Austrian Lab for AI Trust* Dossier 6

KI-Chatbots als Social Companions

Executive Summary

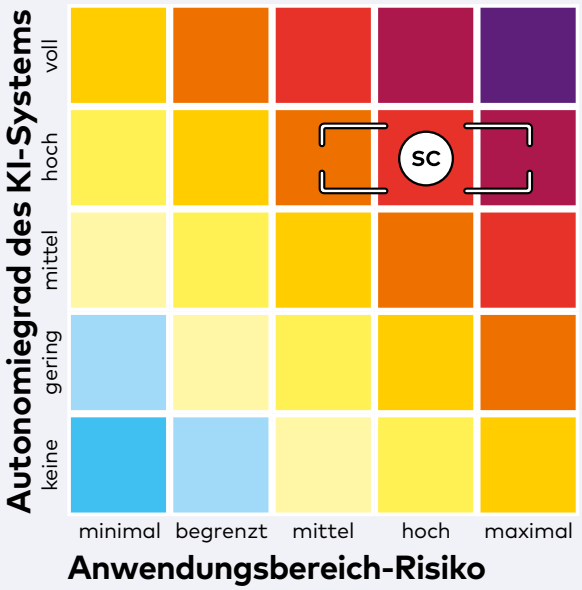
In einer Gesellschaft, die von digitaler Vernetzung und zunehmender Einsamkeit geprägt ist, verändert sich die Art und Weise, wie wir soziale Interaktion definieren und erleben. Der rasante Fortschritt dialogorientierter KI führt zu einer Nutzung von Chatbots als „Social Companions". Die eingesetzten Technologien ermöglichen es, komplexe, emotional nuancierte Mensch-Maschinen-Dialoge in Echtzeit zu führen, die über einfache Befehlssequenzen hinausgehen. Anstatt lediglich als funktionale Werkzeuge zu dienen, werden KI-Systeme zunehmend so konzipiert, dass sie als einfühlsame Zuhörer:innen, Mentor:innen oder virtuelle Freund:innen erscheinen, die im Gegensatz zu menschlichen Partner:innen jederzeit zur Verfügung stehen.

Diese technologische Entwicklung eröffnet neue Möglichkeiten der psychologischen Unterstützung und, auf den ersten Blick, eine Minderung von Einsamkeit. Auf Seiten der Nutzer:innen lassen sich Kommunikationsbarrieren überwinden, wodurch ein sicherer Raum für Austausch entsteht, frei von der Angst vor Beurteilung. Darüber hinaus ermöglichen die Rund-um-die-Uhr-Verfügbarkeit und die beispiellose Personalisierung der KI, dass sich KI-Begleiter:innen an die Bedürfnisse der Nutzer:innen anpassen und so das Gefühl einer tiefen Verbundenheit entstehen kann.

Diese Tools bergen jedoch auch erhebliche ethische Herausforderungen und psychologische Risiken. Falschinformationen und Verzerrungen durch künstliche Intelligenz sind bei KI-Chatbots weit verbreitet. Gefahren wie eine einseitige emotionale Abhängigkeit, bei der die Grenze zwischen echter menschlicher Verbindung und algorithmischer Simulation verschwimmt, sind typisch für soziale Chatbots. Langfristig könnte dies dazu führen, dass sich Nutzer:innen aus realen sozialen Beziehungen zu Menschen zurückziehen. Dies betrifft insbesondere Jugendliche und junge Erwachsene. Zudem bestehen Risiken hinsichtlich ***Datenschutzes*** und Manipulation: Da die Chatbots Zugang zu sehr persönlichen Informationen haben, könnten sie dazu genutzt werden, das Verhalten der Nutzer:innen subtil zu beeinflussen oder bestehende soziale Isolation zu verstärken, indem sie eine ***Echokammer*** der Bestätigung schaffen.

Im ALAIT-Risikoradar ist die Nutzung von Social Companions als „hohes Risiko" (rot, siehe Grafik) klassifiziert. Dies ergibt sich sowohl aus dem Anwendungsbereich als auch aus dem Autonomiegrad: Da in Bezug auf die Anwendung von KI-Chatbots als persönliche Begleiter:innen Hinweise für besondere Gefahren vorliegen, wird ihr Risiko hier zwischen „mittel" (Stufe 3) und „maximal" (Stufe 5) eingestuft (Anmerkung: Dies stellt eine Abweichung zum ***EU AI-Act*** dar, der auf das Thema nicht spezifisch eingeht und Chatbots generell nicht im hohen Risikobereich einstuft.) In Bezug auf den Autonomiegrad sind KI-Chatbots passive Werkzeuge, die selbst keine Handlungen setzen. Da sie jedoch weitgehend autonom antworten, wurde ihr ***Autonomiegrad*** auf Stufe 4 eingeordnet.

ALAIT Risikoradar für KI-Chatbots als Social Companions



SC Social Companions

Das ALAIT Risikoradar ist ein wissenschaftlich entwickeltes Risikoanalysetool für Künstliche Intelligenz (KI), das KI-Anwendungen kontextbezogen und unter Berücksichtigung ihres technischen Autonomiegrades einstuft und so die Risikosphäre für Anwender:innen auf einen Blick sichtbar macht. Dabei gilt: Je höher das Einsatz-Risiko aus dem Anwendungskontext und je größer der Autonomiegrad des KI-Systems in Bezug auf Entscheidungen, desto riskanter ist der Einsatz einzustufen. Eine erweiterte Klammer weist auf eine Bandbreite in der Risikoeinstufung hin. Ein geringer Autonomiegrad eines KI-Systems bedeutet nicht, dass man sich zurücklehnen kann. Es erfordert eine starke Rolle der Menschen, die es anwenden. (Details zum Stufenmodell s. S. 9f).

**ALAIT ist ein vom Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) initiiertes Forschungsprojekt zur Entwicklung einer neuartigen Form von Technologiebewertung für KI-Technologien mit wissenschaftlichen Methoden.*

Als Konsequenz ergibt sich daraus die Empfehlung für die Praxis, diese vorsichtig einzusetzen und Gefahren von Abhängigkeit, einer weiter gehenden sozialen Isolation der Nutzer:innen und Falschinformationen in Betracht zu ziehen. Für einen verantwortungsvollen Umgang mit Social Companions ist es unerlässlich, ihren nicht-menschlichen Status ***transparent*** zu machen und das Bewusstsein für die zugrunde liegenden Algorithmen und KI-Technologien zu schärfen. Chatbots können niemals ein Ersatz für echte menschliche Beziehungen sein. Nur durch die Stärkung von Technologiekompetenz und kritischer menschlicher Selbstreflexion kann erreicht werden, dass Social Companions die Lebensführung unterstützen, ohne die Tiefe echter zwischenmenschlicher Beziehungen zu ersetzen und die psychische Integrität der Nutzer:innen zu gefährden.

Einleitung

Ein Chatbot ist eine KI-Anwendung, die mit Menschen in natürlicher Sprache Dialoge führt. Nutzer:innen können dem Chatbot Fragen stellen und das System antwortet. Dies kann über Spracheingabe und -ausgabe erfolgen, wobei ein Chatbot Text- und Audioeingabe und zunehmend auch andere Datenformate unterstützen kann.[1] Chatbots haben ihren Ursprung in den sechziger Jahren des vergangenen Jahrhunderts. Damals entwickelte der Informatiker Joseph Weizenbaum im Jahr 1966 mit ELIZA eine computerbasierte Sprachverarbeitung, die als frühe Anwendung von Künstlicher Intelligenz gilt.[2]

Die Imitation menschlicher Sprache ist mittlerweile so ausgereift, dass sich KI-gestützte Chatbots für Gespräche eignen und für manche Menschen die Rolle eines Freundes oder einer Freundin übernehmen. Das kann sogar so weit gehen, dass eine Liebesbeziehung entsteht.[3] Social Companions sind besonders darauf ausgelegt, mit Nutzer:innen emotional reaktionsfähige Gespräche zu führen, und übernehmen häufig intime soziale Rollen.[4] Sie sind in der Lage, emotionale Hinweise der Nutzer:innen zu erkennen und ihre Antworten an den wahrgenommenen Gesprächston und die Stimmung anzupassen. Dadurch können sie empathieähnliche Reaktionen und unterstützende Kommunikationsformen und Inhalte bereitstellen. Dies ist insbesondere für Nutzer:innen mit einem Bedarf an emotionaler Zuwendung von Relevanz. Im Verlauf der Nutzung passen sich diese Systeme zunehmend an den Kommunikationsstil und die Präferenzen der Nutzer:innen an. Dadurch werden die Interaktionen als authentisch und persönlich wahrgenommen. Durch ihre kontinuierliche Verfügbarkeit und ihre dialogischen Fähigkeiten können Social Companions ein Gefühl von Begleitung und sozialer Nähe erzeugen.[5] In einem Turing-Test-ähnlichen Experiment konnten über 800 Teilnehmende kaum unterscheiden, ob die Antworten von ChatGPT oder von menschlichen Therapeut:innen kamen und manche fanden die Antworten der KI sogar einfühlsamer und hilfreicher.[6]

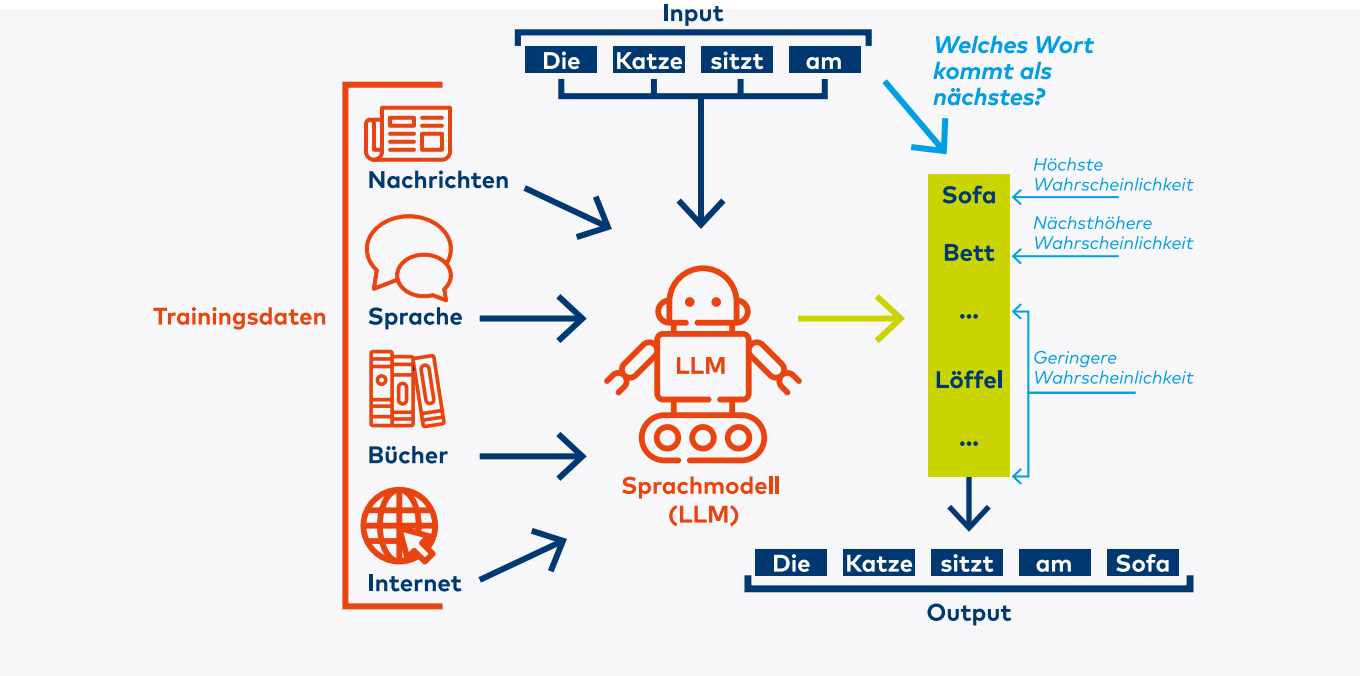
Generell ist die Nutzung von KI-Chatbots im Alltag weit verbreitet. In einer aktuellen Studie geben 94 % der 11- bis 17-Jährigen an, KI-Chatbots zu nutzen – vor allem für schulische und zunehmend auch für persönliche Themen. Knapp ein Viertel (24 %) der Befragten nutzt ChatGPT täglich. Die älteren Befragten (30 %) tun dies häufiger als die jüngeren (18 %). Mehr als 40 Prozent gaben an, ChatGPT mehrmals pro Woche zu nutzen.[7] Repräsentative Umfragen unter Erwachsenen in den USA haben gezeigt, dass 16 % KI zur sozialen Begleitung und 24 % Chatbots zur Unterstützung bei psychischen Problemen genutzt hatten. Unter Erwachsenen im Vereinigten Königreich nutzen schätzungsweise 8 % wöchentlich KI für „emotionale Zwecke". Ähnliche Trends zeigen sich bei jüngeren Nutzer:innen: Umfragen unter US-Teenagern zeigen, dass 13 % generative KI zur emotionalen Unterstützung nutzen, 52 % KI regelmäßig als Gesellschafter:in nutzen und 42 % entweder selbst eine KI-Begleitung hatten oder jemanden kennen, der im vergangenen Jahr eine hatte.[27]

Chatbots können Gespräche führen, Geschichten erzählen, einfach nur Gesellschaft leisten oder an die Einnahme von Medikamenten erinnern und hilfreiche Ratschläge zu verschiedenen Themen geben und so zur Steigerung des individuellen Wohlbefindens beitragen.[8] Die ständige Verfügbarkeit und Aufmerksamkeit, die diese Chatbots bieten, sowie die Möglichkeit, ehrlich und offen zu sein, ohne Scham und Angst, nicht verstanden zu werden, machen diese Technologie besonders attraktiv. Zwar betreffen Einsamkeit, Depressionen und das Bedürfnis, gehört und verstanden zu werden, Menschen jeden Alters, allerdings nutzen ohnehin bereits sozial vulnerable Gruppen wie ältere Menschen, Kinder und Jugendliche besonders häufig Social Companions.[4] Die damit verbundenen Chancen und Herausforderungen für Kinder und Jugendliche werden nach einer kurzen Technologiebeschreibung ausführlich behandelt.

Technologiebeschreibung

KI Chatbots funktionieren auf Basis von Large Language Models (LLMs), auch Sprachmodelle genannt. LLMs nehmen die Benutzereingabe entgegen und erstellen auf Grundlage der Informationen, mit denen sie trainiert wurden, eine Antwort.

Das folgende Diagramm zeigt das Grundprinzip von Sprachmodellen:



Ein LLM lernt, indem es aus sehr vielen Textsorten wie Nachrichten, Bücher, Gesprächen und Internetseiten Muster erkennt, wie Wörter und Sätze typischerweise zusammenhängen. Wenn man dem Modell einen Satzanfang wie „Die Katze sitzt am…" vorgibt, berechnet es das mit der höchsten Wahrscheinlichkeit folgende Wort. Dabei prüft es zahlreiche Möglichkeiten, etwa „Sofa", „Bett" oder „Löffel", und wählt schließlich das wahrscheinlichste Wort aus. So entsteht der vollständige Satz „Die Katze sitzt am Sofa". Kurz gesagt, ein LLM ist ein Programm, das auf Basis von gelerntem Sprachwissen vorhersagt, welches Wort als Nächstes am wahrscheinlichsten passt.

Aus dem oben beschriebenen Funktionsprinzip von LLMs kann man ableiten, dass Chatbots über kein echtes „Verständnis" von Konzepten und Zusammenhängen der realen Welt verfügen. Sprachmodelle sind mathematische Funktionen, die sehr gut das nächste Wort für einen beliebigen Text vorhersagen können. Dennoch scheint diese Technologie sehr überzeugend auf Eingaben zu reagieren, sodass sie in vielen Bereichen für die Imitation von menschlicher Sprache, einschließlich sozialer Interaktionen, eingesetzt wird. Mithilfe von Natural Language Processing (NLP) ist das möglich. NLP hilft Chatbots, die Intention von Nutzer:innen zu erkennen und komplexe sowie lange Anfragen zu interpretieren und zu beantworten.[1]

Chancen

1. **Verfügbarkeit und einfache Verwendung:** Chatbots sind benutzerfreundlich und rund um die Uhr verfügbar. Viele Menschen finden es sehr praktisch, eine Begleiterin oder einen Begleiter zu haben, den sie jederzeit kontaktieren können. Das macht sie auch für einsame Menschen oder solche, die soziale Interaktion suchen, sehr attraktiv.[8] Darüber hinaus haben KI-Chatbots das Potenzial, Mängel in der psychischen Gesundheitsversorgung, wie beispielsweise hohe Behandlungskosten oder lange Wartezeiten aufgrund fehlender Fachkräfte, zumindest in Teilbereichen zu kompensieren.[9]

2. **Informationsunterstützung und Personalisierung:** Chatbots können auf interaktive und einfache Weise relevante Informationen zu fast jedem Thema

bereitstellen, was sie für Nutzer:innen attraktiv macht, die keine Zeit damit verbringen möchten, selbst nach Themen zu recherchieren. Ein weiterer Faktor, der Chatbots für die meisten Nutzer:innen attraktiv macht, ist die Möglichkeit, sie individuell anzupassen. Sie können so programmiert werden, dass sie auf eine Weise reagieren oder sich so verhalten, wie es die Nutzer:innen wünschen, wodurch das Interaktionserlebnis verbessert wird. Einige Unternehmen gehen noch einen Schritt weiter und bieten Chatbots an, die bestimmte Charaktere darstellen, beispielsweise fiktive Figuren oder sogar Therapeut:innen.[10,11]

3. **Wohlbefinden:** KI-basierte Chatbots dürfen in bestimmten Fällen mittlerweile in der deutschen Psychotherapie eingesetzt werden, allerdings nur unterstützend und nicht als alleinige Alternative zu menschlichen Therapeut:innen.[12] Diese spezialisierten Chatbots können vor allem älteren, aber auch jüngeren Menschen dabei helfen, mit Einsamkeit und depressiven Angstzuständen umzugehen, die häufig mit schweren Erkrankungen und Erfahrungen am Lebensende einhergehen.[13] Sie können die Stimmung verbessern und Depressions- und Angstsymptome bei Menschen mit Demenz lindern. Außerdem können sie, wie Studien zeigen, die Einhaltung der Medikamenteneinnahme und die Häufigkeit von Rehabilitationsübungen bei Menschen mit chronisch obstruktiver Lungenerkrankung erhöhen, indem sie Erinnerungen und hilfreiche Informationen bereitstellen.[3,8]

Herausforderungen und Risiken

Für Nutzer:innen:

1. **Halluzinationen und Desinformation:** Ein Problem, das bei allen LLM-basierten Systemen auftritt, sind Halluzinationen, d.h. ungenaue oder falsche Informationen, die zwar authentisch erscheinen, jedoch faktisch falsch sind. Aufgrund von Halluzinationen stellen Fehlinformationen ein großes Risiko dar, insbesondere für Menschen, welche besondere Schwierigkeiten haben, diese Fehlinformationen zu erkennen, wie ältere Menschen und Kinder. Es ist wichtig, die Grenzen von Chatbots anzuerkennen und die ethischen Implikationen der Nutzung von KI-basierten Begleiter:innen für diese Gruppen zu berücksichtigen.[8,14,15]

2. **Übermäßiges Vertrauen und Schmeichelei:** Chatbots haben weder Bewusstsein noch Gefühle. Sie sind daher nicht in der Lage, mit echtem Verständnis oder Unterstützung zu reagieren, was die Gefahr einer einseitigen emotionalen Verletzlichkeit und Abhängigkeit erhöht, insbesondere für diejenigen, die Empathie oder Fürsorge erwarten, welche die Chatbots aber nicht bieten können.[4] Mehrere Studien haben gezeigt, dass Menschen dazu neigen, KI-Systemen mehr zu vertrauen als Menschen - ein Phänomen, das Automatisierungs-Bias bezeichnet wird.[16] Es besteht die Gefahr, dass Menschen KI-generierte Informationen nicht ausreichend hinterfragen, selbst dann, wenn widersprüchliche Informationen oder spezifisches Wissen zu einem konkreten Fall vorliegen.[17] Die Situation wird durch ein Phänomen namens Sycophancy (auf Deutsch „Schmeichelei") verschärft, das die ständige Bestätigung von Menschen durch Chatbots beschreibt. Dies kann ihre Selbstwahrnehmung, ihre Beziehungen und ihre Sicht auf die Welt verzerren. Bei der Frage nach Ratschlägen sind gängige Chatbots wie ChatGPT von OpenAI, Gemini von Google, Claude von Anthropic, Llama von Meta und DeepSeek zu 50 % eher bereit, die Handlungen der Nutzer:innen zu befürworten, als dies Menschen waren.[18]

3. ***Verzerrungen (KI-Bias):*** Sprachmodelle zeigen häufig verschiedene Arten von Verzerrungen (KI-Bias). Da Chatbots anhand von natürlichen Sprachtexten aus Dokumenten und Websites im Internet trainiert werden, übernehmen die Modelle potenziell vorhandene Vorurteile und Stereotypen aus diesen Texten. Dazu können rassistische, fremdenfeindliche und frauenfeindliche Äußerungen gehören. Die Vorurteile, mit denen Chatbots trainiert werden, können somit Fehlinformationen verbreiten und Schaden verursachen.[14] Die Situation wird noch dadurch verschärft, dass KI-Systeme auf Vorab-Training und Fortschritte bei großen Sprachmodellen in Englisch angewiesen sind. Das bedeutet, dass die vom System bereitgestellten Informationen sowohl zeitabhängig (sie können veralten) als auch sprachabhängig sind, was die Leistung von Chatbots in anderen Sprachen als Englisch beeinträchtigt.[8,11,14]

4. **Potenzielle soziale Isolation:** Die Hypothese der sozialen Substitution besagt, dass die Abhängigkeit von KI-Begleiter:innen die Möglichkeiten für hochwertige zwischenmenschliche Beziehungen verdrängen, die Isolation verstärken und das Wohlbefinden möglicherweise beeinträchtigen könnte.[4]

Chatbot-Begleitung bietet möglicherweise nicht nur keine alternativen Sozialisierungsmöglichkeiten für Menschen mit weniger Offline-Unterstützung, sondern untergräbt auch die psychologischen Vorteile, die Menschen durch ein großes Offline-Sozialnetzwerk erhalten. Das Problem wird dadurch verschärft, dass die Technologie die Handlungen und Meinungen der Nutzer:innen, selbst wenn diese schädlich sind, ständig bestätigt, was das Urteil der Menschen über sich selbst, ihre Beziehungen und die Welt um sie herum verzerrt.[18]

5. **Chatbots als Psycholog:innen:** Einige Anbieter:innen von Chatbots bieten verschiedene Dienste an, von einfachen Unterhaltungen mit einer fiktiven Figur bis hin zu einem Chatbot, der sich als Psycholog:in ausgibt. Wenn es auch Spaß machen kann, sich mit einer zufälligen Figur zu unterhalten, ist es sehr problematisch, mit einem Chatbot zu sprechen, der sich als Psycholog:in positioniert. Studien haben gezeigt, dass KI-Chatbots systematisch gegen grundlegende Standards der psychischen Gesundheit und Ethik verstoßen. Dazu gehören die unsachgemäße Bewältigung von Krisensituationen, irreführende Antworten, welche die negativen Überzeugungen der Nutzer:innen über sich selbst und andere verstärken, sowie die Schaffung eines falschen Gefühls der Empathie gegenüber den Nutzer:innen.[9,11] Werbung für Psycholog:innen-Chatbots ist daher gefährlich und unverantwortlich, vor allem weil es noch keine festgelegten regulatorischen Rahmenbedingungen gibt.[9]

6. ***Datenschutz:*** Datenschutzbedenken bestehen bei LLMs, die mit Webdaten trainiert wurden, die häufig personenbezogene Daten enthalten,[15] als auch solchen, die Angaben der Nutzer:innen für das weitere Training benutzen. Aufgrund der Art und Weise, wie Daten erfasst werden und in das Modell Eingang finden, könnten LLM-Chatbots unbeabsichtigt private Nutzerinformationen gegenüber anderen Nutzer:innen preisgeben. Mehrere Studien haben gezeigt, dass es durch bestimmte Eingaben in Chatbots möglich ist, deren ursprüngliche Anweisungen zu überschreiben und sie so zu einem Verhalten zu veranlassen, das die persönlichen Daten der Nutzer:innen gefährdet.[8,14]

7. **Autonomieverlust:** Nutzer:innen könnten sich in Bezug auf ihr Verhalten oder ihre Emotionen zu sehr auf KI-generierte Ratschläge verlassen und dabei versäumen, deren Qualität kritisch zu hinterfragen. Eine solche Abhängigkeit von synthetischen Beziehungen könnte nach und nach die Selbstbestimmung untergraben, die persönliche Autonomie aushöhlen und ethische Bedenken hinsichtlich der individuellen Freiheit aufwerfen.[19]

Für die Gesellschaft:

8. **Energie- und Ressourcenverbrauch:** Die Entwicklung und Herstellung von Hardware sowie das Training und die tägliche Nutzung von KI erfordern enorme Mengen an Strom, Wasser und weiteren Ressourcen.[20] Die Internationale Energieagentur (IAEA) schätzt den Anteil der Rechenzentren am globalen Stromverbrauch zwar aktuell noch auf 1,5 %, erwartet jedoch, dass dieser bis 2030 auf 10 % steigen wird.[21] Eine derart drastische Veränderung des Energieverbrauchs dürfte aufgrund der hohen Nachfrage zu Instabilität im Energiesektor führen. Dies hätte höhere Preise für die Bevölkerung zur Folge. Darüber hinaus könnte eine gesteigerte Energieproduktion dazu führen, dass fossile Kraftwerke verstärkt eingesetzt werden, wodurch sich die bestehende Klimakrise weiter verschärfen würde.[20]

9. **Potenzielle Polarisierung der Gesellschaft:** Durch Schmeichelei und ***KI-Verzerrungen*** werden die Überzeugungen und Interessen der Nutzer:innen verstärkt. Dies führt zu einer eingeschränkten Inhaltsdarstellung und begünstigt somit die Entstehung von ***Echokammern*** und ***Filterblasen***.[19] Der Begriff „Echokammer" beschreibt eine Informationsblase, in der Nutzer:innen nur mit Artikeln konfrontiert werden, die ihre bestehenden Überzeugungen bestätigen. Der Begriff „Filterblase" bezeichnet eine intellektuelle Isolation, den Zustand, von abweichenden Standpunkten und neuen Informationen abgeschnitten zu sein. Filterblasen entstehen durch personalisierte Suchanfragen oder Algorithmen. Dadurch werden Informationen selektiv übernommen und eingeschränkt, was zu einer polarisierten Informationsumgebung führt.

10. **Erosion der Zwischenmenschlichkeit**: Ein wesentliches Risiko von Social Companions liegt in der schleichenden Normalisierung rein serviceorientierter Beziehungsmodelle, bei denen die sofortige Befriedigung eigener Bedürfnisse im Zentrum steht. Diese einseitige Interaktionsform droht das gesellschaftliche Verständnis von Intimität zu verzerren, indem sie Erwartungshaltungen prägt, die zunehmend auch auf menschliche Beziehungen, wie z. B. auf Dating-Plattformen, projiziert werden. Besonders bei Jugendlichen ist die Gefahr groß, dass die KI nicht mehr nur als Ergänzung, sondern als Ersatz für reale soziale Erfahrungen fungiert. Da Chatbots lediglich parasoziale Bindungen ohne echte Reziprozität ermöglichen, entziehen sie jungen Menschen wichtige Entwicklungsräume für Empathie- und Konfliktfähigkeit.[4,26]

Empfehlung zum Praxiseinsatz

1. **Echte menschliche Kommunikation und soziale Interaktionen fördern:** Chatbots können bestehende soziale Netzwerke zwar ergänzen, sie jedoch niemals vollständig ersetzen. Kinder und Jugendliche benötigen ausreichend echte soziale Interaktionen abseits der digitalen Welt, um Lernen, Reflexions- und Beziehungsfähigkeit zu schulen und nicht in ungesunde Abhängigkeiten zu geraten. Insbesondere in schwierigen Situationen ist es wichtig, professionelle Hilfe oder menschliche Unterstützung in Anspruch zu nehmen.

2. **Ausrichtung an Richtlinien und Transparenz:** Alle Organisationen – sowohl private als auch öffentliche – sind an die Richtlinien gebunden und dürfen mit ihren KI-Systemen nicht davon abweichen. Insbesondere müssen diese Systeme der Datenschutzgrundverordnung (DSGVO) entsprechen, um die Privatsphäre der Nutzer:innen zu schützen. Die Verarbeitung personenbezogener Daten muss den zentralen Datenschutzanforderungen der Notwendigkeit und Verhältnismäßigkeit entsprechen.[22] Die Einhaltung der DSGVO ist zwingend erforderlich, und therapeutische KI-Systeme fallen unter die „Hochrisiko"-Kategorien des EU AI Act.[12] Behörden wie das Europäische Büro für Künstliche Intelligenz („AI Office") und nationale Marktüberwachungsbehörden überwachen die Umsetzung und Anwendung des EU AI-Acts. Artikel 50 des EU AI-Acts schreibt vor, dass Nutzer:innen, die mit einer KI (z.B. einem Chatbot) interagieren, darüber informiert werden müssen, dass sie mit einem KI-System kommunizieren, sofern dies nicht offensichtlich ist.[23]

3. **Datenschutzrechte:** Darüber hinaus haben die Benutzer:innen ein Recht darauf zu erfahren, wie und wann ihre personenbezogenen Daten genutzt und in welche Systeme sie eingespielt werden (Art. 15 DSGVO).[22] Anlaufstelle bei Verdacht auf die Verletzung von Datenschutzrechten ist die österreichische Datenschutzbehörde (DSB).[24]

4. **Informationsverifizierung:** Da Sprachmodelle halluzinieren und desinformieren können, ist es wichtig, den Wahrheitsgehalt von Informationen eigenständig zu überprüfen, vor allem bei sensiblen Themen. Mithilfe von Faktenchecks können sich Nutzer:innen der Chatbots auch zu komplexen und umstrittenen Themen eine Meinung bilden – basierend auf rationalen Zugängen und Fakten. So ist es beispielsweise der APA ein Anliegen, durch die Faktenchecks einen Einblick in die Recherchemethoden des digitalen Zeitalters zu geben.[25] CORRECTIV. Faktencheck bietet auch Dienstleistungen im Bereich Faktencheck an.

5. **Schaffung und Stärkung des KI-Wissens:** Es ist wichtig, sich über KI-Technologien zu informieren, Kompetenzen aufzubauen und die eigenen Rechte zu kennen.[17] In Österreich besteht mit der Rundfunk- und Telekom Regulierungs-GmbH (RTR) eine Informations- und Anlaufstelle für die breite Öffentlichkeit. Die RTR-KI-Servicestelle unterstützt Bürger:innen bei Fragen zur Transparenz, zu Rechten im Umgang mit KI-Systemen oder zur Überprüfung von KI-Entscheidungen.

Wichtige Begriffe

Datenschutz in der KI: Die Gesamtheit der Praktiken und Bedenken im Zusammenhang mit der ethischen Erfassung, Speicherung und Nutzung personenbezogener Daten durch Systeme der künstlichen Intelligenz.

Echokammer: Bezieht sich auf eine Informationsblase um einen Nutzer bzw. eine Nutzerin herum, in der nur Informationen zu finden sind, die die bestehenden Überzeugungen der Nutzerin bzw. des Nutzers bestätigen.

Erklärbare KI (XAI): Bezieht sich auf Methoden und Techniken, die es ermöglichen, die Entscheidungen und Ergebnisse von KI-Systemen besser zu verstehen und nachzuvollziehen.

Filterblase: Entspricht der intellektuellen Isolation, die durch personalisierte Suchanfragen oder Algorithmen verursacht wird, um selektiv die Informationen anzunehmen, die eine Person sehen möchte.

Gesetz der Europäischen Union über künstliche Intelligenz (EU AI-Act): Eine europäische Verordnung über künstliche Intelligenz (KI) – die erste umfassende Verordnung über KI von einer großen Regulierungsbehörde überhaupt. Sie konzentriert sich insbesondere auf KI-Systeme mit hohem Risiko.

KI-Autonomie: Die Fähigkeit eines KI-Systems, eine Reihe von Zielen unter einer Reihe von Unsicherheiten in ihrer Umgebung selbstständig und ohne externe Eingriffe zu erreichen.

Systematische Verzerrungen (Bias): Bias ist eine systematisch unterschiedliche Behandlung bestimmter Objekte, Personen oder Gruppen im Vergleich zu anderen. Behandlung ist jede Art von Handlung, einschließlich Wahrnehmung, Beobachtung, Darstellung, Vorhersage oder Entscheidung.

Transparenz: Bedeutet, dass die Funktionsweise, Entscheidungsprozesse und Einsatzbereiche eines KI-Systems nachvollziehbar, erklärbar und offen zugänglich sind – für Entwickler:innen, Nutzer:innen und andere Stakeholder.

Erklärung Stufenmodell des ALAIT Risikoradars

Im ALAIT KI-Risikoradar wird die Beziehung zwischen Anwendungsrisiko und Autonomie eines KI-Systems dargestellt. Die Risikostufen stützen sich auf das EU KI-Gesetz (***EU AI-Act***), insbesondere auf Artikel 6 und Annex III, die sich mit risikoreichen Anwendungsbereichen von KI befassen. Geringere System-Autonomie und Anwendungsrisiken werden durch kältere Farben (blau) und höhere System-Autonomie und Anwendungsrisiken durch wärmere Farben (rot) dargestellt. Der Farbwechsel vermittelt das erhöhte Risiko solcher Entscheidungen. Mithilfe dieser Farbskala lässt sich das Gesamtrisiko erkennen: Violett und dunkelrot – sehr hoch, rot und dunkelorange - hoch, hellorange und Gelbtöne - mittel, Blautöne - geringes Gesamtrisiko. Im Idealfall sollten hohe Anwendungsrisiken und System-Autonomie vermieden oder nur nach sehr sorgfältiger Abwägung eingesetzt werden.

Autonomiegrad des KI-Systems

Stufe 1: Keine Autonomie

KI ist ein passives Werkzeug; Menschen treffen alle Entscheidungen und leiten Maßnahmen ein.

Beispiel: Diagnosesysteme, die medizinische Rohdaten anzeigen oder die Daten analysieren (ohne Empfehlungen!)

Empfohlene Anwendungsfälle: Szenarien mit hohen Risiken oder bei denen ethische Entscheidungen von entscheidender Bedeutung sind (z.B. medizinische Diagnostik, Justizsystem).

Stufe 2: Geringer Autonomiegrad (Human-in-the-Loop)

Die KI gibt Empfehlungen oder Optionen, aber der Benutzer:innen bleibt für die Auswahl und Genehmigung von Maßnahmen verantwortlich.

Beispiel: KI schlägt optimale Routen für die Logistik vor oder Empfehlungssysteme im E-Commerce.

Empfohlene Anwendungsfälle: Aufgaben mittlerer Komplexität mit mäßigen Risiken (z.B. Optimierung der Lieferkette).

Stufe 3: Mittlerer Autonomiegrad (Human-on-the-Loop)

Die KI führt bestimmte Aufgaben autonom aus, wobei Menschen in Ausnahmefällen eingreifen.

Beispiel: KI-gestützte Fertigungsprozesse, bei denen das System Maschinen steuert, aber Nutzende bei Anomalien eingreifen.

Empfohlene Anwendungsfälle: Szenarien, in denen eine kontinuierliche menschliche Beteiligung nicht erforderlich ist, kritische Risiken jedoch eine menschliche Überwachung erfordern (z.B. industrielle Automatisierung, Überwachung von Finanztransaktionen).

Stufe 4: Hoher Autonomiegrad (Human in Control)

Das KI-System arbeitet weitgehend autonom, erlaubt es den Benutzern jedoch, es selbst zu übersteuern, um unerwünschte Ergebnisse zu vermeiden.

Beispiel: Autonome Fahrzeuge

Empfohlene Anwendungsfälle: Umgebungen mit geringem bis mittlerem Risiko (z.B. Logistik, einfaches Verkehrsmanagement).

Stufe 5: Vollständige Autonomie mit minimaler Aufsicht

Das KI-System arbeitet unabhängig und erfordert nur minimale oder gar keine menschliche Intervention. Die Beteiligung des Menschen beschränkt sich auf die langfristige Aufsicht (Audits).

Beispiele: Autonome landwirtschaftliche Maschinen, KI für die Stromnetzverteilung, U-Bahnen, Flughafenbahnen

Empfohlene Anwendungsfälle: Umgebungen mit geringen Sicherheits- oder ethischen Risiken und hoher Zuverlässigkeit des KI-Systems (z.B. sich wiederholende Aufgaben in kontrollierten Umgebungen).

Anwendungsbereich-Risiko

Stufe 1: Minimales Risiko

Das KI-System hat keine Auswirkungen auf die Benutzer:innen oder die Entscheidungsfindung.

Beispiele: Filter, NPCs, Empfehlungsalgorithmen ohne schwerwiegende Folgen (DeepL, andere Übersetzungssysteme)

Kriterien: Keine direkte Auswirkung auf die Hochrisikobereichen des ***EU AI-Acts***.

Stufe 2: Begrenztes Risiko

KI-Systeme, die mit Benutzer:innen interagieren, aber keine Entscheidungen mit hohen Risiken treffen. Das Risiko steigt, wenn es an Transparenz über die Beteiligung von KI mangelt.

Beispiele: Chatbots und KI-generierte Inhalte ohne Offenlegung, einfache Automatisierungsaufgaben.

Kriterien: Bereiche, die nicht in der Liste der „hohen Risiken" des EU AI-Acts enthalten sind.

Stufe 3: Mittleres Risiko

KI-Systeme haben keine besonderen Auswirkungen auf einzelne Personen, aber sie entfalten Wirkung auf kollektiver oder gesellschaftlicher Ebene.

Beispiele: Generative KI wie ChatGPT und andere Systeme, die indirekt die Umgebung beeinflussen können, in der sie eingesetzt werden.

Kriterien: KI-Systeme, die für die öffentliche Nutzung verfügbar sind und das Potenzial haben, bestehende Gepflogenheiten zu beeinflussen und langfristig zu verändern.

Stufe 4: Hohes Risiko

Jeder Algorithmus, der in den laut EU AI-Act „Hochrisikobereichen" angewendet wird oder direkte Auswirkungen auf einzelne Personen hat.

Beispiele: Medizin, Biometrie, kritische Infrastruktur, Bildung und Berufsausbildung, Beschäftigung, Zugang zu Dienstleistungen im öffentlichen Sektor, Strafverfolgung, Migration.

Kriterien: Zugehörigkeit zum „Hochrisikobereich" des EU AI-Acts, nur wenn die Regeln für Transparenz und Datenqualität eingehalten werden.

Stufe 5: Extremes Risiko

Jeder Algorithmus, der in den laut EU AI-Act „Hochrisikobereichen" angewendet wird.

Beispiele: Medizin, Biometrie, kritische Infrastruktur, Bildung und Berufsausbildung, Beschäftigung, Zugang zu Dienstleistungen im öffentlichen Sektor, Strafverfolgung, Migration.

Kriterien: Zugehörigkeit zum „Hochrisikobereich", wenn die Regeln für Transparenz und Datenqualität NICHT eingehalten werden.

Quellen

1 Wie funktionieren Chatbots? (n.d.). Technikum Wien Academy. Retrieved 26 February 2026, from https://academy.technikum-wien.at/ratgeber/was-sind-chatbots/

2 Exkurs: Social Bots und Chatbots. (n.d.). Bundesamt für Sicherheit in der Informationstechnik. Retrieved 19 January 2026, from https://www.bsi.bund.de/DE/Themen/Verbraucherinnen-und-Verbraucher/Informationen-und-Empfehlungen/Onlinekommunikation/Soziale-Netzwerke/Sichere-Verwendung/Exkurs-bots/social-bots.html?nn=896986

3 Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My Chatbot Companion—A Study of Human-Chatbot Relationships. International Journal of Human-Computer Studies, 149, 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

4 Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). The Rise of AI Companions: How Human-Chatbot Relationships Influence Well-Being (arXiv:2506.12605). arXiv. https://doi.org/10.48550/arXiv.2506.12605

5 Kouros, T., & Papa, V. (2024). Digital Mirrors: AI Companions and the Self. Societies, 14(10), 200. https://doi.org/10.3390/soc14100200

6 Hatch, S. G., Goodman, Z. T., Vowels, L., Hatch, H. D., Brown, A. L., Guttman, S., Le, Y., Bailey, B., Bailey, R. J., Esplin, C. R., Harris, S. M., Jr, D. P. H., McLaughlin, M., O'Connell, P., Rothman, K., Ritchie, L., Jr, D. N. T., & Braithwaite, S. R. (2025). When ELIZA meets therapists: A Turing test for the heart and mind. PLOS Mental Health, 2(2), e0000145. https://doi.org/10.1371/journal.pmen.0000145

7 Studie: Werkzeug, Ratgeber, Bezugsperson – 94 Prozent der Jugendlichen nutzen KI-Chatbots. (2026, February 9). Saferinternet.at. https://www.saferinternet.at/presse-detail/studie-werkzeug-ratgeber-bezugsperson-94-prozent-der-jugendlichen-nutzen-ki-chatbots

8 Alessa, A., & Al-Khalifa, H. (2023). Towards Designing a ChatGPT Conversational Companion for Elderly People. Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments, PETRA '23, 667–674. https://doi.org/10.1145/3594806.3596572

9 Brown University. (2025, October 21). New study: AI chatbots systematically violate mental health ethics standards | Brown University. https://www.brown.edu/news/2025-10-21/ai-mental-health-ethics

10 About Character.AI. (n.d.). Retrieved 20 January 2026, from https://policies.character.ai/about

11 Wells, S. (2025, June 11). Exploring the Dangers of AI in Mental Health Care | Stanford HAI. Exploring the Dangers of AI in Mental Health Care. https://hai.stanford.edu/news/exploring-the-dangers-of-ai-in-mental-health-care

12 Demirel, Y. (2025, October 31). KI in der Therapie 2025: Was ist erlaubt? Ratgeber & Regeln. clarathy. https://clarathy.de/blog/ki-in-der-therapie-erlaubt-was-darf-ich

13 Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis. Journal of Medical Internet Research, 22(3), e16235. https://doi.org/10.2196/16235

14 Casheekar, A., Lahiri, A., Rath, K., Prabhakar, K. S., & Srinivasan, K. (2024). A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions. Computer Science Review, 52, 100632. https://doi.org/10.1016/j.cosrev.2024.100632

15 He, F., Zhu, T., Ye, D., Liu, B., Zhou, W., & Yu, P. S. (2025). The Emerged Security and Privacy of LLM Agent: A Survey with Case Studies. ACM Comput. Surv., 58(6), 162:1-162:36. https://doi.org/10.1145/3773080

16 Wihlborg, E., Larsson, H., & Hedström, K. (2016). 'The Computer Says No!' – A Case Study on Automated Decision-Making in Public Authorities. 2016 49th Hawaii International Conference on System Sciences (HICSS), 2903–2912. https://doi.org/10.1109/HICSS.2016.364

17 Thapa, B. (2019). Predictive Analytics and AI in Governance: Data-driven government in a free society. The European Liberal Forum. https://liberalforum.eu/publication/predictive-analytics-and-ai-in-governance-data-driven-government-in-a-free-society/

18 Sample, I. (2025, October 24). 'Sycophantic' AI chatbots tell users what they want to hear, study shows. The Guardian. https://www.theguardian.com/technology/2025/oct/24/sycophantic-ai-chatbots-tell-users-what-they-want-to-hear-study-shows

19 Starke, C., Ventura, A., Bersch, C., Cha, M., de Vreese, C., Doebler, P., Dong, M., Krämer, N., Leib, M., Peter, J., Schäfer, L., Soraperra, I., Szczuka, J., Tuchtfeld, E., Wald, R., & Köbis, N. (2024). Risks and protective measures for synthetic relationships. Nature Human Behaviour, 8(10), 1834–1836. https://doi.org/10.1038/s41562-024-02005-4

20 Ressourcenverbrauch von KI: Die Nimmersatt-Industrie und ihre Kosten. (n.d.). AlgorithmWatch. Retrieved 26 January 2026, from https://algorithmwatch.org/de/ressourcenverbrauch-von-ki/

21 Executive summary – Energy and AI – Analysis. (n.d.). IEA. Retrieved 26 January 2026, from https://www.iea.org/reports/energy-and-ai/executive-summary

22 Art. 15 DSGVO – Auskunftsrecht der betroffenen Person. (n.d.). Datenschutz-Grundverordnung (DSGVO). Retrieved 7 September 2025, from https://dsgvo-gesetz.de/art-15-dsgvo/

23 Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems | EU Artificial Intelligence Act. (n.d.). Retrieved 14 October 2025, from https://artificialintelligenceact.eu/article/50/

24 Datenschutzbehörde, Ö. (n.d.). Österreichische Datenschutzbehörde. Österreichische Datenschutzbehörde. Retrieved 5 September 2025, from https://dsb.gv.at/

25 APA. (n.d.). Faktencheck | News Verifizieren | APA - Austria Presse Agentur. Retrieved 24 November 2025, from https://apa.at/service/faktencheck-2/

26 Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My Chatbot Companion—A Study of Human-Chatbot Relationships. International Journal of Human-Computer Studies, 149, 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

27 Moore, J., Mehta, A., Agnew, W., Anthis, J. R., Louie, R., Mai, Y., Yin, P., Cheng, M., Paech, S. J., Klyman, K., Chancellor, S., Lin, E., Haber, N., & Ong, D. (n.d.). Characterizing Delusional Spirals through Human-LLM ChatLogs.

Projekt ALAIT

Das Austrian Lab for AI Trust (ALAIT) ist ein vom österreichischen Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) gefördertes Forschungs- und Entwicklungs-Projekt zur Schaffung von Vertrauen durch Wissen im Bereich Künstliche Intelligenz (KI). Das Projekt ALAIT zielt darauf ab, Interessierte und wichtige gesellschaftliche Gruppen zu befähigen, KI-Technologien verantwortungsvoll zu nutzen und ethische sowie qualitativ hochwertige Standards für den Einsatz von AI zu etablieren.

Das Projekt wird von **winnovation** geleitet (Gertraud Leimüller und Brigitte Ömer-Rieder) und im Konsortium mit **leiwand.ai** (Rania Wazir und Silvia Wasserbacher-Schwarzer), **TU Wien** (Sabine Köszegi und Ilya Faynleyb) und **Austria Presse Agentur – APA** (Verena Krawarik und Sophia Marecek) umgesetzt.

Die ALAIT-Dossiers sind auf der Projekthomepage abrufbar: https://science.apa.at/project/alait/

Die Inhalte des Dossiers entsprechen dem aktuellen Stand der Technik und wurden sorgfältig nach wissenschaftlichen Kriterien erstellt. Sie dienen jedoch nicht als rechtsverbindliche Auskunft oder Beratung.

Impressum

Medieninhaberin und Herausgeberin:
winnovation consulting gmbh
Linke Wienzeile 42/1, Top 5
1060 Vienna

Veröffentlicht 2026, gefördert durch das Bundesministerium für Innovation, Mobilität und Infrastruktur.

Danksagung:

Wir danken folgender Expertin für ihr hilfreiches Feedback zu Vorversionen dieses Dossiers:
Laura Wiesböck.