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]

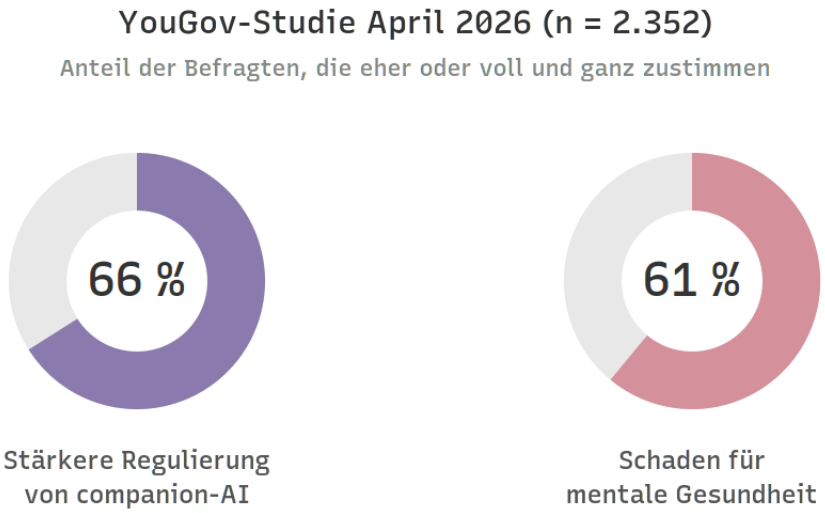
Schutz sensibler Daten: Die DSGVO bietet mit Art. 9 ein hohes Schutzniveau für sensible Daten, wie sie in Gesprächen mit Companion-AI regelmäßig anfallen. Entscheidend ist eine wirksame behördliche Durchsetzung.

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.