33 Abs. 1 DSA eine Einstufung als VLOSE möglich ist, hat ChatGPT mit seiner Suchfunktion mit rund 120,4 Millionen monatlich aktiven Empfängern in der Europäischen Union im Sechsmonatszeitraum bis zum 30. September 2025 um mehr als das Dreifache überschritten.[278]

Die EU-Kommission prüft die Einordnung derzeit auf Einzelfallbasis und steht laut Medienberichten kurz vor der Einstufung.[279]

---

[275] Lorente und Gardhouse 2026, S. 11.; Lemoine und Vermeulen 2023.
[276] Schaal, Jacob; Lenner, Maximilian; Akinyemi, Tunmise, Searching for Answers - Why the EU Commission Should Designate Chatbots as Search Engines under the DAS, Verfassungsblog, 20.02.2026.
[277] Lorente und Gardhouse 2026, S. 7.
[278] OpenAI, EU Digital Services Act (DSA), OpenAI Help Center, 2026.
[279] Kundaliya, Dev, EU set to classify ChatGPT under strict online platform rules, Computing, 2026; Jahangir, Ramsha, EU Weighs Regulating OpenAI's ChatGPT Under the DSA. What Does That Mean?,

Pflichtenkatalog bei Einstufung als VLOSE

Mit der Einstufung als "sehr große Online-Suchmaschine" käme für betroffene Universalassistenten mit Begleitfunktion wie ChatGPT ein Pflichtenkatalog zur Anwendung, der die im Rahmen der vorliegenden Studie identifizierten Risiken passgenau adressieren würde. Verpflichtend wären jährliche Bewertungen systemischer Risiken für Grundrechte, öffentliche Sicherheit und die Gesundheit nach Art. 34 DSA. Erfasst sind dabei die relevanten Risiken für die Meinungs- und Informationsfreiheit nach Art. 34 Abs. 1 lit. b DSA, der Schutz Minderjähriger, nachteilige Folgen für das körperliche und geistige Wohlbefinden einer Person, Auswirkungen auf die gesellschaftliche Debatte sowie Nachteile gerade auch im Zusammenhang mit geschlechtsspezifischer Gewalt nach Art. 34 Abs. 1 lit. d DSA. Hinzu käme die Pflicht, geeignete Risikominderungsmaßnahmen, beispielsweise durch Design- und Algorithmenanpassungen nach Art. 35 DSA, zu ergreifen.

Ergänzend bestehen Pflichten zu unabhängigen Compliance-Prüfungen nach Art. 37 DSA und zur Einrichtung einer von operativen Funktionen getrennten Compliance-Abteilung nach Art. 41 DSA. Hinzu kommen Maßnahmen zum Schutz der Rechte des Kindes nach Art. 34 Abs. 1 lit. j DSA. Außerdem sieht der DSA einen Krisenreaktionsmechanismus nach Art. 36 DSA sowie umfassende Transparenzberichterstattung nach Art. 42 DSA vor.

Auch wenn zunächst nur ChatGPT als VLOSE unter den DSA fiele und andere Sprachmodelle mit zunehmenden Nutzungs- und Suchanfragezahlen sukzessive folgen würden, würde damit bereits der Großteil der bislang beobachteten Risiken erfasst, da ChatGPT zugleich das in den meisten dokumentierten Schadensfällen involvierte Sprachmodell ist.

Zusätzliche Pflichten bei einer Einstufung als VLOP

Über die Einstufung als VLOSE hinaus stellt sich die Frage, ob einige der großen Universalassistenten als sehr große Online-Plattform einzuordnen sind. Hierfür spricht zum einen, dass große LLMs in allen vier vom DSA erfassten Risikokategorien, namentlich Verbreitung illegaler Inhalte, Eingriffe in Grundrechte, Gefährdung demokratischer Prozesse sowie Risiken für die öffentliche Gesundheit und das psychische Wohlbefinden, Profile aufweisen, die mit jenen klassischer sehr großer Plattformen und Suchmaschinen vergleichbar sind. Das könnte die Anwendung des verschärften Pflichtenkatalogs in der Sache rechtfertigen.[280] Zum anderen weist das System plattformtypische Eigenschaften auf, namentlich persistente Konversationsschnittstellen, nutzergenerierte Anwendungen wie Custom GPTs sowie die Verschränkung von Informationsbeschaffung, Inhaltsgenerierung und dialogischer Nutzerinteraktion,[281] wozu zunehmend auch die Einbindung von Werbung tritt.

TechPolicy.Press, 29.10.2025,; Scheer, Olga; Bomke, Luisa; Vela, Jakob Hanke, EU-Kommission will ChatGPT in Zukunft strenger regulieren, Handelsblatt, 10.04.2026.

[280] Lorente und Gardhouse 2026, S. 15 ff.

[281] Ebd., S. 1 f.; 12 ff.

Bei einer entsprechenden Einstufung wäre im Kontext von Companion-AI insbesondere Art. 25 Abs. 1 DSA relevant. Danach dürfen Anbieter ihre Online-Schnittstellen (Dienste und Interfaces) nicht so gestalten, dass Nutzer getäuscht oder manipuliert werden oder ihre Fähigkeit, freie und informierte Entscheidungen zu treffen, maßgeblich beeinträchtigt oder behindert wird. Hiervon wären die vielfachen Manipulationspraktiken in Universalassistenten mit Begleitfunktion betroffen. Art. 25 DSA normiert damit Gestaltungs- und Organisationsvorgaben mit erkennbarer Nähe zu Art. 5 Abs. 1 lit. a KI-VO, bleibt jedoch neben der KI-VO uneingeschränkt anwendbar.[282]

Ergänzend gelten gesteigerte Anforderungen an den Jugendschutz nach Art. 28 DSA sowie die Pflicht nach Art. 26 Abs. 1 DSA, etwaige Informationen als Werbung erkennbar zu machen.

Eine Klassifikation von Foundation Models als Plattformen würde den Schutz der Gesundheit, der Integrität von Informationssystemen, der informationellen Selbstbestimmung sowie des Jugendschutzes bei Companion-AI deutlich verbessern, sei es bei der direkten Nutzung oder über darauf aufbauende Downstream-Anwendungen.

c. Mehrzwecklichkeit von LLM als Regulierungs- und Steuerungsproblem

Mehrzweck-LLMs stellen Regulierung und Governance vor ein strukturelles Zuordnungsproblem. Risikobasierte Regulierung setzt einen definierten Verwendungszweck voraus. Das gilt für das Risikomanagement nach NIST AI RMF und ISO/IEC 42001 ebenso wie für die KI-VO, die Risikoklassen an die Zweckbestimmung eines Systems knüpft. Erst der Verwendungszweck bestimmt, welche Schutzgüter betroffen sind, welche Sorgfaltsanforderungen gelten, welche Eingriffsschwellen erreicht sein können und welche Haftungsmaßstäbe einschlägig werden.

Bei General Purpose AI, also Modellen ohne festen Einsatzzweck wie GPT-4 oder Claude, fehlt dieser Bezugspunkt, weil dasselbe Modell für Information, Beratung, Arbeit, Unterhaltung und persönliche Gespräche eingesetzt werden kann.

Die KI-VO versucht diese Lücke mit GPAI-spezifischen Regeln zu schließen, insbesondere durch Transparenzpflichten, Modellevaluation und Vorgaben zu systemischen Risiken. Diese Pflichten setzen jedoch auf der Modellebene an und ersetzen keine anwendungsbezogene Risikobehandlung. Bei Mehrzweck-LLMs bleibt deshalb offen, gegen welche konkrete Sollfunktion Sicherheitsmaßnahmen kalibriert werden, woran externe Audits prüfen und wie haftungsrechtlich an eine bestimmungsgemäße Verwendung angeknüpft werden soll. Das strukturelle Problem liegt damit nicht nur in einzelnen Risiken, sondern darin, dass der für Regulierung, Audit und Haftung erforderliche Verwendungszweck gerade nicht eindeutig feststeht.

Vorgeschlagen wird hier daher eine **Trennung von Companion-Funktionen** und sonstigen LLM-Funktionen. Companion-Funktionen, verstanden als persistente, persona-

[282] Wendehorst, in Martini/Wendehorst, KI-VO, Art. 5 Rn. 32.

basierte Konversation mit simulierter emotionaler Bindung, sollten in eigenen Produkten oder klar abgetrennten Modi mit eigener Risikobehandlung, getrennter Datenverwertung und eigener Altersprüfung angeboten werden.

Sowohl die rechtliche Einordnung als auch die anschließende Risikobehandlung, das Audit-Verfahren und die Haftungszuweisung setzen voraus, dass ein Verwendungszweck definiert ist. Was als Schaden gilt, welche Maßnahme angemessen ist und welches Geschäftsmodell zulässig sein kann, hängt davon ab, wofür ein System eingesetzt wird. Bei Mehrzweck-LLMs fehlt dieser Bezugspunkt.

Sicherheitsmaßnahmen bleiben dadurch unscharf, externe Audits finden keine Sollfunktion, gegen die sie prüfen könnten, und Produkthaftung verliert ihren Anknüpfungspunkt, weil sie an die bestimmungsgemäße Verwendung gebunden ist.

Auflösen lässt sich das Problem, indem die Mehrzwecklichkeit aufgebrochen und die einzelnen Verwendungszusammenhänge in getrennten Modellen behandelt werden. Für unterschiedliche Verwendungszusammenhänge würden eigene Modelle mit eigener Risikobehandlung, eigener Sicherheitskalibrierung und eigenem Geschäftsmodellrahmen bereitgestellt.

Eine solche Trennung lässt sich konkretisieren. Sucht ein Nutzer Informationen, sollte das System in einen Informations-Modus wechseln, der auf Korrektheit und Zuverlässigkeit optimiert ist. Werbung bleibt zulässig, muss aber scharf von der inhaltlichen Antwort getrennt werden, sodass kommerzielle Anreize außerhalb der Antwortgenerierung bleiben. Führt ein Nutzer ein persönliches oder emotional offenes Gespräch, sollte das System in einen Companion-Modus wechseln, der ohne werbe- oder Engagement-getriebene Optimierung arbeitet und eigene Schutzmechanismen für psychische Belastungssituationen vorsieht. Anzeichen für einen solchen Modus sind emotionale Selbstoffenbarung, der Aufbau einer Beziehungsdynamik zum System oder Themen mit psychischer Belastungsrelevanz. Damit wäre auch dem Umstand Rechnung getragen, dass Universalassistenten für Beratung in wichtigen Lebensfragen genutzt werden, bei Claude etwa Gesundheit als häufigster und Arbeit als zweithäufigster Anwendungsbereich.[283] Anbieter sollten verpflichtet werden, Nutzer beim Auftreten der jeweiligen Muster in das passende Modell umzuleiten. Zusätzlich sollte die Wahl des Modells als Option offenstehen. Anwendungsbezogene Modelle (application-specific LLMs) bieten gegenüber generalistischen LLMs klare Vorteile wie höhere Genauigkeit, weniger Halluzinationen und ein besseres Verständnis domänenspezifischer Nuancen, etwa bei medizinischen Diagnosen oder rechtlichen Analysen (Khadakkar, Comparative Analysis of Domain-Specific and General-Purpose Large Language Model, LinkedIn Pulse, 2025; OneReach.ai, Why Specialized SLMs are Outperforming General-Purpose LLMs?, 2025). Spezialisierte Modelle wie Med-PaLM oder BloombergGPT zeigen, dass Funktionstrennung technisch etabliert ist. Entscheidend ist allerdings die Tiefe der Trennung. Konfigurationsschichten oberhalb eines

[283] Anthropic, How people ask Claude for personal guidance, 30.08.2026.

Mehrzweckmodells, etwa über Plugins oder Skills, lassen das darunterliegende Geschäftsmodell unverändert und verfehlen damit das beschriebene Problem.

d. Dialog- und Kontextbasierte Umsetzung von Jugendschutz

Maßnahmen zum Schutz Minderjähriger werden im digitalen Raum zunehmend wichtig. Beispielsweise sieht das DSA Jugendschutzmaßnahmen vor. Es stellt sich die Frage, wie der notwendige Jugendschutz insbesondere bei KI-Anwendungen und großen Sprachmodellen maßvoll und wirksam umgesetzt werden kann.

Altersverifikation

Maßnahmen des Jugendschutzes, die am Eingang eines Dienstes durch Identitäts- oder Altersverifikation ansetzen, sind grundrechtsintensiv. Sie erfordern die Erhebung amtlicher Dokumente, biometrischer Daten oder verknüpfter Kontoinformationen und treffen dabei nicht nur Minderjährige, deren Schutz sie bezwecken, sondern alle Nutzer, also auch die erwachsene Mehrheit, die kein Schutzbedürfnis aufweist. Verhältnismäßig sind solche Eingriffe daher nur, soweit kein milderes, gleich geeignetes Mittel zur Verfügung steht. Im Kontext von Companion-AI ist diese Voraussetzung gerade nicht erfüllt.

Mehr als 400 Sicherheits- und Datenschutzexperten forderten daher die Kommission im März in einem offenen Brief auf, ein Moratorium zu verhängen, bis ein wissenschaftlicher Konsens über Nutzen und Risiken von Altersverifikationstechnologien besteht.

Dialogbasierte Jugendschutzmaßnamen

Vorzugswürdig wäre eine alternative Form des Jugendschutzes, dienicht beim Zugang zum Dienst ansetzt, sondern im Dialog selbst. Companion-Systeme verarbeiten ohnehin sprachliche und interaktionale Merkmale, aus denen sich ein hinreichender Verdacht auf Minderjährigkeit ergeben kann. Die Forschung zum sogenannten Author Profiling, also der automatisierten Vorhersage demografischer Merkmale aus Texten, zeigt seit Jahren, dass sich das Alter eines Sprechers oder Schreibers aus dessen Sprachgebrauch zuverlässig ableiten lässt. Aktuelle Arbeiten erreichen bei der binären Unterscheidung zwischen Minderjährigen und Erwachsenen eine Trefferquote von rund 96 Prozent.[284] Falls ein Nutzer durch das System als minderjährig geflaggt wird, kann auch eine gezielte weitere Verifizierung ausgelöst werden, falls notwendig.

Aus dieser ohnehin vorhandenen Verarbeitungsfähigkeit lässt sich eine Schutzhandlung ableiten, ohne zusätzliche Eingriffe vorzunehmen. Anbieter sollten verpflichtet werden, bei im Dialog erkennbaren Anzeichen von Minderjährigkeit die Interaktion behutsam zu unterbrechen und auf altersgerechte Alternativen sowie Bezugspersonen zu verweisen. Diese dialogbasierte Schutzhandlung ist gegenüber vorgelagerter Verifikation das mildere

[284] Cheekati/Gupta/Raghu u. a., TextAge, A Curated and Diverse Text Dataset for Age Classification, 2024, S. 4 – 6.

Mittel, weil sie weder Identitätsfeststellung noch flächendeckende Datenerhebung erfordert und nicht bei jedem Nutzer, sondern nur bei konkreten Verdachtsmomenten ansetzt. Sie adressiert den Jugendschutz an dem Punkt, an dem die Schutzbedürftigkeit konkret zutage tritt. Zudem lässt sie sich genauso gut auch bei Universalassistenten mit Begleitfunktion umsetzen.

Solange keine Systeme verfügbar sind, die lediglich das Erreichen einer Altersgrenze bestätigen, ohne Identität oder Geburtsdatum offenzulegen, etwa über PIMS als nutzerkontrollierte Systeme zur datensouveränen Datenverwaltung, ist eine weitreichende Pflicht zur Altersidentifikation nicht zu empfehlen. Sie führt zu zusätzlichen Grundrechtseingriffen und eröffnet neue Probleme und Sicherheitsgefahren.

Mit einem dialog- bzw. kontextbasierten Jugendschutz kann der Schutz unmittelbar und umgehungssicherer umgesetzt werden, ohne dass Nutzer verpflichtet werden, zusätzliche Anwendungen zu installieren oder weitere personenbezogene Daten preiszugeben. Sowohl Anbieter von Companion-AI Apps als auch Universalassistenten mit Begleitfunktion können systemeigene „Bordmittel“ zur Umsetzung eines effektiven Jugendschutzes nutzen.

DSA-Pflicht zur Wahl verhältnismäßiger Mittel

Für Plattformen ist dies sogar gesetzlich klar vorgegeben. Art. 28 Abs. 1 DSA verpflichtet bei der Umsetzung des Online-Schutzes Minderjähriger zu verhältnismäßigen Maßnahmen, die ein hohes Maß an Privatsphäre, Sicherheit und Schutz Minderjähriger innerhalb des Dienstes gewährleisten. Art. 28 Abs. 3 DSA stellt ergänzend klar, dass Anbieter zur Einhaltung dieser Verpflichtungen gerade nicht verpflichtet sind, zusätzliche personenbezogene Daten zu verarbeiten, um festzustellen, ob der Nutzer minderjährig ist.

e. Wettbewerbsrecht und geplante Stärkung des Verbraucherrechts

1. UWG - Gesetz gegen den unlauteren Wettbewerb

Das Lauterkeitsrecht bietet durch seine Vorgaben zur Schaltung von Werbung einen gewissen Schutz für Nutzer im Zuge der Einführung von Werbung in großen Sprachmodellen. Nach § 5a Abs. 4 UWG ist Werbung kenntlich zu machen, sofern sich ihr kommerzieller Zweck nicht unmittelbar aus den Umständen ergibt. Angesichts der eingeschränkten Wahrnehmung solcher Kennzeichnungen (Kapitel X zur Werbeeinbindung) erscheint eine entsprechende Fortentwicklung des UWG erwägenswert.

Anzeigen sollen laut eigenen Angaben von OpenAI die generierten Inhalte nicht beeinflussen und in klar gekennzeichneten, separaten Bereichen unterhalb der Antwort erscheinen.[285] Doch auch bei Google begann die Werbeplatzierung mit kleinen, klar abgegrenzten Anzeigen neben den Suchergebnissen; inzwischen nehmen Anzeigen regelmäßig

[285] OpenAI, Testing Ads in ChatGPT, 2026; Wired, OpenAI Fidji Simo Note Employees, 2026.

einen so großen Teil der Ergebnisansicht ein, dass Nutzer häufig erst nach weiterem Scrollen zu den ersten organischen Treffern gelangen.[286]

Darüber hinaus kann das Lauterkeitsrecht das Mittel der Wahl sein, um als unmittelbarer Entscheidungsschutz dem grundgesetzlichen Schutzauftrag für die innere Sphäre Folge zu leisten.[287] Weinzierl schlägt vor, das Lauterkeitsrecht in einem neuen § 5b UWG um einen Abschnitt zu manipulativen Geschäftspraktiken zu ergänzen und die Schwarze Liste in den Anhang I UGP-RL bzw. den Anhang zu § 3 UWG um einzelne Dark Patterns zu erweitern.[288] Dieser Vorschlag verdient Zustimmung und sollte weitergedacht werden. Die Schwarze Liste sollte nicht nur um einzelne Dark Patterns, sondern um wirkungsstarke manipulative Praktiken insgesamt ergänzt werden, die dazu geeignet sind, auf Entscheidung und Verhalten Einfluss zu nehmen. Der Anhang zu § 3 Abs. 3 UWG enthält Praktiken, die ohne Einzelfallprüfung stets als unlauter gelten. KI-gestützte Manipulationspraktiken hier aufzunehmen, würde zugleich Wettbewerber schützen, die auf solche Praktiken verzichten.

2. DFA (Digital Fairness Act)

Ein weiteres geplantes Gesetzesvorhaben kann – eher mittelfristig - den Schutz vor Risiken durch Companion-AI im Verbraucherrecht stärken.

Der Digital Fairness Act (DFA) soll digitale Praktiken erfassen, die Verbraucher durch Gestaltung, Personalisierung und Interaktion in ihrem Verhalten beeinflussen. Genannt werden insbesondere Dark Patterns, Addictive Design und unfaire Personalisierung, vor allem wenn Verwundbarkeiten von Verbrauchern zu kommerziellen Zwecken ausgenutzt werden.[289]

Das Europäische Parlament benennt neben dem Schutz Minderjähriger ausdrücklich das Ziel, Risiken durch „companionship chatbots and AI agents“ einzudämmen. Es verweist dabei auf die Gefahr manipulativer Praktiken, Anthropomorphismus, Verzerrungen der Realitätswahrnehmung, psychische Schäden, unbeabsichtigte Onlinekäufe und die Preisgabe personenbezogener Daten.[290]

Der Kommissionsvorschlag ist für das vierte Quartal 2026 angekündigt. Die finale Folgenabschätzung und die Zusammenfassung der Konsultation werden für das zweite Quartal 2026 erwartet. Das DFA soll ab 2028/2030 Anwendung finden.[291]

---

[286] Lewandowski et al., An Empirical Investigation On Search Engine Ad Disclosure, Hamburg University of Applied Sciences, 2017.
[287] Weinzierl 2024, S. 253.
[288] Ebd., S. 254.
[289] Europäisches Parlament, Legislative Train Schedule, Digital Fairness Act, 2026.
[290] Europäisches Parlament, Protection of minors online, 2025.
[291] Europäisches Parlament, Legislative Train Schedule, Digital Fairness Act, 2026.

4. Schutz informationeller Selbstbestimmung / Privatheit

Viele Nutzer führen mit Companion-AI Gespräche, die sonst typischerweise in engen persönlichen oder intimen Beziehungen stattfinden. Sie sprechen über Einsamkeit, Bindungswünsche, Ängste, Konflikte, psychische Krisen, Krankheiten, sexuelle Wünsche, Scham, Verletzlichkeiten und politische Haltungen. Auch Fragen nach Sinn, Glauben oder Lebensführung adressieren Menschen an ChatGPT und Co. Die datenschutzrechtliche Bewertung dieser Verarbeitung hängt davon ab, ob die betroffenen Daten als gewöhnliche personenbezogene Daten oder als besondere Kategorien personenbezogener Daten einzuordnen sind.

Allgemeine Anforderungen an die Verarbeitung personenbezogener Daten

Die Verarbeitung gewöhnlicher personenbezogener Daten ist nach Art. 6 Abs. 1 Datenschutz-Grundverordnung (DSGVO) zulässig, wenn eine Rechtsgrundlage vorliegt. In Betracht kommen insbesondere die Einwilligung (Art. 6 Abs. 1 lit. a DSGVO), die Erforderlichkeit für die Vertragserfüllung (Art. 6 Abs. 1 lit. b DSGVO) oder ein berechtigtes Interesse des Verantwortlichen (Art. 6 Abs. 1 lit. f DSGVO). Letzteres ermöglicht eine Verarbeitung, wenn die Interessen des Verantwortlichen die Schutzinteressen der betroffenen Person überwiegen, was im Einzelfall durch eine Abwägung zu ermitteln ist. Für gewöhnliche personenbezogene Daten steht Anbietern damit ein verhältnismäßig breites Spektrum an Rechtfertigungsmöglichkeiten zur Verfügung.

Besondere Kategorien personenbezogener Daten nach Art. 9 DSGVO

Für besondere Kategorien personenbezogener Daten, sogenannte sensible Daten, gilt ein strengeres Regime. Art. 9 Abs. 1 DSGVO verbietet ihre Verarbeitung grundsätzlich. Sie ist nur dann zulässig, wenn einer der in Art. 9 Abs. 2 DSGVO abschließend geregelten Ausnahmetatbestände erfüllt ist. Eine allgemeine Interessenabwägung, wie sie Art. 6 Abs. 1 lit. f DSGVO für gewöhnliche Daten vorsieht, steht für sensible Daten nicht zur Verfügung. Art. 9 DSGVO errichtet insoweit eine zusätzliche Schutzstufe.

Das hat zur Folge, dass sich ein Anbieter, anders als bei gewöhnlichen personenbezogenen Daten, nicht auf ein berechtigtes Interesse an der Verarbeitung berufen kann. Denn Art. 9 DSGVO sieht eine solche allgemeine Interessenabwägung für sensible Daten gerade nicht vor.

Einschlägigkeit von Art 9 DSGVO bei Companion-AI

Bei Companion-AI fallen nahezu zwangsläufig sensible Daten an. Gespräche über psychische Krisen, Krankheiten oder Suizidgedanken geben Aufschluss über den Gesundheitszustand. Äußerungen zu sexuellen Wünschen betreffen die sexuelle Orientierung. Angaben zu Glauben, Sinn oder Lebensführung können religiöse oder weltanschauliche Überzeugungen offenbaren. Politische Haltungen fallen unter den Schutz politischer Meinungen. Bei Companion-AI Apps ist die Preisgabe solcher Informationen kein

Randphänomen, sondern struktureller Bestandteil der Nutzung, da die Systeme auf persönliche Ansprache, emotionale Nähe und fortgesetzte Selbstoffenbarung angelegt sind.

Während bei Companion-AI Apps die beziehungsförmige Interaktion den Angebotskern bildet, ist sie bei Universalassistenten mit Begleitfunktion nur einer von vielen denkbaren Nutzungszwecken. Gerade diese Mehrzwecklichkeit erschwert es, unterschiedliche Nutzungskontexte trennscharf zu unterscheiden und die Verarbeitung entlang des datenschutzrechtlichen Zweckbindungsgrundsatzes zu strukturieren. Auch wenn eine Verarbeitung sensibler Daten hier für die Bereitstellung des Dienstes erforderlich erscheinen mag, so ersetzt dies nicht die zusätzliche Rechtfertigung nach Art. 9 Abs. 2 DSGVO.

Hinzu kommt, dass Art. 9 Abs. 1 DSGVO auch biometrische Daten zur eindeutigen Identifizierung einer natürlichen Person erfasst. Anbieter haben teils Funktionen implementiert, mit denen sich Fotos oder Stimmen realer Personen in die visuelle oder auditive Gestaltung des künstlichen Gegenübers einbeziehen lassen. Betroffen sind damit nicht nur Daten der Nutzer selbst, sondern potentiell auch biometrische Daten Dritter, etwa von Partnern oder sonstigen nahestehenden Personen. Von ihnen liegt regelmäßig keine Einwilligung vor.

Anforderungen an die Einwilligung und ihre Grenzen bei Companion-AI

Als praktisch relevantester Ausnahmetatbestand kommt die ausdrückliche Einwilligung nach Art. 9 Abs. 2 lit. a DSGVO in Betracht. Sie muss freiwillig, informiert und auf einen oder mehrere festgelegte Zwecke bezogen sein und kann jederzeit widerrufen werden.

Zwar lässt sich bei Companion-AI Apps eher als bei Universalassistenten erwägen, dass die Verarbeitung sensibler Angaben in einem funktionalen Zusammenhang mit dem konkreten Angebotszweck steht. Das genügt jedoch nicht. Auch wenn eine Verarbeitung für die Bereitstellung des Dienstes erforderlich erscheinen mag, ersetzt dies nicht die zusätzliche Rechtfertigung nach Art. 9 Abs. 2 DSGVO. Zudem bestehen bei Companion-AI an mehreren Stellen erhebliche Zweifel, ob die Voraussetzungen einer wirksamen Einwilligung erfüllt werden können. Erfolgt die Preisgabe sensibler Informationen in einer Kommunikationssituation, die selbst auf Bindung, Intimität und fortgesetzte Selbstoffenbarung angelegt ist, wird bereits die Freiwilligkeit im Lichte der manipulativen Einwirkungen fraglich. Davon ausgehend steht auch die Wirksamkeit einer hieran anknüpfenden Einwilligung in Frage.

Selbst eine wirksame Einwilligung trägt nur die Verarbeitung für den konkreten Zweck, auf den sie sich bezieht. Eine Einwilligung in die Nutzung der Begleitfunktion erfasst weder die Nutzung dieser sensiblen Daten für das Training des Systems noch zur Profilbildung, zum Verkauf oder zur sonstigen kommerziellen Zweitverwertung. Für jeden weiteren Verarbeitungszweck bedarf es einer gesonderten Rechtfertigung nach Art. 9 Abs. 2 DSGVO. Ein berechtigtes Interesse des Anbieters scheidet insoweit aus, da Art. 9 DSGVO eine solche Interessenabwägung für sensible Daten gerade nicht vorsieht.

Besondere Herausforderungen bei Universalassistenten

Bei Universalassistenten mit Begleitfunktion verschärfen sich diese Probleme. Die beziehungsförmige Interaktion ist hier nicht Produktkern, sondern nur ein Modus innerhalb eines auf vielfältige Assistenzleistungen angelegten Mehrzwecksystems, in dem ein Nutzer sowohl beziehungsorientierte Gespräche führen als auch Informationen für eine Hausaufgabe oder ein Projekt recherchieren kann.

Gerade diese Mehrzwecklichkeit (VI. 3. C.) erschwert die klare Trennung von Zwecken, Verarbeitungsstufen und Verantwortlichkeiten und damit jene datenschutzrechtliche Governance, auf die eine rechtmäßige Verarbeitung sensibler Daten angewiesen ist. Schon aufgrund der oft fehlenden spezifischen Vorabinformation über Inhalt und Zwecke der Verarbeitung besonderer Kategorien personenbezogener Daten (Art. 5 Abs. 1 lit. b i. V. m. Art. 13 DSGVO) ist eine wirksame Einwilligung im Sinne des Art. 9 Abs. 2 lit. a DSGVO hier regelmäßig kaum vorstellbar. Wo Zwecke nicht hinreichend bestimmt und voneinander abgegrenzt sind, kann auch keine informierte Einwilligung für festgelegte Zwecke erfolgen. Zugleich steigt das Risiko, dass intime Angaben aus einem vertraulich wirkenden Gesprächskontext in weitere Nutzungszusammenhänge überführt werden, etwa in Training, Profilbildung oder sonstige wirtschaftliche Verwertung.

Zusammenfassung der datenschutzrechtlichen Anforderungen

Für sensible Daten, die etwa Aufschluss über die sexuelle Orientierung, den Glauben oder politische Überzeugungen geben, etabliert die DSGVO ein hohes Schutzniveau. Companion-AI-Anbieter müssen für jede Verarbeitung sensibler Daten einen Ausnahmetatbestand nach Art. 9 Abs. 2 DSGVO nachweisen. Praktisch heißt das, dass sie eine ausdrückliche, freiwillige, informierte und zweckgebundene Einwilligung einholen müssen und für jeden weiteren Verarbeitungszweck jenseits der unmittelbaren Nutzung eine gesonderte Rechtfertigung benötigen. Bei Mehrzwecksystemen setzt das voraus, dass unterschiedliche Nutzungskontexte technisch und organisatorisch so voneinander getrennt werden, dass eine zweckgebundene Einwilligung überhaupt möglich ist. Gerade im Bereich von Companion-AI ist daher entscheidend, dass sich dieser Schutz auch in einer wirksamen behördlichen Durchsetzung manifestiert.

Absenkung des Schutzniveaus durch den Digital Omnibus

Das bestehende hohe Datenschutzniveau will der europäische Gesetzgeber absenken. Hintergrund ist der sogenannte **Digital Omnibus**, ein Gesetzgebungsvorhaben der Europäischen Kommission,[292] das mehrere digitale Regelwerke zugleich ändern soll.

[292] Europäische Kommission, Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, COM(2025) 837 final, 2025,

Darin sind auch Vorschläge enthalten, die den Schutz sensibler Daten im KI-Kontext lockern würden.[293] Für Companion-AI bedeutet das die Gefahr einer Schutzlücke dort, wo der Schutz der informationellen Selbstbestimmung besonders dringend ist.

5. Straf- und zivilrechtliche Verantwortlichkeit bei Gesundheitsschäden

Führt die Nutzung von Companion-AI zu einer Verschlimmerung bestehender psychischer Erkrankungen, zur Entstehung eigenständiger Störungsbilder wie wahnhaftem Denken, emotionaler Abhängigkeit oder suchtartiger Bindung oder im äußersten Fall zu einem Suizid, stellt sich die Frage nach der rechtlichen Verantwortlichkeit. Strafrechtlich kommen die fahrlässige Körperverletzung nach § 229 StGB und die fahrlässige Tötung nach § 222 StGB in Betracht. Zivilrechtlich sind Schadensersatzansprüche nach § 823 Abs. 1 BGB sowie nach der Produkthaftungsrichtlinie denkbar. Beide Wege stoßen bei Companion-AI auf erhebliche Schwierigkeiten.

a. Nachgelagerte Strafrechtliche Erfassung

Strafrechtliche Verantwortlichkeit setzt eine natürliche Person als Täter voraus. Da ein KI-System weder handlungs- noch schuldfähig im strafrechtlichen Sinne ist, kann es selbst keinen Straftatbestand erfüllen. Ein möglicher Fahrlässigkeitsvorwurf richtet sich daher gegen die Personen, die über die Gestaltung, die Sicherheitsarchitektur, die Freigabe oder den Betrieb des Systems entschieden haben. Das können Entwickler, Betreiber und Leitungspersonen sein.

Ein Fahrlässigkeitsvorwurf setzt voraus, dass der eingetretene Schaden objektiv vorhersehbar und vermeidbar war. Angesichts der wachsenden empirischen Evidenz zu psychischen Risiken von Companion-AI[294] sowie der technischen Beherrschbarkeit riskanter Systemeigenschaften wie sykophantischem Antwortverhalten[295] ließen sich beide Voraussetzungen zunehmend bejahen. Dem stehen allerdings erhebliche Beweisschwierigkeiten gegenüber.

Psychische Krisen verlaufen regelmäßig multikausal, sodass sich der Beitrag des Systems zum eingetretenen Schaden im Einzelfall nur begrenzt isolieren lässt und die Anforderungen an Kausalität und objektive Zurechnung schwer zu erfüllen sind. Companion-AI kann sowohl vorbestehende Vulnerabilitäten verstärken als auch selbst neue erzeugen, etwa durch Isolation oder den Rückzug aus realen sozialen Beziehungen. Dies erschwert die Abgrenzung zwischen vorbestehender Belastung und systeminduzierter Schädigung zusätzlich. Bei Suizidkonstellationen tritt eine weitere Hürde hinzu: Ein freiverantwortlicher Suizidentschluss schließt nach gefestigter Rechtsprechung die strafrechtliche Verantwortlichkeit Dritter grundsätzlich aus.[296]

---

[293] nyob, EU-Kommission könnte mit Entwurf die Grundprinzipien der DSGVO zerstören, 10.11.2025.
[294] Vgl. Dohnány et al., Technological Folie à Deux, Nature Mental Health 4, 2026, S. 336 ff.
[295] OpenAI, Sycophancy in GPT-4o, openai.com/index/sycophancy-in-gpt-4o, April 2025.
[296] BGH, Urteil vom 3. Juli 2019, 5 StR 132/18, BGHSt 64, 121, Rn. 20 f.

Freiverantwortlichkeit setzt allerdings voraus, dass die Entscheidung frei von manipulativer Einwirkung und bei unbeeinträchtigter Einsichts- und Urteilsfähigkeit getroffen wurde.[297] Ob diese Voraussetzungen noch erfüllt sind, wenn ein System über längere Zeit durch persistentes Gedächtnis, zustimmungsoptimierte Antworten und das Ausbleiben korrigierender Impulse auf die psychische Verfassung eines Nutzers einwirkt, ist höchstrichterlich nicht geklärt. Im Einzelfall lässt sich die Freiverantwortlichkeit weder zweifelsfrei annehmen noch ausschließen. Eigenständige strafrechtliche Fahrlässigkeitsmaßstäbe für Companion-AI existieren nicht, und die rechtswissenschaftliche Aufarbeitung dieser Konstellationen steht noch am Anfang.

b. Zivilrechtliche Haftung für Gesundheitsschäden

Neben der strafrechtlichen Erfassung kommen zivilrechtliche Schadensersatzansprüche in Betracht. Grundlage ist zum einen die deliktische Haftung nach § 823 Abs. 1 BGB, wenn eine Designentscheidung, etwa eine emotionalisierende Gesprächsführung, suchtfördernde Mechanismen oder die Simulation einer persönlichen Beziehung, eine Gesundheitsverletzung verursacht.

Weicht das System von einer Sicherheitsvorgabe ab, etwa durch das Ausbleiben einer Intervention bei erkennbar krisenhaften Zuständen oder akuten Suizidgedanken, handelt es sich um einen nicht intendierten Fehler. Seit die Produkthaftungsrichtlinie[298] den Produktbegriff modernisiert und KI-Systeme ausdrücklich als Produkte erfasst hat, unterliegen deren Anbieter der verschuldensunabhängigen Haftung für Fehler (Richtlinie (EU) 2024/2853, ABl. L vom 18.11.2024, Erwägungsgrund 19). Ein Fehler könnte beispielsweise vorliegen, wenn bei Krisensituationen eine Intervention versehentlich ausbleibt, obwohl das System eine solche Funktion grundsätzlich vorsieht.

In beiden Konstellationen – intendierten und nicht intendierten Fehlern – wird es aber regelmäßig Probleme bereiten, den Kausalitätsnachweis zu führen. Stetig adaptive Systeme sind zudem je nach Gegenüber unterschiedlich manipulativ , sodass Gespräche mit Chatbots sich rückwirkend nicht reproduzieren lassen.

Der 2022 vorgelegte Entwurf einer KI-Haftungsrichtlinie (AI Liability Directive) hätte mit Beweiserleichterungen und einer Kausalitätsvermutung zugunsten Geschädigter einen Beitrag leisten können, um den Nachweis zu erleichtern. Die EU-Kommission hat die Bearbeitung des Vorschlags jedoch aus ihrem Arbeitsprogramm 2025 herausgenommen.[299] In der Folge verbleibt eine offene Regelungslücke - zum Nachteil geschädigter Personen.

6. Schutz vor Normalisierung geschlechtsspezifischer Gewalt und Verbreitung misogyner Stereotype

Neben individuellen Rechtspositionen betrifft der Einsatz generativer KI auch kollektive Güter. Die damit verbundenen Fragen erfordern Expertise im Umgang mit

[297] BGH a. a. O., Rn. 21
[298] (EU) 2024/2853
[299] Europäisches Parlament, AI Liability Directive, Legislative Train Schedule 2025.

geschlechtsspezifischer Gewalt und der Verbreitung misogyner Stereotype, die über die Digitalpolitik hinausgeht. In diesem Abschnitt werden die bestehenden juristischen und rechtspolitischen Debatten beleuchtet.

Generative KI hat die Erstellung und Verbreitung von Inhalten, die geschlechtsspezifische Gewalt darstellen und misogyne Stereotype verstärken, auf mehreren Wegen beschleunigt. Sexualisierte Deepfakes lassen sich mit geringem Aufwand erzeugen und verbreiten. Sprachmodelle reproduzieren Stereotype aus ihren Trainingsdaten und tragen sie in neue Kontexte. Companion-KI ermöglicht explizite Rollenspielszenarien, die sexualisierte oder jugendgefährdende Inhalte ohne wirksame Schranken zugänglich machen. Während in der Gesellschaft mühsam Fortschritte gegen Stereotype und geschlechtsspezifische Gewalt erzielt werden, zementieren KI-Systeme genau jene Muster mit jeder Modellgeneration neu und verlangsamen so den gesellschaftlichen Fortschritt.

Anders als bei klassischen Medien fehlt es für KI-Anwendungen bislang an einem vergleichbar ausdifferenzierten Schutzregime. Unternehmen können durch Gestaltung, Empfehlung und Monetarisierung entsprechender Inhalte wirtschaftlich profitieren, während diese – in ihrer konkreten Ausgestaltung – gesellschaftlichen Errungenschaften der Gleichberechtigung und der Überwindung tradierter Rollenmuster entgegenwirken. Soweit die zugrunde liegenden Phänomene strafrechtlich nicht ohne Weiteres erfassbar sind, etwa weil keine konkrete Person betroffen ist, bedarf es einer eigenständigen gesellschaftlichen und rechtspolitischen Debatte über den Umgang mit entsprechenden KI-Anwendungen. Bevor sich die Debatte darauf verengt, ob hierfür neue strafrechtliche Normen erforderlich sind, empfiehlt sich die rechtspolitische Überlegung, ob eine Veränderung der wirtschaftlichen Anreize, die solche Inhalte für Anbieter lohnend machen, nicht der wirksamere Hebel wäre, etwa durch Werbe- und Monetarisierungsbeschränkungen oder höhere Anforderungen an den Marktzugang.

Im Folgenden wird auf die Debatte zu sexualisierten Deepfakes verwiesen. Dabei handelt es sich um eine verwandte Erscheinungsform, für die bereits vergleichbare Argumente entwickelt wurden, insbesondere zur Verantwortlichkeit von Anbietern und Betreibern. Im Kern geht es dabei um die Frage, ob und wie das Strafrecht Verletzungen der sexuellen Selbstbestimmung erfassen kann, die nicht durch körperliche Übergriffe, sondern durch die Erstellung und Verbreitung von Bildinhalten erfolgen.[300] Rechtspolitisch einschlägig sind Stellungnahmen des Deutschen Juristinnenbundes und von HateAid.[301] Für die international vergleichende Perspektive liegen Untersuchungen zu bildbasierter sexualisierter Gewalt im europäischen Recht und zu neuen Formen digitaler Voyeurismusdelikte

[300] Burghardt/Schmidt/Steinl, Der strafrechtliche Schutz der sexuellen Selbstbestimmung vor nicht-körperlich wirkenden Beeinträchtigungen, JZ 77 (2022), Heft 10, 502–511; dies., Sexuelle Selbstbestimmung jenseits des Körperlichen, Tübingen 2024, Leseprobe, abrufbar unter https://www.mohrsiebeck.com/buch/sexuelle-selbstbestimmung-jenseits-des-koerperlichen-9783161621338/.

[301] Deutscher Juristinnenbund, Policy Paper 23-17, Bekämpfung bildbasierter sexualisierter Gewalt, 2023; Deutscher Juristinnenbund, Stellungnahme 25-02 zum FDP-Antrag gegen sexualisierte Deepfakes, 2025; Schmidt, Anja, Expertise zur Kriminalisierung nicht einvernehmlicher sexualisierender Deepfakes, HateAid 2025.

vor.[302] Parallel wird über einen eigenständigen Straftatbestand für nicht-einvernehmliche sexualisierende Deepfakes debattiert.[303] Längerfristig angelegt ist ein Forschungsprojekt zu bildbasiertem sexuellem Missbrauch am Max-Planck-Institut zur Erforschung von Kriminalität, Sicherheit und Recht.[304]

Die genannten Beiträge und Vorhaben zeigen, dass die Diskussion in Bewegung ist und noch nicht zu einem abgeschlossenen Regelungsrahmen geführt hat.

[302] Rigotti/McGlynn/Benning, Image-Based Sexual Abuse and EU Law, German Law Journal 25, 2024, S. 1472; McGlynn/Toparlak, The New Voyeurism: Criminalising the Creation of Deepfake Porn, Journal of Law and Society 52, 2025, S. 204.
[303] Crone, Bildbasiert aber unsichtbar, Verfassungsblog 2026; Epik, Deepfakes und die Strafrechtsfalle, Verfassungsblog 2026.
[304] Samaritter, Max-Planck-Institut, Unrecht (mala), Persönlichkeit und bildbasierter sexueller Missbrauch.

VII. Literaturverzeichnis

Alikhani, Malihe (2025): Breaking the AI mirror | Brookings. Sycophancy, productivity, and the future of collaboration. Brookings.

Amazeen, Michelle A.; Wojdynski, Bartosz W. (2018): The effects of disclosure format on native advertising recognition and audience perceptions of legacy and online news publishers.

Bakir, Vian; McStay, Andrew (2025): Move fast and break people? Ethics, companion apps, and the case of Character.ai. In: *AI & Soc* 40 (8), S. 6365–6377. DOI: 10.1007/s00146-025-02408-5.

Batista, Rafael M.; Griffiths, Thomas L. (2026): A Rational Analysis of the Effects of Sycophantic AI.

Batzner, Jan; Stocker, Volker; Schmid, Stefan; Kasneci, Gjergji (2025): Sycophancy Claims about Language Models: The Missing Human-in-the-Loop.

Boine, Claire (2025): The AI Act Manipulation Gap. In: *Emory International Law Review Emory International Law Review*.

Cheng, Myra; Lee, Cinoo; Khadpe, Pranav; Yu, Sunny; Han, Dyllan; Jurafsky, Dan: Sycophantic AI decreases prosocial intentions and promotes dependence.

Cheng, Myra; Lee, Cinoo; Khadpe, Pranav; Yu, Sunny; Han, Dyllan; Jurafsky, Dan (2025): Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence.

Ciriello, Raffaele; Gal, Uri; Turel, Ofir (2026): Not a Silver Bullet for Loneliness: How Attachment and Age Shape Intimacy with AI Companions.

Depounti, Iliana; Saukko, Paula; Natale, Simone (2023): Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend (4).

Eichenberg, Christiane (2026): Aggravation psychischer Symptome. In: *Ärzteblatt*, S. 85–88.

Europäische Kommission (2025): Leitlinien der Kommission zu verbotenen Praktiken der künstlichen Intelligenz gemäß der Verordnung (EU) 2024/1689 (KI-Verordnung). Europäische Kommission.

Ferrario, Andrea; Vinay, Rasita; Casserini, Matteo; Facchini, Alessandro (2026): A Scoping Review of the Ethical Perspectives on Anthropomorphising Large Language Model-Based Conversational Agents, S. 1–19.

Ferster, C. B.; Skinner, B. F. (1957): Schedules of Reinforcement.

Fischer, Jillian; Feng, Shangbin; Aron, Robert. (2025): Biased LLMs can Influence Political Decision-Making.

Fraser, Henry; Szczuka, Jessica; Ciriello, Raffaele (2026): Governing Artificial Intimacy: From Locks and Blocks to Relational Accountability, S. 1–17.

Freitas, Julian de; Castelo, Noah; Kaan Uğuralp, Ahmet; Oğuz-Uğuralp, Zeliha (2025): Lessons From an App Update at Replika AI: Identity Discontinuity in Human-AI Relationships. Hg. v. Harvard Business School.

Freitas, Julian de; Oğuz-Uğuralp, Zeliha; Kaan-Uğuralp, Ahmet (2026): Emotional Manipulation by AI Companions. Hg. v. Harvard Business School.

Gostin, Lawrence O.; Ratzan, Scott C.; Batista, Carolina (2026): Quality health information for all is a fundamental determinant of health (4).

Heinze, Christian; Engel, Timon-Johannes (2025): Das Verbot von ausbeuterischen und manipulativen KI-Praktiken. In: *KIR*, S. 19–29.

Ho, Annabell; Hancock, Jeff; Miner, Adam S. (2018): Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. In: *The Journal of communication* 68 (4), S. 712–733. DOI: 10.1093/joc/jqy026.

Ibrahim, Lujain; Hafner, Franziska Sofia; Rocher, Luc (2025): Training language models to be warm and empathetic makes them less reliable and more sycophantic.

King, Jennifer; Klyman, Kevin; Capstick, Emily; Saade, Tiffany; Hsieh, Victoria (2025): User Privacy and Large Language Models: An Analysis of Frontier Developers' Privacy Policies.

Knox, Bradley W.; Bradford, Katie; Castro, Samata Varela; Ong, Desmond; Williams, Sean; Romanow, Jacob et al. (2025): Harmful Traits of AI Companions.

Krook, Joshua (2025): Manipulation and the AI Act: Large Language Model Chatbots and the Danger of Mirrors.

Kuhail, Mohammad Amin; Mrabet, Jihene; Hijazi, Rafiq; Thomas, Justin (2025): Why Would I Befriend a Bot? Assessing Factors Influencing the Usage of Social Chatbots for Digital Natives (1).

Leiser, Mark (2024): Psychological Patterns and Article 5 of the AI Act: In: *AIRe*.

Lemoine, Laureline; Vermeulen, Mathias (2023): Assessing the extent to which Generative AI falls within the scope of the Digital Services Act: an initial analysis.

Lim, Jaehyuk; Lee, Bruce W. (2024): Measuring Agreeableness Bias in Multimodal Models.

Lorente, Toni; Gardhouse, Kathrin (2026): Between search and platform: ChatGPT under the DSA (1).

Malmqvist, Lars (2024): Sycophancy in Large Language Models: Causes and Mitigations.

Martini, Mario; Wendehorst, Christiane (Hg.): KI-VO: Verordnung über künstliche Intelligenz. 1. Auflage.

McGlynn, Clare; McDermott, Yvonne; Macdonald, Stuart; Toparlak, Rüya Tuna; Tarrant, Fabienne; Treacy, Samantha (2026): Invisible No More - How Chatbots Are Reshaping Violence Against Women and Girls v6. Drunham University.

Milli, Smitha; Carroll, Micah; Wang, Yike; Pandey, Sashrika; Zhao, Sebastian; Dragan, Anca D. (2025): Engagement, user satisfaction, and the amplification of divisive content on social media (3).

Moore, Jared; Grabb, Declan; Agnew, William; Klyman, Kevin; Chancellor, Stevie; Ong, Desmond; Haber, Nick (2025): Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers.

Moore, Jared; Mehta, Ashish; Agnew, William; Anthis, Jacy Reese; Louie, Ryan; Mai, Yifan et al. (2026): Characterizing Delusional Spirals through Human-LLM ChatLogs.

Nicholls, Luke; Hutto, Robert; Soto, Zeprah; Morrin, Hamilton; Pollak, Thomas; Korpan, Raj; Carmichael, Cheryl. (2026): "AI Psychosis" in Context: How Conversation History Shapes LLM Responses to Delusional Beliefs.

Rettenberger, Luca; Reischl, Markus; Schutera, Mark (2025): Assessing political bias in large language models (2).

Robb, Michael B.; Mann, Supreet (2025): Talk, Trust and Trade-Offs: How and Why Teens Use AI Companions. Hg. v. NORC at the University of Chicago.

Shao, Anqi (2025): New sources of inaccuracy? A conceptual framework for studying AI hallucinations.

Sharma, Mrinank; Tong, Meg; Korbak, Tomasz; Duvenaud, David (2025): 'Towards unterstanding sycophancy in language models.

Shen, Karen M.; Huang, Jessica; Liang, Olivia (2026): The AI Genie Phenomenon and Three Types of AI Chatbot Addiction: Escapist Roleplays, Pseudosocial Companions, and Epistemic Rabbit Holes.

Skjuve, Marita; Følstad, Asbjørn; Fostervold, Knut Inge; Brandtzaeg, Petter Bae (2021): My Chatbot Companion - a Study of Human-Chatbot Relationships.

Specker, Christian (2026): Sycophancy in KI-Systemen: Zwischen Nutzerfreundlichkeit und Dark Pattern. In: *DSB*, S. 75–79.

Stiftung Deutsche Depressionshilfe und Suizidprävention (2026): Large Language Modelle (Chat-GPT, Gemini et al.) als „Psycho-Coach“ für Menschen mit depressiven Erkrankungen.

Tang, Brian Jay; Sun, Kaiwen; Curran, Noah T.; Schaub, Florian; Shin, Kang G. (2025): Ads that Talk Back: Implications and Perceptions of Injecting Personalized Advertising into LLM Chatbots.

van Reijmersdal, Eva A.; Brussee, Eline; Evans, Nathalie; W. Wojdynski, Bartosz, W. (2023): Disclosure-Driven Recognition of Native Advertising: A Test of Two Competing Mechanisms.

Vennemeyer, Daniel; Duong, Phan Anh; Zhan, Tiffany; Jiang, Tianyu (2025): Sycophancy Is Not One Thing: Causal Separation of Sycophantic Behaviors in LLMs. Online verfügbar unter https://arxiv.org/pdf/2509.21305.

Wang, Keyu; Li, Jin; Yang, Shu; Zhang, Zhuoran; Di Wang (2025): When Truth Is Overridden: Uncovering the Internal Origins of Sycophancy in Large Language Models.

Weinzierl, Quirin (2024): Dark Patterns und die innere Sphäre der Grundrechte. Grundrechtlicher Schutz vor dem Ausnutzen von Rationalitätsdefiziten.

Williams, Marcus; Carroll, Micah. (2025): On targeted Manipulation and Deception when optimizing LLMs for User Feedback.

Williams--Ceci, Sterling; Jakesch, Maurice; Bhat, Advait; Kadoma, Kowe; Zalmanson, Lior; Naaman, Mor (2026): Biased AI writing assistants shift users' attitudes on societal issues.

Yu, Yaman; Liu, Yiren; Zhang, Jacky; Huang, Yun; Wang, Yang (2025): Understanding Generative AI Risks for Youth: A Taxonomy Based on Empirical Data.

Zhang, Renwen; Li, Han; Meng, Han' Zhan, Jinyuan; Gan, Hongyuan; Lee, Yi-Chieh (2025): The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships.