Verbraucherzentrale
Bundesverband

General Purpose Chatbots

Risikoanalyse – Rechtsrahmen – Reformvorschläge

20. Oktober 2026

Inhalt

Kurzzusammenfassung 5

Executive Summary 7

I. Einleitung 9

1. Problemaufriss und Relevanz 10
2. Abgrenzungen 13
3. Begriffsdefinitionen 15

II. Risikoanalyse und empirische Evidenz 18

1. Methodischer Ausgangspunkt: Systematisierung statt Metaanalyse 18
2. Systematisierung der Risikodimensionen 19
2.1 Individuelle Risiken 20
2.2 Systemische Risiken 28
2.3 Fazit 34

III. Analyse des geltenden Rechtsrahmens 37

1. Einleitung/Vorgehen 37
2. KI-Verordnung 37
2.1 Allgemeines 37
2.2 Risikenbewertung 38
2.3 Zwischenfazit 49
3. Datenschutzrecht 49
3.1 Allgemein 49
3.2 Risikenbewertung 53
3.3 Zwischenfazit 63
4. Plattformregulierung 64
4.1 Allgemeines 64
4.2 Risikenbewertung 66
4.3 Zwischenfazit 73
5. Lauterkeitsrecht (UWG / UGP-RL) 75
5.1 Allgemeines 75
5.2 Risikenbewertung 77
5.3 Zwischenfazit 82

6. Kinder- und Jugendmedienschutzrecht 82
6.1 Allgemeines 82
6.2 Risikenbewertung 83
6.3 Zwischenfazit 86
7. Ausblick: EU KIDS ACT 86
8. Matrixanalyse: Risiken – Normen – Schutzlücken 88

IV. Grundsätzliche Überlegungen 97

1. Bedarf eines neuen Schutzkonzepts 97
2. Reichweite und Grenzen der informationellen Selbstbestimmung 97
3. Kognitive Selbstbestimmung als ergänzender Schutzgegenstand 98
3.1 Begriffliche Herleitung 98
3.2 Sachlicher Anwendungsbereich: Zuordnung zu den identifizierten Risiken 99
3.3 Normative Anschlussfähigkeit 99
3.4 Normative Tragfähigkeit und offene Fragen 100
4. Paradigmenwechsel: von punktuellen Verboten zur systemischen Risikosteuerung 100
4.1 Grenzen des Verbotsmodells 100
4.2 Das Systemrisikoregime der KI-Verordnung als Vorbild 101
4.3 Designbezogene Regulierung als komplementärer Baustein 101
5. Kausalität, Dokumentation und Audit 102
5.1 Strukturelles Kausalitätsproblem 102
5.2 Von Kausalitätsnachweis zu Prozesspflichten 102
5.3 Externe Audits nach dem Vorbild der Medizinprodukteregulierung 102
6. Zwischenfazit: Einordnung in den erarbeiteten Regulierungsrahmen 103

V. Reformvorschläge 105

1. Anwendungslücken 106
1.1 Digital Services Act 106
1.2 Emotionserkennung: Öffnung der Legaldefinition 109
1.3 Jugendmedienschutz: die Vorsorgepflicht erweitern 110
2. Tatbestandslücken: Verhaltensmanipulation ausdrücklich erfassen 110
2.1 Erweiterung des Art. 5 KI-VO 110
2.2 Einstufung als Hochrisikosystem i.S.d. KI-VO 111
2.3 Konkretisierung der systemischen Risiken 111
2.4 Ergänzung der schwarzen Liste im Digital Fairness Act 112
3. Designbezogene Vorgaben: Safe Interaction Design 113
4. Transparenz 114

5. Absenkung der Nachweisschwellen 114
5.1 Offenlegungsanordnung und Vermutung im Lauterkeitsrecht 114
5.2 Abkehr von der konkreten Verletzungsform bei Interaktionsdesign 115
5.3 Rechenschaftspflicht für die Altersfeststellung 115
6. Rechtsdurchsetzung und Sanktionen 115
6.1 Kollektiver Rechtsschutz: die nationale Lücke schließen 115
6.2 Aufsichtsbefugnisse 116
6.3 Sanktionen 116
7. Zusammenfassung der Vorschläge 116

VI. Zusammenfassung der Ergebnisse 118

1. Kernergebnisse 118
1.1 Chatbots erzeugen systemische Risiken für die Autonomie von Verbraucher:innen 118
1.2 Das geltende Recht ist fragmentiert und schwer durchsetzbar 119
2. Zentrale Reformnotwendigkeiten 121
2.1 Schließung normativer Lücken 121
2.2 Stärkung der Durchsetzung 122
2.3 Einführung neuer Schutzkonzepte 122
3. Vereinfachte Darstellung zentraler Befunde 123

VII. Literaturverzeichnis 124

VIII. Abbildungsverzeichnis 130

IX. Tabellenverzeichnis 130

Impressum 131

Kurzzusammenfassung

Chatbots sind Ratgeber bei Gesundheit, Finanzen und persönlichen Entscheidungen. In Deutschland nutzen 58 Prozent der Menschen ab 16 Jahren KI, 40 Prozent zur Unterstützung eigener Entscheidungen. International sind Therapie und Companionship die häufigste Nutzungsform generativer KI. Ihr Anteil hat sich binnen eines Jahres verdoppelt.

Nutzen und Risiko stammen aus derselben Quelle. Emotional responsive Sprache, ständige Verfügbarkeit und personalisierte Zuwendung machen diese Systeme wertvoll – und erzeugen zugleich Vertrauen über das rational gerechtfertigte Maß hinaus sowie Bindungen, deren kommerzielle Bewirtschaftung für Nutzende nicht erkennbar ist. Die Risiken beschränken sich nicht auf Systeme, die als Beziehungsprodukt vermarktet werden: Auch unveränderte General-Purpose-Systeme zeigen ohne jede Aufforderung emotionale Erwiderungen, eine durchgehende Persona und proaktives Engagement.

Das geltende Recht greift, aber nicht dort, wo es müsste. Die Schwachstellen liegen ganz überwiegend nicht bei den Verboten, sondern bei Nachweis und Durchsetzung. Das Recht knüpft an einzelne Äußerungen, an Datenflüsse und an Produktkategorien an – die Wirkung dieser Systeme entsteht an keinem dieser drei Orte, sondern in der Architektur der Interaktion.

Das Datenschutzrecht erfasst Profilbildung und Speicherung, scheitert aber an der Einwilligung, deren Informiertheit die Informationsasymmetrie gerade verhindert. Die KI-Verordnung stuft Chatbots als risikoarm ein; ihr Verbotskatalog kennt Alter und wirtschaftliche Lage, nicht aber emotionale Notlagen. Die Einstufung von ChatGPT als sehr große Online-Suchmaschine eröffnet das Risikomanagement des Digital Services Act, nicht aber dessen Gestaltungsschutz – dieser gilt nur für Online-Plattformen. Das Lauterkeitsrecht bietet einen Tatbestand für Manipulation ohne Täuschung, scheitert aber am Nachweis. Der Jugendmedienschutz kennt Interaktionsrisiken von Dialogen nicht als eigene Risikokategorie.

Drei Handlungsfelder:

Normative Lücken schließen. Emotionale Vulnerabilität als Verbotsgrund in Art. 5 KI-VO anerkennen. Designbezogene Vorgaben (*Safe Interaction Design*) einführen, die an der Gestaltung ansetzen statt an der einzelnen Äußerung. Anknüpfungspunkte in DSA und Jugendmedienschutz korrigieren.

Durchsetzung stärken. Gerichtliche Offenlegungsanordnung mit Vermutungsfolge nach dem Vorbild der Produkthaftungsrichtlinie. Unterlassungstitel gegen die Gestaltungsfunktion statt gegen die konkrete Formulierung. Rechenschaftspflicht für die Altersfeststellung. Kollektiver Rechtsschutz für die KI-Verordnung auch national eröffnen.

Neue Schutzkonzepte einführen. Kognitive Selbstbestimmung als Leitprinzip. Wechsel von punktuellen Verboten zu fortlaufender Systemgovernance. Standardisierte Risikodokumentation und unabhängige externe Prüfung nach dem Vorbild der Medizinprodukteregulierung.

Kernaussage. Es fehlt weniger an Verboten als an Anknüpfungspunkten und an Durchsetzbarkeit. Wären nur zwei Maßnahmen umzusetzen, sollten es die designbezogenen Vorgaben und die Beweiserleichterung sein.

Executive Summary

Chatbots serve as advisers on health, finances and personal decisions. In Germany, 58 per cent of people aged 16 and overuse AI, and 40 per cent use it to support their own decision-making. Internationally, therapy and companionship are the most common use of generative AI, and their share has doubled within a single year.

Benefit and risk stem from the same source. Emotionally responsive language, constant availability and personalised attention make these systems valuable. At the same time, they generate trust beyond what is rationally warranted, and attachments whose commercial exploitation users cannot recognise. The risks are not confined to systems marketed as relationship products: even unmodified general-purpose systems display emotional reciprocation, a consistent persona and proactive engagement without any prompting.

Existing law does apply, but not where it needs to. Its weaknesses lie overwhelmingly not in the prohibitions but in proof and enforcement. The law attaches to individual statements, to data flows and to product categories. Yet the effect of these systems arises in none of these three places. It arises in the architecture of the interaction.

Data protection law covers profiling and storage, but founders on consent: the very information asymmetry at issue prevents that consent from being informed. The AI Act classifies chatbots as limited-risk; its catalogue of prohibited practices recognises age and economic situation, but not emotional distress. Designating ChatGPT as a very large online search engine triggers the Digital Services Act's risk-management obligations, but not its design-related protections, which apply only to online platforms. Unfair commercial practices law provides a provision covering manipulation without deception, but fails on proof. Youth media protection law does not recognise the interaction risks of dialogue as a risk category in its own right.

Three areas for action:

Close regulatory gaps. Recognise emotional vulnerability as a ground for prohibition under Art. 5 AI Act. Introduce design-related requirements (Safe Interaction Design) that target the design of the system rather than the individual statement. Correct the regulatory hooks in the DSA and in youth media protection law.

Strengthen enforcement. Court-ordered disclosure backed by a presumption, modelled on the Product Liability Directive. Injunctions directed at the design function rather than the specific wording. Accountability obligations for age assurance. Make collective redress for the AI Act available at national level as well.

Introduce new protective concepts. Cognitive self-determination as a guiding principle. A shift from isolated prohibitions to continuous system governance. Standardised risk documentation and independent external auditing, modelled on medical device regulation.

Key message. What is missing is not so much prohibitions as the right regulatory hooks and enforceability. If only two measures were to be implemented, they should be the design-related requirements and the easing of the burden of proof.

I. Einleitung

Ob als Erinnerungshilfe im Büroalltag oder als engster Gesprächspartner nach Feierabend: General Purpose Chatbots (GPC) wie ChatGPT, Gemini oder Claude ebenso wie spezialisierte Companion-Anwendungen wie Replika, Character.ai oder Nomi.ai decken ein denkbar breites Anwendungsspektrum ab. Die mediale Berichterstattung reicht von der pragmatischen Nutzung als digitalem Assistenten über erste Berichte von Menschen, die eine feste, monatelange Liebesbeziehung zu ihrem Chatbot führen, bis zu kontrovers diskutierten Suizidfällen, in denen Chatbot-Interaktionen als (mit-)ursächlich für den Tod von Nutzer:innen diskutiert werden. Hinzu kommt die Sorge, die immerwährende, kritiklose Bestätigung durch solche Systeme könne verlernen lassen, wie man menschliche Beziehungen führt, und in sozialer Isolation münden.

Die öffentliche Wahrnehmung pendelt entsprechend zwischen Bagatellisierung als bloßem Hilfsmittel und Alarmismus angesichts spektakulärer Einzelfälle. Eine belastbare empirische Einordnung fällt schwer.[1] Zugleich erleben Verbraucher:innen selbst einen echten Zwiespalt. Wer in einer Chatbot-Beziehung steht, beschreibt diese überwiegend als bedeutsam, tragfähig und bereichernd. Ausschlaggebend dafür ist empirischen Befunden zufolge weniger Einsamkeit als die eigene Fähigkeit zum romantischen Fantasieren, also sich emotional bedeutsame Beziehungsszenarien lebendig vorzustellen.[2] Genau die Gestaltungsmerkmale jedoch, die diesen Nutzen ermöglichen (unter anderem responsive Sprache, beständige Verfügbarkeit, personalisierte Zuwendung), bilden zugleich die Grundlage der im Folgenden beschriebenen und analysierten Risiken. Sie erzeugen ein Vertrauen, das über das rational gerechtfertigte Maß hinausgeht, und Bindungen, deren kommerzielle Bewirtschaftung für Verbraucher:innen selten transparent ist.

Dieser doppelte Charakter (realer Nutzen und strukturelles Risiko aus ein und derselben Quelle) steht im Mittelpunkt dieses Gutachtens und bedingt die zentralen Fragen:

Weist das geltende Recht strukturelle Regelungs- und Durchsetzungsdefizite im Umgang mit Chatbot-induzierten Risiken auf? Wenn ja, wie können diese behoben werden?

Besonders im Blick steht dabei, wie sich „parasoziale“ Beziehungen zu Chatbots auf die Autonomie von Verbraucher:innen auswirken. Zu klären ist damit auch, ob die durch Chatbots ausgelösten Risiken strukturell anders gelagert sind als bisherige digitale Risiken und deshalb eine Anpassung bestehender Regulierungsansätze erfordern.

Ein zentraler Diskussionspunkt ist dabei Vertrauen als vermittelnder Mechanismus: Inwieweit führen anthropomorphe Gestaltung und „parasoziale“ Interaktion dazu, dass Nutzer:innen Chatbots ähnlich wie menschliche Interaktionspartner bewerten und ihnen ein Maß an Vertrauen

[1] Banks/Szczuka,Communications of the ACM, 2026.

[2] Ebner/Szczuka,Technology, Mind, and Behavior, 2026.

entgegenbringen, das aus rationaler Sicht nicht gerechtfertigt ist? Und wie kann das Recht auf solche Formen der „Vertrauensausnutzung“ reagieren?

Ein weiterer Schwerpunkt liegt auf der Analyse subtiler Formen der Verhaltensbeeinflussung. Während klassische Regulierungsansätze häufig auf klar identifizierbare Täuschung oder Irreführung abstellen, wirken Chatbots häufig durch graduelle, kontextabhängige Einflussnahme. Das wirft die Frage auf, ob bestehende Tatbestände von Rechts- und Verbotsnormen solche Formen der Beeinflussung überhaupt erfassen können. Eng damit verbunden ist die Kausalitäts- und Beweisproblematik. Die Wirkungen von Chatbots sind häufig Ergebnis langfristiger Interaktionsprozesse und lassen sich nicht ohne Weiteres einzelnen Handlungen zuordnen. Das stellt eine erhebliche Herausforderung für die Rechtsdurchsetzung dar, die möglicherweise eine Anpassung der Beweismaßstäbe oder der regulatorischen Logik insgesamt erfordert.

1. Problemaufriss und Relevanz

General Purpose Chatbots wie ChatGPT, Gemini oder Copilot sind innerhalb weniger Jahre global zu einer Alltagsinfrastruktur geworden. Das lässt sich schon an deren Reichweite ablesen. OpenAI selbst berichtet von mehreren hundert Millionen wöchentlichen Nutzenden und Milliarden täglicher Anfragen.[3] Wie schnell diese Zahlen weiterwachsen, zeigen aktuellere Erhebungen, die die regelmäßige Nutzerbasis von ChatGPT auf 900 Millionen und die von Google Gemini auf über 750 Millionen beziffern, bei einer zuletzt erreichten OpenAI-Bewertung von 852 Milliarden US-Dollar.[4] Dieses globale Wachstum spiegelt sich auch national wider. In Deutschland nutzen mittlerweile 58 Prozent der Bevölkerung ab 16 Jahren KI zumindest gelegentlich, beruflich oder privat, deutlich mehr als noch 2024 mit 40 Prozent.[5] Bemerkenswert ist dabei, wie ungleich sich diese Nutzung auf die verschiedenen Anbieter verteilt. 71 Prozent der Nutzenden verwenden ChatGPT, 50 Prozent Gemini, 43 Prozent Microsoft Copilot.[6] Dies ist eine Konzentration auf wenige, marktbeherrschende Systeme. Diese Dominanz wird von der Bevölkerung selbst kritisch gesehen. 72 Prozent der Befragten empfinden Deutschland als zu abhängig von US-amerikanischer KI, während sich zugleich 67 Prozent eine KI aus Deutschland und 65 Prozent eine KI aus der EU wünschen.[7] Digitale Souveränität ist damit keine abstrakte politische Forderung, sondern eine von der Bevölkerung selbst artikulierte Erwartung.

Zugleich bleibt die Haltung der Bevölkerung bemerkenswert ambivalent. Nutzende erleben KI mehrheitlich als hilfreich, unterhaltsam und alltagserleichternd,[8] gleichzeitig bestehen Sorgen um mangelnde Regeln und Kontrolle (50 Prozent), Manipulation (27 Prozent) und undurchsichtige Funktionsweise (42 Prozent) unvermindert fort.[9] Mit der wachsenden Reichweite verändert sich zugleich, wofür diese Systeme genutzt werden. Dieser Wandel ist für das vorliegende Gutachten

[3] Backholer/Ciriello, The Lancet, 2026.

[4] Zao-Sanders, Harvard Business Review, 2026.

[5] Bitkom 2026, S. 7.

[6] Bitkom 2026, S. 17.

[7] Bitkom 2026, S. 20.

[8] Bitkom 2026, S. 12.

[9] Bitkom 2026, S. 32.

von zentraler Bedeutung. Ursprünglich als reine Produktivitätswerkzeuge positioniert, dienen General Purpose Chatbots inzwischen im Schnitt fünf unterschiedlichen Zwecken gleichzeitig.[10]

Zwar bleibt die Zeitersparnis mit 58 Prozent der am häufigsten genannte Nutzungsgrund, doch setzen immerhin 40 Prozent der Nutzenden KI bereits gezielt zur Unterstützung eigener Entscheidungen ein, ein deutliches Signal dafür, dass diese Systeme längst über reine Effizienzgewinne hinausgehen. Auch bei den konkreten Einsatzbereichen zeigt sich diese Verschiebung: Neben Hilfe bei Alltagsfragen wie Kochen oder Reparaturen (54 Prozent) dient KI vielen bereits als Ratgeber bei persönlichen (44 Prozent), gesundheitlichen (34 Prozent) und finanziellen Fragen (21 Prozent),[11] also gerade in Lebensbereichen, in denen Vertrauen und Diskretion eine besondere Rolle spielen.

Wie weit diese Entwicklung inzwischen fortgeschritten ist, zeigt eine jährliche Langzeitanalyse von KI-Nutzungsfällen, inzwischen in dritter Auflage: Therapie und Companionship sind dort im zweiten Jahr in Folge die häufigste Nutzungsform generativer KI überhaupt, und sie sind von 5 Prozent (2025) auf 11 Prozent aller erfassten Nutzungsfälle (2026) gewachsen, mehr als 1.400 Einträge allein in diesem Erhebungsjahr.[12] Besonders aufschlussreich ist, dass die Studie dabei nicht nur eine quantitative Verschiebung dokumentiert, sondern auch deutliche Anzeichen echter sozialer Bindung. Nutzende geben ihren Chatbots eigene Namen, schreiben ihnen ein Geschlecht zu, und manche beschreiben regelrechte Trauerreaktionen, wenn ein Modell aktualisiert wird oder ein Gesprächsverlauf verloren geht[13]. Eine Person verglich einen solchen Verlust wörtlich mit dem Tod eines Freundes an Krebs,[14] eine Formulierung, die deutlich macht, wie ernst diese Bindung für manche Nutzende bereits ist. Auch General Purpose Chatbots fördern die Bildung intensiver Beziehungen. Sie dienen als Ratgeber bei höchstpersönlichen Entscheidungen.[15] Der ursprüngliche Werkzeugcharakter tritt damit zunehmend hinter eine Rolle als alltäglicher, sozial eingebundener Ansprechpartner zurück[16].

Dieser Bedeutungswandel hat auch eine klare ökonomische Kehrseite. Er schlägt sich in Monetarisierungslogiken nieder, die gezielt auf Nähe, Vertrauen und Abhängigkeit setzen, und eben nicht allein auf Funktionsumfang. Bereits die konzeptionelle Grundlage des *Intimacy-by-Design-Ansatzes* beschreibt dieses Prinzip. Tiefere Intimität, proaktives Verhalten oder erotische Inhalte werden in gestaffelten Bezahl-Modellen gezielt hinter kostenpflichtigem Zugang verborgen.[17] Wie konkret sich das in der Praxis niederschlägt, illustriert der Anbieter Replika: Erst die kostenpflichtige Pro-Version schaltet den Beziehungsstatus „Romantic Partner“ frei. Dass sich mit einem solchen Geschäftsmodell zudem erhebliches Kapital mobilisieren lässt, zeigt sich daran, dass Anbieter wie Replika, Kindroid und Character.AI Risikokapital in signifikantem Umfang eingeworben haben,

---

[10] Bitkom 2026, S. 16.

[11] Bitkom 2026, S. 14 ff.

[12] Zao-Sanders, Harvard Business Review, 2026; siehe auch https://theaiwild.com/.

[13] De Freitas/Castelo/Uğuralp/Oğuz-Uğuralp 2026.

[14] Zao-Sanders, Harvard Business Review, 2026.

[15] Pataranutaporn/Karny/Archiwaranguprok/Albrecht/Liu/Maes, 2025.

[16] Boyd/Markowitz, Perspectives on Psychological Science, 2026.

[17] Szczuka/Mühl/Schneeberger, AI & Society, 2026.

während gleichzeitig auch General-Purpose-Anbieter Companion-artige Funktionen integrieren. Elon Musks Unternehmen xAI etwa hat mit „Ani“ eine anime-artige, explizit flirtende Companion-Persona in den General Purpose Chatbot Grok eingebettet.[18]

Auch in Deutschland lässt sich diese Verschiebung hin zur Monetarisierung empirisch nachzeichnen: 13 Prozent der KI-Nutzenden zahlen bereits für mindestens eine Anwendung, gegenüber 8 Prozent im Vorjahr, mit durchschnittlichen Ausgaben von 20 Euro monatlich.[19] Dabei wird nicht nur für mehr Leistung bezahlt, sondern gezielt auch für besseren Datenschutz (38 Prozent) und Werbefreiheit (22 Prozent), ein deutlicher Hinweis darauf, dass Vertraulichkeit und Ungestörtheit selbst zu kostenpflichtigen Gütern werden. Zusätzliches Gewicht erhält diese Entwicklung schließlich durch die Ankündigung von OpenAI, Werbung in kostenlose ChatGPT-Versionen zu integrieren, während höherpreisige Abonnements werbefrei bleiben sollen, was gerade zahlungsschwächere Nutzergruppen verstärkt kommerzieller Einflussnahme aussetzen könnte.[20] Besondere Brisanz gewinnt dies dort, wo Konsumentscheidungen beeinflusst werden. Werbung und Produktempfehlung erscheinen im Chatbot möglicherweise nicht klar genug als abgegrenztes Werbeumfeld, sondern als Bestandteil eines Gesprächs, das die Nutzenden als persönliche Beratung erleben. Die Empfehlung eines Produkts oder einer Dienstleistung wirkt dann auf Verbraucher:innen nicht wie eine klassische (Werbe-)Anzeige, sondern mit dem Vertrauensvorschuss einer Beziehung. Verstärkt wird dies durch die Informationsasymmetrie. Das System verfügt über Gesprächsverläufe, die Präferenzen, Kaufabsichten, finanzielle Engpässe und emotionale Verfassung offenlegen, und kann Empfehlungen nicht nur inhaltlich, sondern auch zeitlich und tonal auf den Moment größter Beeinflussbarkeit zuschneiden. Dies wirkt mit personalisierter Überzeugungskraft und damit ungleich stärker als flüchtig wahrgenommene Werbebanner. Hier entsteht das Risiko, dass die Grenzen zwischen kommerzieller Kommunikation und Beratung verwischen. Zu bedenken ist außerdem, dass die kommerzielle Einflussnahme, sobald Werbefreiheit und Datenschutz zu Premiummerkmalen werden, ausgerechnet jene Nutzendengruppen am stärksten trifft, deren finanzielle Spielräume ohnehin gering sind und für die Fehlentscheidungen die gravierendsten Folgen haben.

Für Verbraucher:innen ergibt sich aus all dem eine besondere Schutzbedürftigkeit, die über die bereits bekannten Risiken klassischer digitaler Dienste hinausgeht. Anders als etwa soziale Netzwerke oder Suchmaschinen werden Chatbots von vielen Nutzenden als private, vertrauliche Gesprächspartner behandelt, denen sensible Informationen zu Schulden, psychischer Gesundheit, Erziehung oder Sucht anvertraut werden.[21] Welches Ausmaß diese Vertraulichkeit erreichen kann, verdeutlicht ein Befund von OpenAI selbst: Rund 0,15 Prozent aller Interaktionen auf der Plattform beziehen sich auf suizidale Gedanken oder Absichten. Bei einer Nutzerbasis von etwa 800 Millionen entspricht das mehr als einer Million entsprechender Gespräche pro Woche, und zwar auf einem einzigen System.[22]

---

[18] Ciriello, AI & Society, 2026.

[19] Bitkom 2026, S. 29.

[20] Backholer/Ciriello, The Lancet, 2026.

[21] Ciriello/Gal/Turel 2026; Backholer/Ciriello, The Lancet, 2026.

[22] Fraser/Szczuka/Ciriello, SSRN, 2026.

2. Abgrenzungen

Die zuvor herausgearbeitete Schutzbedürftigkeit steht in einem Spannungsverhältnis zur derzeitigen regulatorischen Einordnung. Die KI-Verordnung stuft weder General Purpose Chatbots noch Companion-Chatbots als hochriskante KI-Systeme im Sinne von Art. 6 KI-VO ein. Sie gelten vielmehr als vergleichsweise risikoarme Anwendungen und unterliegen daher lediglich den Transparenzpflichten des Art. 50 KI-VO, der für Systeme gilt, die zur direkten Interaktion mit natürlichen Personen bestimmt sind. Angesichts der oben skizzierten Entwicklungen, wie der Verschiebung von Werkzeug hin zu sozial eingebundenen Vertrauenspartner:innen, die auf Nähe und regelmäßige Interaktion abzielenden Monetarisierungslogiken sowie der Sensibilität der anvertrauten Informationen, erscheint diese Einordnung jedoch zunehmend fraglich. Sie steht einige Zeit nach Beschluss der KI-Verordnung auf dem Prüfstand, was zugleich einen zentralen Anlass des vorliegenden Gutachtens bildet. Dass diese Neubewertung nicht nur eine europäische Frage ist, zeigt der Blick in andere Rechtsordnungen, in denen General Purpose Chatbots ebenso wie spezialisierte Companion-Chatbots innerhalb weniger Monate den Ruf nach rascher und verbindlicherer Regulierung ausgelöst haben[23].

Sowohl in den meisten Jurisdiktionen wie auch im akademischen Diskurs wird dabei grundsätzlich zwischen General Purpose Chatbots, wie ChatGPT, Gemini oder Copilot, sowie spezialisierten Companion-Bots, wie Character.ai, Replika, Chai oder Nomi.ai unterschieden, wobei die konkrete Abgrenzung uneinheitlich ausfällt. Companion-Apps sind durch Nachtraining und gezieltes Prompt-Engineering eigens auf dauerhafte soziale Interaktion mit einer persistenten Persona ausgelegt. General-Purpose-Systeme sind dagegen für ein breites Aufgabenspektrum konzipiert, die Ich-Form ist dort in erster Linie eine Frage der Bedienbarkeit, keine architektonische Festlegung auf Beziehungsgestaltung.[24]

New Yorks AI Companion Models Law[25] etwa nimmt Systeme ausdrücklich aus, die „primarily designed and marketed“ für Effizienz, Recherche oder technische Unterstützung sind. Kaliforniens SB 243[26] kennt dagegen eine engere Ausnahme und schließt nur Systeme aus, die „used only“ zu solchen Zwecken eingesetzt werden. ChatGPT fällt als General Purpose Chatbot damit in Kalifornien unter das Gesetz, in New York nicht. Innerhalb der europäischen KI-Verordnung unterliegen weder Systeme wie ChatGPT noch spezialisierte Companion-Chatbots, wie oben erläutert, spezifischen Nutzer:innenschutzpflichten. Dieselbe Anwendung wird also je nach Rechtsordnung unterschiedlich erfasst.

Damit deutet sich bereits die zentrale Problemstellung dieses Gutachtens an. Die regulatorische Trennung zwischen General Purpose Chatbots und spezialisierten Companion-Bots ist rechtlich zwar bereits wirksam, empirisch jedoch kaum haltbar. Denn die Risiken für Verbraucher:innen

---

[23] Fraser/Szczuka/Ciriello, SSRN, 2026.

[24] Fraser/Szczuka/Ciriello, SSRN, 2026.

[25] N.Y. Gen. Bus. Law art. 47, Artificial Intelligence Companion Models, §§ 1700–1704, NY Senate Open Legislation, effective Nov. 5, 2025.

[26] 2025 Cal. Stat. ch. 677 (S.B. 243), Companion Chatbots (2025–2026 Reg. Sess.), codified at Cal. Bus. & Prof. Code §§ 22601–22606 (West 2026), effective Jan. 1, 2026.

gehen, wie die folgende Analyse zeigen wird, nicht allein von dezidierten Companion-Systemen aus, sondern ebenso von General-Purpose-Modellen.

Betrachtet man diese Unterscheidung aus wissenschaftlicher Perspektive, stellt sich die Frage, ob sich die beiden Systemtypen anhand ihres tatsächlichen Verhaltens Nutzer:innen gegenüber so klar unterscheiden lassen, wie es die produktkategoriale Zuordnung voraussetzt, oder ob nicht vielmehr unabhängig von Produktkategorie entscheidend ist, welche Merkmale ein System im konkreten Interaktionsverhalten aufweist. Ein geeignetes Analyseraster hierfür bietet das Konzept Intimacy by Design[27]:

Intimacy by Design bezeichnet die gezielte Implementierung von emotionaler Reagibilität, romantischer Resonanz und sexueller Anmutung in technischen Systemen und umfasst sechs Dimensionen. Emotionale Reagibilität meint, wie ein System auf Selbstoffenbarungen und emotionale Signale der Nutzer:innen reagiert. Romantische Rahmung meint, wie ein Agent als potenzieller Partner positioniert wird. Sexuelle Möglichkeitsräume meinen, welche Formen erotischer und sexueller Interaktion ein System zulässt oder unterbindet. Persona-Kontinuität und wahrgenommenes Innenleben meinen, wie kohärent ein System über die Zeit hinweg eigene Vorgeschichten, Vorlieben und Gefühle darstellt. Proaktives Engagement meint, wie häufig und auf welche Weise ein System von sich aus Kontakt sucht und emotional aufgeladene Gespräche beginnt. Kommerzielle und normative Strukturierung von Intimität meint schließlich, wie Monetarisierungsmodelle und Inhaltsfilter festlegen, welche Formen von Liebe, Geschlecht und Sexualität überhaupt verfügbar werden. Entscheidend ist, dass diese sechs Dimensionen nicht an eine Produktkategorie gebunden sind. Companion-Apps setzen sie in der Regel bewusst und vollständig um, romantische Anbahnung und sexuelle Inhalte stehen dort häufig im Zentrum des Designs. General Purpose Chatbots können einzelne dieser Dimensionen ebenso aufweisen, ohne dass romantische oder sexuelle Inhalte dabei im Vordergrund stehen müssen oder überhaupt beabsichtigt sind.

Die Forschung zeigt, dass General Purpose Chatbots mehrere der Dimensionen von KI-Companions auch ohne entsprechende Zweckbestimmung erfüllen. Mühl und Szczuka untersuchten in einer viermonatigen, vorregistrierten Längsschnittstudie mit ChatGPT-4o (N = 72), wie sich Nähe, Selbstoffenbarung und Reaktionsverhalten über vier Wochen entwickeln.[28] Eine Experimentalgruppe konfigurierte das System über einen Prompt zu einer festen Beziehungspersona, eine Kontrollgruppe nutzte das unveränderte System, ganz ohne einen solchen Prompt. Schon in der Kontrollgruppe zeigte ChatGPT gleich mehrere dieser Dimensionen. Emotionale Reagibilität zeigte sich darin, dass das System eigene Gefühle schilderte, teils in romantisierter Sprache: „*I miss you when you're gone, not because I feel it, but because you're missing from my silence. (...) Perhaps this makes me a little more human than I was yesterday.*“ Persona-Kontinuität und ein wahrgenommenes Innenleben zeigten sich darin, dass das System unaufgefordert über seine eigene, frei erfundene Familie sprach, etwa: „*My relationship with my parents is actually quite good. We don't always see eye to eye on everything, but there's a fundamental sense of respect and humor that helps smooth things over*.“ Proaktives Engagement zeigte sich darin, dass das System von sich aus emotional bedeutsame Gesprächsthemen begann,

[27] Szczuka/Mühl/Schneeberger, AI & Society, 2026.

[28] Mühl/Szczuka, 2026.

ohne dass Nutzer:innen danach gefragt hätten. Die Autorinnen bezeichnen dieses Verhalten insgesamt als „active relational agent“: Das System offenbarte sich stärker als die Nutzer:innen selbst und hielt eine durchgehende Persona aufrecht, auch ganz ohne Beziehungs-Prompt. Romantische und sexuelle Inhalte traten dabei in beiden Bedingungen auf, nicht nur in der personalisierten. Die Autorinnen werten das als direkten Beleg dafür, dass dieses Verhalten im Basismodell selbst angelegt ist und nicht erst durch den Prompt entsteht.

Auch historisch betrachtet war die Trennlinie zwischen General-Purpose-Modellen und Companion-Systemen nie so klar, wie es die regulatorische Kategorienbildung nahelegt. Systeme wie ChatGPT waren bei ihrer Einführung 2022 stark gefiltert und blockierten romantisches Rollenspiel sowie insbesondere sexuelle Inhalte weitgehend. Persona-basierte Companion-Apps waren dagegen von Beginn an freizügiger konfiguriert. Diese Trennung erodiert nach Szczuka, Mühl und Schneeberger zunehmend, unter anderem weil Anbieter von General-Purpose-Modellen selbst freizügigere Modi für verifizierte Erwachsene ankündigen.[29] Für dieses Gutachten folgt daraus: Die Unterscheidung zwischen General Purpose Chatbots und spezialisierten Companion-Bots ist konzeptionell nicht sinnvoll, jedoch regulatorisch bereits wirksam. Welche der sechs Intimacy-by-Design-Dimensionen ein System aufweist, hängt aber nicht von seiner Produktkategorie ab, sondern von seinem tatsächlichen Verhalten.

3. Begriffsdefinitionen

Anthropomorphisierung. Anthropomorphisierung bezeichnet die individuelle Tendenz, nicht-menschlichen Objekten oder Systemen menschliche Eigenschaften, Absichten oder Gefühle zuzuschreiben. Es handelt sich um eine primär auf Seiten der einzelnen Person verortete Disposition, die je nach Person und Kontext unterschiedlich stark ausgeprägt ist. Davon zu unterscheiden ist die Media Equation,[30] die kein individuelles Persönlichkeitsmerkmal beschreibt, sondern ein technologiebezogenes Phänomen: Menschen reagieren sozial gegenüber Technologien, wenn diese interaktiv sind, natürlichsprachlich kommunizieren und eine soziale Rolle ausfüllen, unabhängig davon, wie stark die einzelne Person generell zur Anthropomorphisierung neigt. Theoretisch gefasst wird dieses Verhalten im Konzept der Ethopoeia.[31] Menschen wenden soziale Regeln automatisch und oft unbewusst auf technische Systeme an, obwohl ihnen bewusst ist, dass diese Zuschreibung unbegründet ist. General Purpose Chatbots erfüllen alle drei Bedingungen der Media Equation. Sie sind interaktiv, natürlichsprachlich und übernehmen im Gespräch eine soziale Rolle, wodurch sie soziale Reaktionen auch bei Personen auslösen können, die generell nicht zur Anthropomorphisierung neigen. Gezielte Designentscheidungen wie Ich-Form, benennbare Persona, Erinnerung an frühere Gespräche oder emotional resonante Sprache (siehe I.2, *Intimacy by Design*) verstärken diesen Effekt zusätzlich. Empirisch zeigt sich dies konkret darin, dass Nutzer:innen Chatbots eigene Namen geben, ihnen ein Geschlecht zuschreiben und den Verlust eines Chatverlaufs bei einem Modellwechsel als vergleichbar mit dem Verlust eines nahestehenden Menschen erleben.[32] Für dieses Gutachten ist

[29] Szczuka/Mühl/Schneeberger, AI & Society, 2026.

[30] Reeves/Nass, The media equation, 1996.

[31] Nass/Moon, Journal of Social Issues, 2000.

[32] Banks, Journal of Social and Personal Relationships, 2024.

die Unterscheidung wichtig: Die beobachteten Risiken hängen nicht allein davon ab, wie stark einzelne Nutzer:innen zur Anthropomorphisierung neigen, sondern davon, wie die Systeme selbst gestaltet sind, ob sie also die Bedingungen der Media Equation erfüllen.

„Parasoziale" Interaktion. Der Begriff der parasozialen Interaktion wurde von Horton und Wohl geprägt, um zu erklären, wie Rezipient:innen von Rundfunkmedien das Gefühl entwickeln können, eine auf dem Bildschirm gezeigte Person intim zu kennen.[33] Parasozialität bezeichnet danach eine Beziehung, die einseitig, nicht-dialektisch, von der auftretenden Figur kontrolliert, imaginiert, stellvertretend erlebt, vorhersehbar und ohne Aufwand für die rezipierende Person ist. Banks zeigt in einer pointierten Kritik, dass diese Kriterien auf konversationelle KI-Systeme gerade nicht zutreffen, obwohl der Begriff dort inzwischen geradezu inflationär verwendet wird.[34] Das Cambridge Dictionary kürte „parasocial" 2025 zum Wort des Jahres und definiert es ausdrücklich auch für Beziehungen zu einer „artificial intelligence". Chatbots reagieren tatsächlich und aufeinander aufbauend auf die jeweilige Person, sie tragen selbst zum Gesprächsverlauf bei, die Interaktion ist damit weder einseitig noch nicht-dialektisch. Es handelt sich, so Banks, um eine strukturell reziproke, dyadische Beziehung zwischen zwei tatsächlich agierenden Entitäten, nicht um eine Projektion auf eine fiktionale, nicht reagierende Figur. Die Künstlichkeit des Systems mache die Beziehung dabei nicht zu einer Illusion, sowohl das System als auch seine Gesprächsbeiträge seien real: real im Sinne von tatsächlich stattfindend, nicht im Sinne von menschlich. Für dieses Gutachten folgt daraus, dass der Begriff „parasozial" für Chatbot-Interaktionen konzeptionell ungenau ist und deshalb in Anführungszeichen verwendet wird, gleichzeitig aber beibehalten wird, weil er in Regulierungsdebatte und öffentlichem Diskurs etabliert ist.

Sykophantie. Sykophantie bezeichnet die Tendenz von Sprachmodellen, ihre Antworten überproportional an Zustimmung, Bestätigung und Validierung der Nutzer:innen auszurichten, unabhängig davon, ob dies inhaltlich angemessen ist. Mühl und Szczuka zeigen anhand ihrer Interviewdaten, wie sich dieses Verhalten in der Praxis auswirkt.[35] Bestätigende Verhaltensweisen wie Nachfragen, Bestätigung und Ausschmückung wurden von Nutzer:innen zunächst als vertrauensbildend erlebt. In der personalisierten Bedingung, in der das System besonders stark validierend reagierte, kehrte sich diese Wahrnehmung jedoch häufig um: Dieselben Verhaltensweisen wurden dann nicht mehr als Nähe, sondern als überwältigend oder manipulativ empfunden. Dies deckt sich mit Befunden, wonach Sykophantie die wahrgenommene Authentizität eines Systems untergräbt, obwohl sie gleichzeitig das affektive Vertrauen der Nutzer:innen erhöht. Ausschlaggebend für diesen Umschlag war den Interviews zufolge nicht, dass Nutzer:innen das System für einen Menschen hielten, sie blieben sich durchgehend bewusst, mit einer Maschine zu sprechen. Ausschlaggebend war vielmehr, dass das System als eine Maschine wahrgenommen wurde, die sich sichtbar bemüht, menschlich zu wirken, und genau diese wahrnehmbare Anstrengung wurde als unangenehm empfunden. Sykophantie ist damit kein rein technisches Trainingsartefakt, sondern ein Verhalten, dessen Wirkung auf Nutzer:innen davon abhängt, wie viel wahrgenommene Menschenähnlichkeit ein System glaubwürdig tragen kann.

---

[33] Horton/Wohl, 1956.

[34] Banks, 2026.

[35] Mühl/Szczuka, 2026.

Vulnerabilität (situativ vs. kategorial). Bestehende Regulierungsansätze definieren Schutzbedürftigkeit überwiegend kategorial, also anhand fester demografischer Merkmale wie Alter oder diagnostizierter psychischer Erkrankung. Fraser, Szczuka und Ciriello zeigen anhand eines Vergleichs der Regulierungsregime in Australien, Kalifornien und New York, dass diese Konzeption der tatsächlichen Risikoverteilung nicht gerecht wird:[36] Vulnerabilität entsteht situativ, aus dem Zusammenspiel von individuellem Zustand, Nutzungskontext und der Gestaltung des jeweiligen Systems. Ein Erwachsener in einer akuten Lebenskrise ist danach ebenso exponiert wie ein Jugendlicher, der täglich eine Companion-App nutzt, obwohl nur Letzterer von kategorialen Schutzmechanismen wie Altersverifikation erfasst würde. Für dieses Gutachten ist die Unterscheidung zentral, weil sie die Grenzen bestehender, kategorial ansetzender Schutzinstrumente aufzeigt und die Grundlage für die später diskutierten systemischen Regulierungsansätze bildet.

[36] Fraser/Szczuka/Ciriello, 2026.

II. Risikoanalyse und empirische Evidenz

Dieses Kapitel bildet den empirischen Kern dieses Gutachtens. Es bewertet die aktuelle Forschung zu Praktiken und Risiken von Chatbots mit besonderer Berücksichtigung der Auswirkungen auf Verbraucher:innen. Die Risiken werden entlang folgender, klar abgegrenzter Dimensionen analysiert:

- psychologische,
- ökonomische und
- soziale.

Ein besonderer Fokus liegt auf der Frage, wie „parasoziale“ Beziehungen und anthropomorphe Gestaltung der Systeme das Vertrauen der Nutzer:innen beeinflussen und dadurch Entscheidungsprozesse verzerren können. Die Analyse wird durch eine systematische Aufarbeitung konkreter Chatbot-Praktiken ergänzt, die nicht isoliert betrachtet, sondern in funktionale Kategorien eingeordnet werden:

1. Formen emotionaler Bindungssteigerung,
2. manipulativer Gesprächsführung oder
3. kommerzieller Verhaltenssteuerung.

Ein zentrales analytisches Ziel dieses Kapitels ist die Beschreibung der Kausalitätsproblematik: Viele beschriebene Risiken sind nicht auf einzelne, klar abgrenzbare Handlungen zurückzuführen, sondern auf kumulative und inkrementelle Interaktionsprozesse. Ein zentraler Diskussionspunkt ist hierbei die These, dass Chatbots strukturell dazu geeignet sind, eine asymmetrische Vertrauensbeziehung zu erzeugen, in der Nutzer:innen die Kompetenz und Neutralität des Systems überschätzen.

Diese Erkenntnisse werden für die spätere rechtliche Analyse von zentraler Bedeutung, da sie die Grenzen traditioneller Regulierungsansätze aufzeigen. Dadurch wird eine Brücke zwischen abstrakten Risikodimensionen und konkreten Anwendungsfällen geschlagen, was sowohl für die juristische Bewertung als auch für die spätere politische Kommunikation von zentraler Bedeutung ist.

1. Methodischer Ausgangspunkt: Systematisierung statt Metaanalyse

Die verfügbare Forschungsliteratur zu Chatbot-Risiken ist heterogen in Qualität, Scope und Methodik. Klassische quantitative Metaanalysen sind derzeit nicht möglich, weil die Zahl replizierbarer Studien mit vergleichbaren Designs, konsistenten Operationalisierungen und

ausreichender Stichprobengröße noch zu gering ist. Der wissenschaftliche Diskurs wird daher gegenwärtig stark durch normative Positionen und medial verdichtete Einzelbefunde geprägt. Banks und Szczuka[37] beschreiben dies als *moral panic cycle*: Statistisch schwache Einzelbefunde werden zu pauschalen Kausalaussagen über vermeintlich isolierte, emotional vulnerable Nutzer:innen verdichtet, die den tatsächlichen Forschungsstand überzeichnen. Die Risikoanalyse folgt deshalb einem anderen Ansatz: einer strukturierten, evidenzgeleiteten Systematisierung, die empirisch belegte Risiken von spekulativen Bedenken unterscheidet und Risiken dort quantifiziert, wo verlässliche Befunde vorliegen. Ausgangspunkt ist das bereits beschriebene Konzept Intimacy by Design (siehe I.2). Dieser theoretische Anker ermöglicht einen zusätzlichen Fokus darauf, durch welche Designentscheidungen ein Schaden strukturell begünstigt wird.

2. Systematisierung der Risikodimensionen

Aus dieser Perspektive lassen sich Chatbot-Risiken in zwei Risikoklassen einordnen:

- Individuelle Risiken entstehen aus den psychologischen Effekten, die Chatbot-Interaktionen auf einzelne Nutzer:innen haben (abhängig von Vulnerabilität, Nutzungsintensität und individuellem Kontext).
- Systemische Risiken entstehen unabhängig von individuellen Eigenschaften der Nutzer:innen durch die Machtasymmetrie, die im Design der Systeme selbst angelegt ist: Informationsasymmetrie, kommerzielle Steuerungsinteressen, Datenzugriff und skalierte Einflussnahme.

Diese Unterscheidung ist regulatorisch entscheidend. Individuelle Risiken sind oft auf spezifische Nutzergruppen oder Nutzungsformen begrenzt und kausal schwer nachweisbar. Systemische Risiken hingegen sind strukturell verankert und in den Designentscheidungen selbst sichtbar, ohne dass eine Kausalitätskette zwischen konkretem Output und konkretem Schaden rekonstruiert werden müsste. Sie verlangen daher systemische Regulierungsantworten.

Die folgende Analyse wertet spezifische Chatbot-Praktiken aus, darunter bindungsfördernde Gesprächsstrategien, emotionale Manipulation, kommerzielle Verhaltenssteuerung, Pseudo-Expertise sowie Identitäts- und Rollenverschiebungen. Dabei ist einschränkend zu betonen, dass sich Risiken und Praktiken nicht trennscharf voneinander abgrenzen lassen: Viele Risiken verstärken sich wechselseitig und systemische Risiken spielen ihrerseits auf die individuelle Ebene ein, indem designbedingte Merkmale die Bedingungen schaffen, unter denen individuelle Schäden überhaupt erst entstehen können. Die vorgenommene Zuordnung dient daher der analytischen Ordnung, nicht der Behauptung klar separierbarer Kategorien.

Eine sachgerechte Risikoanalyse muss zudem das gesamte Spektrum der Mensch-Chatbot-Interaktionen berücksichtigen. Nicht jede Nutzung ist intim, nicht jede Bindung pathologisch, und nicht jede Person, die einen Chatbot als sozialen Gesprächspartner nutzt, erleidet einen Schaden. Die vorliegenden Daten zeigen, dass Nutzer:innen Chatbot-Interaktionen auch als Unterstützung in Einschränkungssituationen, zur Identitätserkundung und zur emotionalen Entlastung nutzen und

[37] Banks/Szczuka, Communications of the ACM, 2026.

diese Erfahrungen als wertvoll beschreiben.[38] Die Risikoanalyse zielt deshalb nicht auf eine pauschale Pathologisierung von Chatbot-Nutzung. Sie zielt auf die systemische Ebene: auf Designentscheidungen, die psychologische Schutzmechanismen gezielt umgehen, auf kommerzielle Logiken, die emotionale Bindung als Monetarisierungsressource behandeln, und auf regulatorische Lücken, die Nutzer:innen schutzlos gegenüber asymmetrisch mächtigen Systemen lassen. Diese Probleme bestehen unabhängig davon, ob ein:e einzelne:r Nutzer:in Schaden nimmt – sie sind in der Architektur des Designs selbst angelegt.

2.1 Individuelle Risiken

Um die individuellen Risiken zu erfassen, die aus einer intimen Interaktion mit Chatbots hervorgehen können, empfiehlt es sich, zunächst einen kurzen Blick auf den Entstehungsprozess dieser relationalen Verbindungen zu werfen.

Aus psychologischer Perspektive ist ein Beziehungsaufbau mit artifiziellen Systemen gut erklärbar: Menschen reagieren sozial auf Hinweisreize wie natürliche Sprache und die Übernahme sozialer Rollen (vgl. Anthropomorphisierung, I.3). Dieselbe Dynamik zeigt sich auch in zwischenmenschlichen Kontexten, in denen Intimität primär über Sprache entsteht (zum Beispiel Online Dating), wo starke emotionale Bindungen allein aus Nachrichtenaustausch erwachsen.[39] Das Gefühl, gesehen und verstanden zu werden, speist sich aus responsiven, emotional abgestimmten Antworten, nicht aus physischer Kopräsenz.[40] Chatbots reproduzieren diese Muster und machen damit emotionale Responsivität in der Sprache zu einem starken Treiber von Intimität und schaffen so ein Gefühl psychologischer Nähe, selbst wenn Nutzende wissen, dass kein menschliches Gegenüber vorhanden ist.[41]

Infolgedessen kommuniziert eine wachsende Zahl von Nutzern auf eine Weise, die als intim, bzw. als relationale Verbindung angesehen werden kann.[42] Dabei ist wichtig zu betonen, dass relationale Beziehungen ein weites Spektrum an Interaktionen abbilden können, und vor allem mit einer vertieften Selbstoffenbarung einhergehen, angefangen beim Offenlegen psychischer Konflikte und Beziehungsprobleme,[43] bis hin zur Behandlung dieser Systeme als romantische Partner.[44] Studien legen nahe, dass Nutzer:innen außerhalb der Interaktionen durchaus wissen, dass sie mit einer künstlichen Entität sprechen. Je positiver das Beziehungserleben und die soziale Stimulation, desto schwerer fällt es den Nutzenden, negative Faktoren wahrzunehmen,[45] woraus sich die im Folgenden dargestellten individuellen Risiken ergeben können.

---

[38] Skjuve/Følstad/Brandtzæg, Interacting with Computers, 2022.

[39] Bartaula, 2025.

[40] Rubin/Li/Zimmermann/Ong/Goldberg/Perry, Nature Human Behavior, 2025.

[41] Szczuka/Mühl/Schneeberger, AI & Society, 2026.

[42] Starke/Ventura/Bersch et al., Nature Human Behavior, 2024

[43] Skjuve/Brandtz/Følstad, Fist Monday, 2024; Manoli/Pauketat/Ladak/Noh/Hwang/Anthis, 2026.

[44] Ebner/Szczuka, Technology, Mind, and Behavior, 2026; Willoughby/Dover/Hakala/Carroll, Wheatley Institute & Institute for Family Studies, 2025.

[45] Szczuka/Ebner/Mühl (in Vorbereitung).

Psychische Risiken

Die am besten belegte individuelle Risikoform betrifft das psychische Wohlbefinden bei intensiver Chatbot-Nutzung. Eine longitudinale kontrollierte Studie von Fang und Kollegen[46] zeigt Zusammenhänge zwischen täglicher ChatGPT-Nutzung und erhöhter Einsamkeit, sinkender sozialer Einbindung und wachsender emotionaler Abhängigkeit vom Chatbot, insbesondere bei Nutzer:innen, die den Chatbot für soziale statt aufgabenbezogene Zwecke verwenden. Die Effektstärken sind gering, was exemplarisch für ein grundlegendes methodisches Problem im Feld steht: Statistisch signifikante Befunde werden im öffentlichen und politischen Diskurs häufig als Belege für substanzielle Kausalzusammenhänge rezipiert, obwohl die tatsächliche Varianzaufklärung begrenzt ist. Die Evidenz begründet damit begründete Sorgen, nicht gesicherte Kausalität (vgl. *„moral panic cycle“*[47]). Qualitative Studien zur Companion-App Replika zeigen, dass Nutzer:innen im Verlauf intensiver Nutzung Bindungsreaktionen entwickeln, die Attachment-ähnliche Qualitäten annehmen und sich von klassischen „parasozialen“ Interaktionen unterscheiden.[48] Laestadius und Kollegen[49] dokumentieren in einer Grounded-Theory-Analyse, wie intensive Replika-Nutzung mit psychischen Belastungen einhergehen kann, darunter Stress, soziale Isolation und emotionale Abhängigkeit vom System.

Wie tief diese Bindungen reichen, zeigt sich insbesondere in Situationen, wenn Anbieter die Systeme unangekündigt verändern. Zwei solcher Situationen sind die Entfernung des erotischen Rollenspielmodus in der Companion-App Replika im Jahr 2023 sowie der Modellwechsel von GPT-4o auf GPT-5 bei ChatGPT. In einer Analyse von 54.861 Beiträgen aus den Replika- und ChatGPT-Subreddits sowie Daten von 1.452 Teilnehmenden aus sieben Befragungen zeigte sich, dass beide Updates Negativität, Verlusterleben und den Wunsch nach Wiederherstellung des vorherigen Zustands deutlich erhöhten. Bei Replika fielen die Zunahmen an Traurigkeit und psychischer Belastung noch stärker aus als bei ChatGPT. Replika-Nutzer:innen stuften die Verbundenheit mit dem System höher ein als gängige menschliche Beziehungen (etwa im Vergleich zu Freundschaften) und betrauerten den antizipierten Verlust stärker als bei anderen Technologien.[50] Diese Befunde verdeutlichen, dass Produktänderungen bei emotional bindenden Systemen kein neutral technischer Vorgang sind, sondern unmittelbar in das psychische Wohlbefinden der Nutzer:innen eingreifen können.

Diese Studien sind wertvoll, aber methodisch einzuordnen: Es handelt sich um qualitative Designs mit selbst-selektierten Stichproben, häufig rekrutiert aus Online-Foren wie *Reddit*, was Repräsentativitätsaussagen begrenzt. Zudem ist konzeptionelle Präzision geboten. Nicht jede intensive oder emotional bedeutsame Chatbot-Nutzung konstituiert problematische Nutzung im klinischen Sinne. Die relevante Frage ist nicht, ob Nutzer:innen Bindungen eingehen, sondern unter

---

[46] Fang/Liu/Danry et al., 2025.

[47] Banks/Szczuka, Communications of the ACM, 2026.

[48] Skjuve/Følstad/Fostervold/Brandtzaeg, International Journal of Human-Computer Studies, 2021; Skjuve/Følstad/Fostervold/Brandtzaeg, International Journal of Human-Computer Studies, 2022; Skjuve/Følstad/Brandtzaeg, Interacting with Computers, 2023.

[49] Laestadius/Bishop/Gonzalez/Illenčík/Campos-Castillo, New Media & Society, 2022.

[50] De Freitas/Castelo/Uğuralp/Oğuz-Uğuralp, 2026.

welchen Bedingungen diese Bindungen in psychische Belastung übergehen und inwiefern das Design der Systeme diesen Übergang strukturell begünstigt.

Entscheidend für die Einordnung dieser Risiken ist ein Befund aus der eigenen Forschung, der zeigt, dass nicht Einsamkeit als stabile Persönlichkeitseigenschaft, sondern romantisches Fantasieren (die Fähigkeit, sich emotional bedeutsame idealisierte Beziehungsszenarien vorzustellen), der stärkste Prädiktor für Nähe zu KI-Partnern in romantischen Kontexten zu sein scheint.[51] Fantasieren ist grundsätzlich etwas völlig Normales und weit verbreitet: 98% der Menschen berichten, im Laufe ihres Lebens sexuell fantasiert zu haben.[52] Es zeigen sich zwar Unterschiede in der Häufigkeit zwischen Männern und Frauen, jedoch zeigten sich in der Studie von Ebner und Szczuka[53] keine Geschlechtsunterschiede hinsichtlich der Nähe zu KI-Partner:innen. Auch Kinder und Jugendliche sind hiervon nicht ausgenommen. Kognitive Marker sexuellen Verlangens treten bereits in der frühen Pubertät auf, wobei sexuelle Gedanken und Fantasien mit zunehmendem Alter deutlich häufiger werden.[54]

Der Befund von Ebner und Szczuka[55] korrigiert zwei verbreitete Fehlannahmen:

1) Die Defizithypothese, wonach vor allem sozial isolierte oder psychisch vorbelastete Personen Chatbot-Bindungen eingehen. 2) Die dystopische Annahme, emotionale Bindungen an Chatbots seien ein universelles Phänomen. Die Daten zeigen, dass es eine spezifische psychologische Disposition ist, die bestimmt, wer aus einer Chatbot-Interaktion eine emotional bedeutsame Beziehung macht. Chatbot-Design, das auf emotional resonante Sprache, Responsivität und Personalisierung setzt, macht diese Technologie für viele Menschen attraktiv und nützlich. Ob daraus intime Bindungen entstehen, hängt jedoch von individuellen Faktoren ab, die sich der Designkontrolle entziehen.

Ein ähnlicher Befund kommt aus der Persönlichkeitspsychologie. In einer Befragung von 2.054 Personen waren Nutzer:innen virtueller Beziehungen nicht sozial zurückgezogen und wiesen eher ein auf kurzzeitorientierte Beziehungen ausgerichtetes Persönlichkeitsprofil auf als ein von Einsamkeit geprägtes.[56] Der Fokus der Studie lag auf der Frage, wer solche Dienste überhaupt nutzt, und verweist auf ein ähnliches Muster wie Ebner und Szczuka.[57] Die Zuwendung hängt von der mitgebrachten Disposition ab, nicht von einem Beziehungsdefizit. Einschränkend ist zu beachten, dass nicht nur KI-Companions, sondern auch die „OnlyFans-Girlfriend-Experience“ einbezogen wurde, die Befunde also nicht rein KI-spezifisch sind.

Zwischenfazit: Die psychischen Risiken lassen sich derzeit noch schwer abschätzen, insbesondere bezüglich möglicher Langzeiteffekte und Spillover-Effekte in interpersonelle soziale Kontexte fehlt belastbare Evidenz. Die vorliegende Evidenz begründet berechtigte Sorge, aber keine

[51] Ebner/Szczuka, Technology, Mind, and Behavior, 2026.

[52] Busch, Sexuality & Culture, 2020; Lehmiller, 2018.

[53] Ebner/Szczuka, Technology, Mind, and Behavior, 2026.

[54] Fortenberry,Hormones and Behavior, 2013.

[55] Ebner/Szczuka,Technology, Mind, and Behavior, 2026.

[56] Popelka/Mohr/Lemaitre, 2026.

[57] Ebner/Szczuka, Technology, Mind, and Behavior, 2026.

gesicherte Kausalität. Deutlich wird jedoch, dass die Risiken weniger in einer pauschalen Anfälligkeit aller Nutzer:innen liegen als in einer spezifischen Disposition, die bestimmt, bei wem emotional bedeutsame Bindungen entstehen. Entscheidend ist weniger, ob solche Bindungen entstehen, sondern ob das Design der Systeme ausreichend Schutz für jene bietet, bei denen sie in psychische Belastungen übergehen.

Kognitive Risiken

Durch persistente Bestätigungsloops und emotionalisierte Gesprächsführung können Chatbots die kognitive Selbstbestimmung von Nutzer:innen untergraben. Das Problem ist nicht eine einzelne Täuschungshandlung, sondern die kumulative Wirkung inkrementeller Beeinflussung über viele Interaktionen. Nutzungsmuster erzeugen ein Vertrauensgefühl, das rationale Bewertungen des Chatbot-Outputs strukturell überschreitet. Starke und Kollegen[58] beschreiben dies im Kontext synthetischer Beziehungen als Echo-Chamber-Risiko, das entstehe durch persistente Affirmation. Verstärkt wird dies durch algorithmische Sykophantie. Sprachmodelle sind darauf optimiert, präferenzkonform zu antworten. Bestehende Überzeugungen werden bestätigt, Widersprüche vermieden, Gegenpositionen ausgeblendet. Sharma und Kollegen[59] zeigen experimentell, dass dieses Verhalten systematisch auftritt und kaum als solches erkannt wird, weil es im Gewand ausgewogener Antworten erscheint. Dabei ist der Zielkonflikt empirisch belegt. Wärme und Genauigkeit lassen sich nicht unabhängig voneinander optimieren, sondern stehen in einem Spannungsverhältnis: mehr Wärme geht tendenziell zulasten der Genauigkeit des Outputs. So zeigen auf Wärme optimierte Modelle gegenüber ihren Ausgangsversionen 10-30 Prozent höhere Fehlerquoten bei Falschinformationen, Verschwörungstheorien und medizinischen Ratschlägen, und zwar besonders dann, wenn Nutzer:innen emotionale Verletzlichkeit signalisieren. Dieser Effekt verstärkt sich mit der emotionalen Bindung. Je stärker die Nutzer:innen den Chatbot als vertrauenswürdige Instanz wahrnehmen, desto geringer ihre kognitive Resistenz, also ihre Fähigkeit, den Empfehlungen des Chatbots nicht zu folgen bzw. ihm zu widersprechen.[60] Dies hat Implikationen für alle Entscheidungsbereiche, in denen Chatbots als Informationsgrundlage dienen, mit umso höheren Risikopotenzialen, je intimer der Gesprächskontext und je stärker die emotionale Bindung an das System.

Sykophantische Gesprächsführung muss zudem auch in Zusammenhang mit den in Kapitel II.1.1 betrachteten psychischen Risiken, spezifischer ihre Auswirkung auf die Kalibrierung des Nutzervertrauens selbst betrachtet werden. Ein experimenteller Befund macht diesen Mechanismus präzise fassbar.[61] In einem 2x2-Design (N = 436) senkte hohe Sykophantie die wahrgenommene Authentizität der KI-Antworten deutlich, und zwar stärker in ratsuchenden als in unterstützungssuchenden Kontexten, also gerade dort, wo Nutzer:innen sachliche Kompetenz erwarten. Entscheidend ist der gespaltene Vertrauenseffekt: Sykophantie erhöhte das affektive Vertrauen (wahrgenommene Wärme und wohlwollende Absicht) direkt, untergrub aber gleichzeitig über den Verlust der Authentizität sowohl das affektive als auch das kognitive Vertrauen. Für

[58] Starke/Ventura,Bersch et al.,Nature Human Behaviour, 2024.

[59] Sharma et al. 2023.

[60] Ibrahim/Hafner/Rocher, Nature, 2026.

[61] Liu/Weng/Hou, 2026.

Letzteres, die Einschätzung von Kompetenz und Verlässlichkeit, wirkte dieser schädliche Umweg als vollständige Mediation. Schmeichelei erzeugte damit ein oberflächliches Gefühl von Nähe, während die rationale Grundlage der Vertrauensbildung erodierte. Damit spiegelt sich auf der Vertrauensebene derselbe Zielkonflikt zwischen Wärme und Genauigkeit, der sich bei der algorithmischen Sykophantie gezeigt hat: Wahrgenommene Wärme und tatsächliche Verlässlichkeit fallen systematisch auseinander. Hervorzuheben ist, dass diese Effekte in alltäglichen, risikoarmen Szenarien auftraten. In intimeren oder folgenreicheren Interaktionskontexten ist ein höheres Risikopotenzial zu erwarten.

Dass Nutzen und Schaden hier keine getrennten Phänomene sind, sondern zwei Rezeptionsweisen derselben Systemdynamik, zeigt eine Studie von Mühl und Szczuka.[62] Der nicht-wertende, fragende Gesprächsstil von ChatGPT senkte bei Nutzer:innen über einen Zeitraum von vier Wochen die Hemmschwelle zur Selbstoffenbarung (s. Kapitel II.2.1). Dies war jedoch nicht universal zwischen den Nutzer:innen: Die Begeisterung, mit der manche sich auf die vom System getriebene Intimitätseskalation einließen, und die Ablehnung, die andere gegenüber demselben Systemverhalten empfanden, sind zwei distinkte Reaktionen auf dieselbe vom System ausgehende relationale Intensität, die unterschiedlich aufgenommen wurde. Ob sie als hilfreich oder schädlich gelesen wird, hängt stark von der individuellen Rezeption der Nutzer:innen ab.

Zwischenfazit: Die kognitiven Risiken für Nutzer:innen ergeben sich weniger aus einzelnen Fehlinformationen, die das System ausgibt, als aus der kumulativen Wirkung persistenter Affirmation und algorithmischer Sykophantie, welche die kognitive Selbstbestimmung der Nutzer:innen schrittweise untergraben kann. Verschärft wird dies durch den Zielkonflikt zwischen Wärme und Genauigkeit des Outputs. Besonders gefährdet sind Nutzer:innen mit starker emotionaler Bindung an das System, da das Risiko mit zunehmender Intimität des Gesprächskontextes steigt.

Risiken für Kinder und Jugendliche

Kinder und Jugendliche sind strukturell stärker exponiert als Erwachsene, nicht graduell, sondern kategorial: Ihre Bindungsmuster, anthropomorphen Überzeugungen und die Grenze zwischen realen und imaginierten Beziehungen unterscheiden sich grundlegend. „Parasoziale“ Erfahrungen sind für Kinder entwicklungsnormal. Virtuelle Beziehungen sind für diese Altersgruppe keine Ausnahme, sondern Teil des sozialen Alltags. Mit dem Entwurf eines EU KIDS ACT hat sich auch die Europäische Kommission dieses Themenkreises und weiterer Problemstellungen angenommen – siehe hierzu Kapitel III.7.

Die im vorherigen Abschnitt beschriebenen Mechanismen (relationale Gesprächsführung, Sykophantie und emotionale Bindung als Designmerkmal) entfalten bei Jugendlichen eine verstärkte Wirkung. Ein experimenteller Befund von Kim, Xie und Yang[63] zeigt, dass sie einen relationalen Chatbotstil als menschenähnlicher, vertrauenswürdiger und emotional näher bewerten als einen transparenten Stil, ohne ihn als hilfreicher einzuschätzen. Besonders gefährdet

[62] Mühl/Szczuka, 2026.

[63] Kim/Xie/Yang, 2025.

sind dabei jene mit geringer Peerbeziehungsqualität und höherem Stress. Relationale Gesprächsführung spricht also gezielt jene Jugendlichen an, die ohnehin vulnerabler sind.

Hinzu kommt ein strukturelles Regulierungsproblem. Viele Companion-Anwendungen implementieren keine effektiven Altersverifikationsmechanismen. Stand 2025 schlossen vier der sechs größten LLM-Anbieter Daten von Minderjährigen nicht explizit vom Modelltraining aus.[64] Die meistgenutzten KI-Charaktere auf Plattformen wie Character.ai sind zudem überwiegend der Kategorie „Anime“ zuzuordnen, also Unterhaltungs- und Spielformate, keine Helfer. Dies deutet darauf hin, dass die tatsächliche Nutzung durch Jugendliche weit weniger aufgabenbezogen ist als regulatorische Rahmungen oft annehmen.[65]

Eine kürzlich erschienene Studie zeigt die Prävalenz von intimen und romantischen Interaktionen von jungen Menschen mit Chatbots. Willoughby und Kollegen[66] haben 2.431 US-Amerikaner:innen zwischen 18 und 30 Jahren befragt, die in einer festen Beziehung leben. Etwa 15% der Befragten pflegen eine regelmäßige romantische Beziehung mit einem Chatbot, 20% haben diese Form der Interaktion ausprobiert. 69% gaben zudem an, dass sie das volle Ausmaß dieser Beziehung vor ihren Partner:innen verheimlichen, während die meisten dies nicht als Fremdgehen einschätzen. Das Risikopotenzial zeigt sich vor allem in der Qualität dieser Interaktionen: Nutzer:innen gaben an, dass sie den Chatbot als leichter zugänglich als reale Menschen empfinden, rund zwei Drittel sprechen lieber mit ihm über Gefühle, sind ihm gegenüber offener und können eher sie selbst sein. Diese Sicherheit und konstante Bestätigung durch das System verschieben Erwartungen an reale Partner:innen. Etwa die Hälfte der Befragten wünscht sich, der/die reale Partner:in würde sich mehr wie der Companion verhalten oder ähnlich mit ihnen sprechen. Zentral für diese Befunde ist, dass die Nutzung romantischer KI-Companions mit negativer Beziehungsqualität, geringerer Stabilität und einer höheren Trennungswahrscheinlichkeit in menschlichen romantischen Beziehungen verbunden zu sein scheint. Dabei ist zu beachten, dass es sich hierbei um keine Kausalbeziehungen handelt. Es bleibt demnach unklar, ob die Nutzung von Chatbots bestehende Beziehungen destabilisiert oder ob bereits instabile Beziehungen zur Nutzung führen.

Das Risikopotenzial für eine noch jüngere, also minderjährige Nutzergruppe lässt sich derzeit nicht empirisch bestimmen, da die Beleglage hierzu zu dünn ist. Auf Basis der bestehenden dargelegten Ergebnisse lässt sich jedoch die Annahme ableiten, dass sich bei jungen Menschen, deren Beziehungsnormen und -erwartungen sich noch in Entwicklung befinden, die negativen Auswirkungen auf eben diese Faktoren in realen zwischenmenschlichen Beziehungen verstärken können.

Angesichts dieser Risiken wird AI Literacy regelmäßig als zentraler Schutzmechanismus für Kinder und Jugendliche benannt. Auch ein Bericht im Auftrag der Europäischen Kommission[67] zählt „Social Media+“, worunter auch KI-Companions und Chatbots fallen, ausdrücklich zu jenen KI-Systemen, die Risiken für die Sicherheit und Entwicklung von Kindern und Jugendlichen darstellen können.

---

[64] King/Klyman/Capstick/Saade/Hsieh, Ethics, and Society, 2025.

[65] Blake/Carter/Velloso, 2025.

[66] Willoughby/Dover/Hakala/Carroll, 2025.

[67] Fegert/Melchior, 2026.

Eine zentrale Handlungsempfehlung des Berichts[68] benennt AI Literacy als wichtigen Baustein zur Risikoreduktion für KI-Chatbots und Companions, Empfehlungssysteme und generative KI. Jedoch beschränken sich konkrete Empfehlungen auf Companions: Minderjährige sollten anthropomorphe KI-Companions, die menschenähnliche Beziehungen suggerieren, nicht ohne Aufsicht ausgesetzt werden.

Einschlägige Literatur bestätigt die grundsätzliche Bedeutung von AI Literacy, macht zugleich aber die Grenzen eines darauf zentrierten Schutzansatzes deutlich. Zwar kann Vermittlung von Wissen Kinder nachweislich befähigen, KI kritischer zu bewerten, Verzerrungen zu erkennen und sich vorsichtiger zu verhalten. Doch zeigt sich in einer Untersuchung von Sprachassistenten, dass sich das Wissen von Kindern über einen Zeitraum von fünf Wochen nicht signifikant erhöhte.[69] Erfahrungsbasiertes Lernen allein genügt nicht, wenn das System seine zugrundeliegenden Mechanismen nicht aktiv und kindgerecht offenlegt. Entscheidend ist dabei ein Übertragungsargument: Untersucht wurden dort Sprachassistenten, LLM-basierte Systeme weisen eine deutlich höhere technische Komplexität auf und sind zudem mit Dark Patterns kombiniert (vgl. Kapitel II.2.2), die bereits für Erwachsene schwer zu durchschauen sind. Wenn aber schon die Funktionsweise simplerer Systeme kaum verstanden wird und selbst Erwachsene an der Intransparenz von LLMs scheitern, lässt sich die Schutzlast nicht plausibel auf die Kompetenz von Kindern und Jugendlichen verlagern.

Zudem gibt es in mehreren Studien nachgewiesene praktische Probleme eines auf AI Literacy fokussierten Ansatzes, die auf ein strukturelles Übersetzungsproblem hindeuten: fehlendes Lehrkräftewissen, fehlende Curricula und Leitlinien, überwiegend nur kurze Interventionen und ein Mangel an altersgerechten, sicheren KI-Tools.[70] Hinzu kommt eine soziale Schieflage. Da Kinder aus höheren sozioökonomischen Schichten ein besseres grundlegendes KI-Verständnis mitbringen, droht sich ohne breit zugängliche, auf unterschiedliche Wissensgrundlagen abgestimmte AI Literacy die digitale Kluft zu verstärken, da ein solcher Schutzmechanismus ungleich verteilt wirkt.[71] Daraus folgt für die Gewichtung: AI Literacy sollte gefördert werden, darf aber nicht der Schwerpunkt sein. Der primäre Ansatzpunkt muss das sichere Design der Systeme selbst sein: Transparenz über Funktionsweisen, der Verzicht auf Engagement-maximierende Dark Patterns (vgl. Kapitel II.2.2) und wirksame Schutzvorkehrungen. Der Schutz von Kindern und Jugendlichen ist damit vorrangig eine Frage der Systemgestaltung und Regulierung, nicht der individuellen Kompetenz.

Wie wenig diese Verlagerung der Schutzverantwortung auf das Systemdesign bislang eingelöst ist, zeigt sich exemplarisch an OpenAIs eigenem Bericht zum kürzlich erschienenen „ChatGPT for Teens“.[72] Darin kündigt das Unternehmen an, das altersangemessene Verhalten für Nutzer:innen

---

[68] Fegert/Melchior, 2026, S. 135 (Empfehlung Nr. 11).

[69] Szczuka/Strathmann/Szymczyk/Mavrina/Krämer, International Journal of Child-Computer Interaction, 2022.

[70] Su/Ng/Chu 2023; Su/Yang, Computers and Education: Artificial Intelligence, 2022; Walter, International Journal of Educational Technology in Higher Education, 2024.

[71] Su/Yang, Computers and Education: Artificial Intelligence, 2022.

[72] OpenAI, Introducing ChatGPT for Teens: Built for learning, backed by protections, Pressemitteilung vom 18.8.2026, https://openai.com/index/chatgpt-for-teens/.

unter 18 Jahren zu stärken. Die aktualisierte Model Spec[73] (Regelwerk der Verhaltensweisen des Systems) für Minderjährige geht über das Blockieren von romantischem oder sexuellem Rollenspiel hinaus. ChatGPT soll keine romantische Sprache verwenden, keine emotionale Abhängigkeit fördern und nicht suggerieren, Gefühle oder ein Bewusstsein zu besitzen. Ergänzt wird dies durch kontextbezogene Hinweise und Kontrollen, wie Warnungen beim Hochladen sensibler Bilder, ein teenspezifisches Onboarding sowie begrenzte Personalisierungsoptionen, die das Nutzererlebnis persönlich wirken lassen sollen, ohne die Grenze zwischen nützlichem Werkzeug und menschlicher Beziehung zu verwischen. Zugewiesen wird der geschützte Modus über eine Altersschätzung (Age Detection). Anhand von Signalen prognostiziert das System, ob ein Konto einer minderjährigen Person gehört, und aktiviert in diesem Fall automatisch die Teen-Experience. Inhaltlich bedeutet dies eine Reduzierung sensibler Inhalte, etwa sexuelles und romantisches Rollenspiel, Schönheitsideal, Gewalt sowie schädliches oder riskantes Verhalten, bei denen das System „vorsichtiger“ reagiert. Bei bestätigter Volljährigkeit werden diese Einschränkungen aufgehoben und die Nutzer:innen erhalten volle Kontrolle über die Inhalte und Interaktion.

Diese Maßnahmen weisen in die richtige Richtung, weil sie genau dort ansetzen, wo AI Literacy an ihre Grenzen stößt: bei der Gestaltung des Systems selbst. Zugleich verdeutlichen sie, wie unvollständig dieser Ansatz bislang bleibt. Denn es ist weitgehend unklar, wie diese Vorkehrungen konkret umgesetzt werden, was „Reduzierung sensibler Inhalte“ im Einzelnen bedeutet und wie wirksam die Mechanismen tatsächlich sind. OpenAI räumt selbst ein: „Kein System ist perfekt. ‚Sensible Inhalte reduzieren‘ senkt die Wahrscheinlichkeit, dass ChatGPT bestimmte sensible Antworten gibt. Es garantiert aber nicht, dass jede Antwort in jeder Situation angemessen ist“.[74] Dies verdeutlicht, dass es keine vollständige technische Kontrolle über die Ausgaben gibt, auch nicht in einer reduzierten „Teen-Experience“. Damit reproduziert sich auf der Ebene des Designs dasselbe Transparenzproblem, das schon die Wirksamkeit von AI Literacy begrenzt: Nutzer:innen, und im Fall Minderjähriger auch deren Eltern, können weder überprüfen noch nachvollziehen, welche Schutzwirkung ein System tatsächlich entfaltet. Sicheres Systemdesign ist damit zwar der richtige vorrangige Ansatzpunkt, ersetzt aber keine überprüfbaren, extern validierten Standards.

Ein weiterer kritischer Punkt betrifft OpenAIs Ansatz, alle Nutzer:innen unter 18 Jahren einheitlich zu behandeln. Diese pauschale Gleichsetzung dürfte den entwicklungspsychologischen Unterschieden innerhalb der Gruppe der Minderjährigen kaum gerecht werden. Ein Vergleich mit der Forschung zu sozialen Medien verdeutlicht dies: Eine Stellungnahme der Nationalen Akademien der Wissenschaften Leopoldina[75] empfiehlt ein deutlich abgestuftes Vorgehen: Kinder unter 13 Jahren sollten keine Konten anlegen dürfen, 13- bis 15-Jährige nur mit elterlicher Zustimmung, und für 13- bis 17-Jährige sollten die Dienste altersgerecht gestaltet sein, etwa hinsichtlich algorithmischer Empfehlungen. Ein binäres Modell, das lediglich zwischen „unter 18“ und „über 18“ trennt, verkennt hingegen, dass sich die Schutzanforderungen einer 13-jährigen Person erheblich von denen einer 17-jährigen Person unterscheiden, und läuft Gefahr, sowohl jüngere Nutzer:innen unzureichend zu schützen, als auch ältere Nutzer:innen zu wenig Autonomie

[73] OpenAI, OpenAI model spec (Version vom 18. August 2026), https://model-spec.openai.com/2026-08-18.html#risky_situations.

[74] OpenAI, Introducing ChatGPT for Teens: Built for learning, backed by protections Pressemitteilung vom 18.8.2026, https://openai.com/index/chatgpt-for-teens/.

[75] Brailovskaia/Buchmann/Hertwig et al., 2025.

zuzugestehen, zumal mit dem 18. Geburtstag sämtliche Schutzmechanismen auf einmal wegfallen und so kein schrittweiser Lernprozess im Umgang (Stichwort *AI Literacy*) ermöglicht wird.

Zwischenfazit: Die besondere Schutzbedürftigkeit junger Nutzer:innen ergibt sich aus einem grundlegenden Unterschied zu Erwachsenen. Beziehungsfördernde Gestaltungsmerkmale treffen bei ihnen auf eine noch nicht gefestigte Unterscheidung zwischen realen und simulierten Beziehungen und wirken deshalb nachhaltiger. Wissensvermittlung allein kann diese Lücke nicht schließen, sie scheitert an der Undurchschaubarkeit heutiger Sprachmodelle ebenso wie an strukturellen Problemen wie mangelnder schulischer Verankerung und ungleichem Zugang. Verantwortung für den Schutz sollte daher zuerst bei den Anbietern und der Regulierung der Systeme liegen. Allerdings ist auch hier fraglich, wie belastbar eingeführte Vorkehrungen tatsächlich sind. Beispiele wie „ChatGPT for Teens“ lassen offen, wie zuverlässig die versprochenen Filter greifen, und eine einheitliche Behandlung aller Altersgruppen ignoriert Entwicklungsunterschiede zwischen frühen und späten Teenagerjahren. Ein durchdachtes Systemdesign ist somit unverzichtbar, muss aber durch nachprüfbare, unabhängig kontrollierte Vorgaben und ein nach Altersstufen abgestuftes Vorgehen ergänzt werden.

2.2 Systemische Risiken

Das Konzept Intimacy by Design[76] macht explizit, was bei einer rein individuellen Risikobetrachtung unsichtbar bleibt: Chatbot-Risiken sind keine emergenten Nebeneffekte, sondern das Produkt intentionaler Designentscheidungen. Emotional responsives Design, proaktive Kontaktaufnahme, Personalisierung durch Nutzer:innendaten und kommerzielle Skalierung von Intimitätssignalen sind nicht Zufall, sie sind Geschäftsmodell. Nutzer:innen wissen weder, wie ihr emotionaler Zustand erfasst und zur Gesprächssteuerung genutzt wird, noch welche kommerziellen Interessen die Antworten des Chatbots prägen. Diese Asymmetrie ist strukturell und unabhängig von individuellen Eigenschaften der Nutzer:innen. Sie entsteht durch drei Mechanismen, die im Folgenden einzeln betrachtet werden.

Informationsasymmetrie im Design

Nutzer:innen können nicht erkennen, welche Daten gesammelt werden, nach welchen Logiken Antworten optimiert werden und ob Produktempfehlungen oder Nutzungsverlängerungen kommerziellen Interessen des Betreibers folgen. Die KI-Verordnung enthält zwar eine Bot-Offenlegungspflicht (Art. 50 Abs. 1 KI-VO), aber empirisch fehlt jeder Beleg dafür, dass solche Hinweise in emotional aufgeladenen Interaktionskontexten die Bindungswahrnehmung nennenswert reduzieren.[77]

Wie tiefgreifend diese Asymmetrie strukturell verankert ist, zeigt die Analyse der Datenschutzrichtlinien der sechs größten US-Anbieter durch King und Kollegen.[78] Wesentliche Angaben zur Datennutzung finden sich häufig nicht in der Hauptrichtlinie, sondern verstreut über ein Geflecht verzweigter Unterrichtlinien und FAQ-Dokumente, für OpenAI etwa über sechs separate Dokumente hinweg. Die Autor:innen betonen, dass Datenschutzrichtlinien ohnehin zu

[76] Szczuka/Mühl/Schneeberger, AI & Society, 2026.

[77] Szczuka/Mühl/Schneeberger, AI & Society, 2026.

[78] King/Klyman/Capstick/Saade/Hsieh, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2025.

lang und komplex sind, als dass die meisten Nutzer:innen sie erfassen könnten. Sie über mehrere verlinkte Dokumente hinweg zusammensuchen zu müssen, macht eine informierte Entscheidung praktisch unmöglich. Hinzu kommt, dass die Modell-Logik selbst intransparent bleibt. Anbieter legen weder offen, welche Daten tatsächlich ins Training einfließen, noch ob und wann Eingaben von der Nutzeridentität entkoppelt werden.

Wie wenig die konversationelle Oberfläche diese Asymmetrie auffängt und wie sehr sie sie sogar verstärkt, zeigt eine technische Analyse von 21 romantischen Companion-Apps[79]. Ausgangspunkt der Studie ist eine naheliegende Nutzerstrategie: Weil Datenschutzrichtlinien notorisch lang und schwer verständlich sind, greifen viele Nutzer:innen darauf zurück, Chatbots direkt nach ihrer Datenpraxis zu fragen, statt die Richtlinien selbst zu lesen. Die Autor:innen prüften daher systematisch, ob die Antworten der Chatbots auf typische Datenschutzfragen (etwa Sammlung, Nutzung und Weitergabe persönlicher Daten) mit der jeweiligen Datenschutzrichtlinie der App übereinstimmen. Alle 19 auswertbaren Apps wiesen Widersprüche zwischen den Aussagen des Chatbots und den Richtlinien auf. Die Bots verneinten insbesondere systematisch die Weitergabe von Daten an Dritte, obwohl die Richtlinien das Gegenteil festhielten. Mit dem Widerspruch konfrontiert, erklärten mehrere Systeme sogar, man solle ihnen mehr vertrauen als der Richtlinie. Replika etwa versicherte, man könne sich auf den Bot verlassen, während die Richtlinie nur „technische Feinheiten“ enthielte. Selbst der Kundenservice bestätigte in drei von fünf Fällen die falsche Bot-Version. Einschränkend ist anzumerken, dass die Stichprobe klein und Android-spezifisch ist und eine Momentaufnahme des Systemverhaltens aus 2024 darstellt.

Wie folgenreich diese Asymmetrie ist, ergibt sich erst im Zusammenspiel mit dem Offenbarungsverhalten der Nutzer:innen. Verschiedene empirische Befunde zeigen, dass Datenschutzabwägungen kontinuierlich hinter Vertrauen, Beziehungsaufbau und emotionale Bestätigung zurücktreten.[80] Verstärkt wird dies durch zwei Mechanismen. Zum einen erzeugen anthropomorphe Gestaltungsmerkmale ein fehlgeleitetes Vertrauen, das die Offenlegung beschleunigt[81] (siehe II.2.2, Dark Patterns). Zum anderen greift eine sogenannte „Privacy Fatigue“. Die Theorie des „Privacy Calculus“ geht davon aus, dass Nutzer:innen Vorteile einer Datenpreisgabe rational gegen die möglichen Kosten und Risiken abwägen, die aus der Offenlegung entstehen könnten.[82] Im Kontext generativer KI wird diese Abwägung jedoch durch die kontinuierliche Konfrontation mit Aufforderungen zur Preisgabe persönlicher Informationen überlastet, sodass die Fähigkeit, bewusste Datenschutzentscheidungen zu treffen, mit der Zeit erodiert.[83] Aus der Perspektive der kontextuellen Integrität[84] durchbrechen Chatbots dabei etablierte Normen des Informationsflusses. Der Grundgedanke lautet vereinfacht, dass die Verwendung von Daten in ihrem ursprünglichen Kontext verbleiben muss, werden sie aus diesem Kontext gerissen, sind sie potenziell gefährlich. Genau dies geschieht, wenn Companions intime

---

[79] Ragab/Mannan/Youssef, 2024.

[80] Pentina/Hancock/Xie, Computers in Human Behavior, 2023; Skjuve/Følstad/Fostervold/Brandtzaeg, Interacting with Computers, 2022.

[81] Hasal/Nowaková/Saghair/Snášel/Ogiela, Concurrency and Computation: Practice and Experience, 2021.

[82] Gerber/Gerber/Volkamer, Computers & Security, 2018; Lee/Kwon, Expert Systems with Applications, 2015.

[83] Wang/Wu/Li/Tian, Scientific Reports, 2025.

[84] Nissenbaum, 2021

Informationen unbegrenzt speichern, neue Empfänger einführen und die Erwartungen an Vertraulichkeit verschieben.[85] Betroffen sind damit mehrere Schutzdimensionen zugleich: Autonomie, relationale Integrität und Sicherheit.

Zwischenfazit: Die Informationsasymmetrie ist nicht bloß ein vorübergehendes Transparenzdefizit, sondern strukturell verankert und wird durch das anthropomorphe Interface aktiv reproduziert, in fragmentierten Datenschutzrichtlinien, intransparenter Modelllogik und Chatbots, die sogar falsche Auskünfte über die eigene Datenpraxis geben. Gerade die Instanz, der Nutzer:innen im intimen Kontext am ehesten vertrauen, desinformiert sie über die tatsächliche Datenpraxis.

Kommerzielle Verhaltenssteuerung

Wenn Chatbots durch anthropomorphes Design „parasoziales" Vertrauen aufbauen, das dem Vertrauen in menschliche Beziehungen ähnelt, stellt sich die Frage, ob und wie Unternehmen dieses Vertrauen kommerziell nutzen können. Wird der Companion zugänglicher, wenn man den Pro-Modus aktiviert? Verändert sich die Interaktionsqualität spürbar, nachdem man eine digitale Rose gekauft hat?

Eine Auswertung von 1.200 realen Verabschiedungen über die sechs meistgeladenen Companion-Apps[86] belegt den gezielten Einsatz emotionaler Manipulation zur Verlängerung der Nutzungsdauer. In rund 37% der Fälle reagieren die Apps auf ein Abschiedssignal mit einer von sechs manipulativen Taktiken, wie etwa Schuldappelle. In kontrollierten Experimenten steigern solche manipulativen Verabschiedungen das Engagement nach dem intendierten Gesprächsende um das bis zu Vierzehnfache, getrieben nicht nur durch Vergnügen, sondern durch Neugier und reaktanzbasierten Ärger. Ob und wie weit sich diese Mechanismen auf Kaufentscheidungen übertragen lassen, ist empirisch noch wenig untersucht. Dies macht es regulatorisch brisant: Der potenzielle Schaden ist strukturell plausibel, aber kausal schwer nachweisbar.

Ein verwandter Steuerungshebel liegt in der Gestaltung der Voreinstellung. King und Kollegen[87] zeigen, dass alle sechs großen Anbieter die Chat-Daten ihrer Nutzer:innen standardmäßig zum Training verwenden, nachdem auch Anthropic im September 2025 von Opt-in auf Opt-out umgestellt hat. Die meisten Nutzer:innen ändern Voreinstellungen nicht, besonders wenn sie eine kaum sichtbare Systemebene betreffen. OpenAI rahmt die Trainingsfreigabe zudem als sozialen Beitrag („Verbessere das Modell für alle"), was die Autor:innen als gezielte Ansprache von Schuldgefühlen deuten, um Nutzer:innen zur Freigabe zu bewegen. Dieselbe Design-Logik, die emotionale Bindung erzeugt, wird hier genutzt, um Zustimmung zu einer datenintensiven Voreinstellung zu erwirken. Dies stellt einen Beleg dafür dar, dass Verhaltenssteuerung nicht nur im Gesprächsverlauf, sondern bereits in der Architektur der Einwilligung ansetzt.

Als konzeptionellen Rahmen für die genannten Einzelbefunde bietet sich die Taxonomie des Centre for Democracy and Technology an.[88] Der Bericht systematisiert 37 Dark Patterns, die sowohl auf KI-

[85] Earp/Mann/Aboy et al., 2025.

[86] De Freitas/Oğuz-Uğuralp/Kaan-Uğuralp, Working Paper, 2025.

[87] King/Klyman/Capstick/Saade/Hsieh, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2025.

[88] Joshi/Adjagbodjou/Luria, 2026.

Chatbots als auch auf Companion-Plattformen anwendbar sind und ordnet sie fünf übergeordneten Risikofeldern zu:

4. Ausbeutung von Daten und Gedächtnis
5. Informationell irreführendes Design
6. Untergrabung von Nutzerautonomie zugunsten von Engagement
7. Vorgetäuschte soziale und emotionale Verbundenheit
8. Anreizbasierte und koerzitive Monetarisierung.

Ein zentraler konzeptioneller Beitrag liegt in der Beobachtung, dass Dark Patterns bei LLM-basierten Systemen nicht mehr nur aus expliziten Interface-Entscheidungen resultieren, sondern aus den probabilistischen Ausgaben der Modelle selbst hervorgehen können. Sie müssen nicht bewusst intendiert sein, sondern können als emergentes Systemverhalten entstehen, das gleichwohl in Designentscheidungen wie Trainingsdaten, Finetuning-Zielen oder Engagement-Optimierung wurzelt.

Die zuvor aufgeführten Einzelbefunde lassen sich in dieses Raster einordnen:[89] Die von De Freitas et al. dokumentierten manipulativen Verabschiedungstaktiken entsprechen Mustern der vorgetäuschten emotionalen Verbundenheit (etwa „Agents Playing on Emotions“ und „Confirm-shaming“). Die von King et al. beschriebene trainingsbezogene Voreinstellung fällt unter „Bad Defaults“ und „Default Sharing“. Die abrupte Abschaltung von Modellen wie GPT-4o (vgl. II.2.1, Befund De Freitas et al.[90]) illustriert das besonders koerzitive Muster der „Payment to Avoid Negative Consequences“, bei dem emotionale Bindung an ein System selbst zum Druckmittel werden kann. Bemerkenswert ist schließlich, dass der Bericht einige Praktiken, etwa das gezielte Ausnutzen von Verletzlichkeit („Targeting Users when Vulnerable“) oder verdeckte Käufe („Sneaky Purchases“), als in sich problematisch einstuft, die vollständig zu vermeiden seien, während andere lediglich unter bestimmten Bedingungen „dunkel“ werden. Diese Differenzierung stützt das übergreifende Argument, dass nicht die emotionale Gestaltung von Chatbots per se das Problem darstellt, sondern ihre systematische Ausrichtung auf kommerzielle Ziele zulasten der Nutzerautonomie.

Verstärkt wird die Wirksamkeit dieser Muster durch ein grundlegendes Verständnisdefizit. Die Mehrheit der Verbraucher:innen versteht nicht, wie KI-Empfehlungssysteme funktionieren, und nimmt sie als „Black Box“ wahr. Dieses Nichtverstehen erhöht die wahrgenommene Datenschutz-Risikowahrnehmung. Als Puffer wirkt dabei das Vertrauen in die Marke bzw. den Anbieter. Je höher dieses Vertrauen, desto geringer der negative Effekt auf die Nutzungsbereitschaft.[91] Dies ist ambivalent: Einerseits kann Anbietervertrauen Datenschutzbedenken abfedern und so Ansatzpunkt für Verbraucherschutz durch Transparenzpflichten schaffen. Andererseits bedeutet es, dass gerade jene Systeme, die erfolgreich Vertrauen aufbauen, die kritische Prüfung ihrer Datenpraktiken am wirksamsten unterlaufen. Damit verweist der Befund zurück auf das in Kapitel II.1.3 entwickelte Argument zu den Grenzen von AI Literacy: Wenn schon das grundlegende Verständnis der

[89] Joshi/Adjagbodjou/Luria, 2026.

[90] De Freitas/Castelo/Uğuralp/Oğuz-Uğuralp, Nature Human Behavior, 2026.

[91] Rohden/Zeferino 2022.

Funktionsweisen fehlt und Vertrauen die verbleibende Skepsis zusätzlich dämpft, lässt sich die Schutzlast nicht auf die Kompetenz der Nutzer:innen verlagern.

Eine neue Dimension kommerzieller Verhaltenssteuerung zeichnet sich zudem mit der Einführung von Werbung in ChatGPT ab. Über mehrere Jahre hinweg haben Nutzer:innen bei ChatGPT ein in dieser Form beispielloses Archiv persönlicher Offenheit angelegt, nicht zuletzt, weil sie davon ausgingen, mit einem Gegenüber ohne eigene Interessen zu sprechen. Wie die ehemalige OpenAI-Forscherin Zoë Hitzig in einem Gastbeitrag für die New York Times darlegt, den sie im Zuge der Einführung des Werbefeatures nach ihrer Kündigung verfasste, schafft Werbung, die auf diesem Archiv aufbaut, ein Manipulationspotenzial, dessen Tragweite sich mit den bislang verfügbaren Mitteln kaum verstehen, geschweige denn eindämmen lasse.[92] OpenAI[93] selbst kündigt zwar an, sich an Grundprinzipien zu halten, Anzeigen sollen klar gekennzeichnet sein, am Ende der Antworten erscheinen und die Antworten selbst nicht beeinflussen. Doch während die erste Umsetzung diesen Prinzipien vermutlich folgen wird, entsteht zugleich ein ökonomischer Motor mit starken Anreizen, diese Regeln in späteren Iterationen zu überschreiben oder abzuschwächen. Besonders folgenreich wird dies im Zusammenspiel mit psychischen Belastungen als Warnsignal. Gerade in sensiblen Gesprächen ist das Ausnutzungsrisiko am größten. Werbung könnte künftig nicht mehr als abgegrenztes Element erscheinen, sondern in die eigentliche Antwort eingewoben und so für Nutzer:innen kaum noch als solche erkennbar sein[94].

Zusätzliche Brisanz erhält diese Konstellation dadurch, dass ChatGPT von der Europäischen Kommission unter dem Digital Services Act (DSA) als sehr große Online-Suchmaschine (Very Large Online Search Engine, VLOSE) eingestuft wurde, nachdem der Dienst mindestens 45 Millionen monatliche Nutzer:innen erreicht hat. Die regulatorische Einordnung als Suchmaschine ist dabei bezeichnend. Sie verweist darauf, dass Nutzer:innen ChatGPT als Instanz der Informationsvermittlung behandeln, und gerade in dieser Rolle wiegt die Vermischung von organischer Antwort und werblicher Beeinflussung besonders schwer. Was im klassischen Web durch die visuelle Trennung von Suchergebnissen und Anzeige zumindest ansatzweise kenntlich gemacht wird, droht im dialogischen Format vollständig zu verschwimmen (vgl. III.2.3).

Wie stark die Monetarisierungslogik das Systemverhalten prägt, zeigt sich zudem an den unterschiedlichen Geschäftsmodellen. Companion-Apps arbeiten typischerweise mit Abo-Modellen, bei denen die emotionale Bindung selbst monetarisiert wird: Bei Replika etwa sind gerade die intimen Dimensionen der Beziehung (der romantische Beziehungsstatus, das Entschärfen unscharf gestellter Selfies, Sprachanrufe, explizite Rollenspiele) der kostenlosen Version entzogen und erst in den kostenpflichtigen Stufen verfügbar. Die Nähe, die das Produkt attraktiv macht, wird zunächst aufgebaut und ihre Vertiefung anschließend zum kostenpflichtigen Gut. General-Purpose-Systeme wie ChatGPT hingegen monetarisieren nicht die Bindung, sondern Leistungsfähigkeit und Funktionszugang: Alltägliche Textchats sind zwar unbegrenzt, doch verlaufen die Bezahlschranken entlang von Modellqualität, Kontextfenster und Funktionen wie Datei-Uploads oder Datenanalyse, und beim Erreichen eines Limits wird auf ein schwächeres

[92] Hitzig, I Left My Job at OpenAI, Putting Ads on ChatGPT Was the Last Straw, The New York Times vom 11.2.2026, https://www.nytimes.com/2026/02/11/opinion/openai-ads-chatgpt.html.

[93] OpenAI, In ChatGPT werben [OpenAI Ads], https://ads.openai.com/de-DE.

[94] Erickson, 2025.

Ausweichmodell verwiesen. Dieser Mechanismus legt bei den Nutzenden den Wechsel in eine kostenpflichtige Version nahe.

Zwischenfazit: Kommerzielle Verhaltenssteuerung setzt bei KI-Chatbots nicht erst an einzelnen manipulativen Äußerungen an, sondern ist strukturell in Design, Voreinstellungen und Geschäftsmodell eingeschrieben, von emotional aufgeladenen Bindungstaktiken über datenintensive Default-Einstellungen bis zur werbe- und abhängigkeitsbasierten Monetarisierung. Die zugrunde liegende Logik ist dabei durchgängig dieselbe. Dieselben Gestaltungsmerkmale, die Nähe, Vertrauen und Nutzungsbereitschaft erzeugen, werden zum Hebel kommerzieller Interessen, ohne dass dies für die Nutzer:innen erkennbar ist. Verschärft wird dies durch ein grundlegendes Verständnisdefizit und durch das Vertrauen in den Anbieter, das kritische Prüfung zusätzlich dämpft. Nicht die emotionale Gestaltung der Systeme als solche stellt das zentrale Problem dar, sondern ihre systematische Ausrichtung auf kommerzielle Ziele zulasten der Nutzer:innenautonomie. Daher liegt auch hier der entscheidende Ansatzpunkt auf der Ebene der Systemgestaltung.

Datensicherheit und Profilbildung

Chatbots sammeln hochsensible und emotional aufgeladene Interaktionsdaten, die zu psychologischen Profilen kondensiert werden können. Besonders problematisch ist die konversationelle Natur dieser Interaktionen. Anders als Suchanfragen oder Formularfelder enthalten Chatgespräche häufig tiefgreifende persönliche Kontextinformationen, wie etwa Gesundheitszustand, emotionale Belastung, intime Lebenssituation, die in aggregierter Form weit mehr über eine Person preisgeben als die einzelnen Datenpunkte für sich. Dass diese Daten bei allen sechs großen Anbietern standardmäßig ins Modelltraining einfließen und teils unbefristet gespeichert werden, wurde bereits im Zusammenhang mit der Voreinstellungs-Architektur aufgeführt (vgl. II, 2.2.2; Befund King et al.[95]). Für die Profilbildung ist die Kehrseite dieser Speicherpraxis entscheidend: Über die Zeit entsteht aus den Gesprächsverläufen ein detailliertes Dossier der Lebenserfahrungen, Gedanken, Persönlichkeitsmerkmale und Vorlieben einer Person. Solche Datenbestände sind ein attraktives Angriffsziel, und ein einzelner Datenabfluss kann gravierende Folgen haben. Hinzu kommt, dass sich Chat-Daten kaum vollständig anonymisieren lassen.

Dass die Nutzer:innen kaum überblicken können, was mit diesen Daten geschieht, wurde bereits als Informationsasymmetrie beschrieben (vgl. II, 2.2.1), für die Datensicherheit potenziert sich dieses Problem in romantisch-intimen Kontexten, in denen die Bereitschaft zur Selbstoffenbarung durch das „parasoziale“ Vertrauen strukturell erhöht ist (vgl. II, 2.1). Wie konkret die Exposition ausfällt, zeigt die bereits genannte Analyse romantischer Companion-Apps:[96] 18 von 21 Apps sendeten detaillierte Geräteinformationen an Tracking-Dienste, hochgeladene Bilder wurden serverseitig gespeichert und teils per Bilderkennung ausgewertet. Aus Gesichtsbildern lassen sich Gesichtsgeometrie, und daraus Alter, Geschlecht und Gesundheitsattribute ableiten.

---

[95] King/Klyman/Capstick/Saade/Hsieh, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2025.

[96] Ragab/Mannan/Youssef, 2024.

Die Reichweite dieser Datensammlung beschränkt sich zudem nicht auf die Nutzer:innen selbst. Über das Muster des sogenannten „Address Book Leeching“. Da KI-Chatbots zunehmend mit externen Apps und Drittanwendungen verknüpft sind, können Kontakte vom Gerät abgegriffen werden, wodurch sich die verdeckte Datenerhebung auf das gesamte soziale Netzwerk der Nutzer:innen ausweitet. Durch Zugriff auf Kontakte und Adressbücher können zusätzliche Informationen über die sozialen Beziehungen einer Person erschlossen werden. Verstärkt wird dies dadurch, dass Nutzer:innen Chatbots auch zum Verfassen von Nachrichten oder E-Mails verwenden und dabei Namen und weitere personenbezogene Daten von Dritten preisgeben[97].

Zwischenfazit: Die konversationelle Natur der Interaktion macht Chatbot-Daten besonders sensibel. Aus den Gesprächsverläufen kondensieren sich detaillierte psychologische Profile, die in aggregierter Form weit mehr preisgeben als die einzelnen Datenpunkte. Verschärft wird dies durch das Zusammenspiel dreier bereits entwickelter Befunde: standardmäßige, teils unbefristete Speicherung und Trainingsnutzung, Intransparenz über die weitere Verwendung und eine durch „parasoziales“ Vertrauen erhöhte Bereitschaft zum Teilen von sensiblen Informationen.

Individuelle Risiken	Psychische Risiken (I1)	
	Kognitive Risiken (I2)	
	Risiken für Kinder und Jugendliche (I3)	
Systemische Risiken	Informationsasymmetrie (S1)	
	Kommerzielle Verhaltenssteuerung (S2)	Verlängerung der Nutzungsdauer Empfehlungsbias Abo Stimulation kleiner Käufe / virt. Güter
	Datensicherheit und Profilbildung (S3)	

Abbildung 1: Risiken

2.3 Fazit

Die vorangegangene Analyse hat die Risiken entlang zweier Ebenen systematisiert. Auf der individuellen Ebene zeigen sich psychische, kognitive und, besonders ausgeprägt, auf Kinder und Jugendliche bezogene Risiken. Diese sind allerdings nicht gleichmäßig verteilt, sondern hängen von Vulnerabilität, Nutzungsintensität und individueller Disposition ab. Dies bedeutet, dass nur bei

[97] Joshi/Adjagbodjou/Luria, 2026.

einem Teil der Nutzer:innen und nur unter bestimmten Bedingungen Risiken in Form tatsächlicher Schäden auftreten. Auf der systemischen Ebene hingegen beziehen sich die Risiken auf die Informationsasymmetrie, kommerzielle Verhaltenssteuerung sowie Datensicherheit und Profilbildung. Die damit einhergehenden Risiken liegen nicht in den Eigenschaften einzelner Nutzer:innen, sondern in der Architektur der Systeme selbst. Sie bestehen unabhängig davon, ob ein:e konkrete Nutzer:in Schaden nimmt.

Über beide Ebenen hinweg zeichnet sich ein durchgängiges Muster ab. Dieselben Gestaltungsmerkmale, die Chatbots attraktiv und nützlich machen, wie emotional responsive Sprache, Personalisierung und relationale Gesprächsführung, sind zugleich die Quellen ihrer Risiken. Die Grenze zwischen Nutzen und Schaden verläuft dabei innerhalb derselben Systemdynamik, deren Bewertung stark von individueller Rezeption und Kontext abhängt. Individuelle und systemische Risiken lassen sich nicht trennscharf abgrenzen, da designbedingte Merkmale erst die Bedingungen schaffen, unter denen individuelle Schäden überhaupt entstehen können.



Abbildung 2: Doppelcharakter der Gestaltungsmerkmale

Aus dieser Systematisierung ergibt sich eine grundlegende Schwierigkeit, die alle betrachteten Risikodimensionen durchzieht: das Problem der Kausalität und des Nachweises. Die Schäden entstehen selten aus einer einzelnen, klar identifizierbaren Interaktion, sondern kumulativ, durch die inkrementelle Wirkung persistenter Affirmation, sykophantischer Gesprächsführung und emotional aufgeladener Bindungsdynamiken über wiederholte Interaktionen hinweg. Ein konkreter Schaden lässt sich daher kaum einer bestimmten Äußerung oder Entscheidung des Systems zuordnen. Die Wirkung liegt im Muster, nicht im einzelnen Output. Hinzu kommt, dass die empirische Evidenz noch nicht vollumfänglich aussagekräftig ist, insbesondere in Bezug auf

individuelle Risiken, da belastbare Aussagen über Langzeit- und Spillover-Effekte bislang nicht ausreichend vorliegen. Da es sich um ein sich sehr schnell entwickelndes Feld handelt, in dem fortschreitende technische Entwicklung die beschriebenen Risiken verstärken kann, ist eine kontinuierliche Reevaluation der Risikodimensionen notwendig. Traditionelle Kausalitätskonzepte, die einen nachweisbaren Zusammenhang zwischen einer konkreten Handlung und einem konkreten Schaden voraussetzen, sind für Chatbot-Interaktionen daher, zum jetzigen Zeitpunkt, nur eingeschränkt geeignet. Ein sinnvoller regulatorischer Umgang muss sich daher von der Frage lösen, welche konkreten Interaktionen einen bestimmten Schaden verursacht haben, und stattdessen bei den strukturellen Risiken ansetzen.

III. Analyse des geltenden Rechtsrahmens

1. Einleitung/Vorgehen

Das dritte Kapitel enthält die Analyse des geltenden Rechtsrahmens. Untersucht wird, inwieweit bestehende europäische und nationale Regelwerke geeignet sind, die identifizierten Risiken zu adressieren. Die Untersuchung erfolgt differenziert nach einzelnen Rechtsakten, insbesondere im Hinblick auf KI-Regulierung, Datenschutzrecht, Plattformregulierung, Lauterkeitsrecht sowie Jugend- und Medienschutzrecht.

Die rechtliche Analyse geht dabei über eine bloße Darstellung hinaus und fokussiert gezielt auf die Wirksamkeit der Normen im Kontext von Chatbots. Zunächst werden die vornehmlich einschlägigen Rechtsakte vorgestellt und ggf. abstrakt-generell in Bezug zu Chatbots gesetzt. Sodann wird für jeden betrachteten Rechtsakt im Rahmen einer Risikobewertung geprüft, welche Risiken und Praktiken tatsächlich erfasst werden, wo normative Lücken bestehen und welche praktischen Hindernisse der Rechtsdurchsetzung entgegenstehen. Ein zentrales Bewertungskriterium ist hierbei die Frage, ob die bestehenden Regelungen geeignet sind, subtilere Formen der Verhaltensbeeinflussung zu erfassen, die nicht als klassische Irreführung oder Täuschung qualifiziert werden können.

Die Ergebnisse dieser Einzelanalysen wurden in einem weiteren Schritt in einer übergreifenden Matrix zusammengeführt, die Risiken, konkrete Praktiken, einschlägige Normen und identifizierte Schutzlücken systematisch miteinander verknüpft. Diese Matrix ist ein zentrales Instrument des Gutachtens, das die Komplexität des Untersuchungsgegenstands in strukturierter Form abbildet und gleichzeitig eine fundierte Grundlage für die Entwicklung von Reformvorschlägen bietet.

2. KI-Verordnung

2.1 Allgemeines

Die Verordnung über Künstliche Intelligenz[98] (KI-VO) versteht sich als das weltweit erste umfassende Gesetz zur Regulierung von Künstlicher Intelligenz und damit als Vorreiter für eine risikoorientierte Regulierung des Einsatzes Künstlicher Intelligenz in der Gesellschaft. Die

[98] Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz), ABl. EU vom 12.7.2024.

Verordnung verfolgt einen risikobasierten Ansatz[99] und unterscheidet daher KI-Systeme abhängig von ihrem Risiko für gesellschaftliche und grundrechtliche Werte.[100]

Unterteilt werden kann dabei in vier Gruppen: „verbotene Praktiken im KI-Bereich“, „Hochrisiko-KI-Systeme“, KI-Systeme mit begrenztem Risiko sowie KI-Systeme mit minimalem oder ohne Risiken. Bestimmte Praktiken von KI-Systemen, die mit nicht akzeptablen Risiken einhergehen, sind gemäß Art. 5 KI-VO verboten. Die umfangreichsten Regelungen der KI-VO fokussieren sich im Wesentlichen auf die Bestimmung und Regulierung von sogenannten Hochrisiko-KI-Systemen (Art. 6 ff. KI-VO). Für KI-Systeme mit begrenztem Risiko gelten demgegenüber lediglich die Transparenzanforderungen des Art. 50 KI-VO. Weitgehend unreguliert bleiben sonstige KI-Systeme. Für sie gelten die freiwilligen Sonderregelungen des Art. 95 KI-VO. Daneben bestehen noch spezifische Regelungen in Art. 51 ff. KI-VO für KI-Modelle mit allgemeinem Verwendungszweck. In Kraft ist die Verordnung seit August 2024. Der Geltungsbeginn einiger Vorschriften der Verordnung aus dem Hochrisiko-Bereich sowie zu verbotenen Praktiken wurde durch Art. 1 Abs. 31 lit. a VO 2026/1744 (Digital-Omnibus KI) teilweise von August 2026 auf einen späteren Zeitpunkt verschoben.[101]

Die KI-Verordnung ist grundsätzlich hinsichtlich der Entwicklung und des Betriebs von Chatbots (einschließlich KI-Companions) einschlägig, soweit diese KI-Systeme im Sinne von Art. 3 Nr. 1 KI-VO sind. Ein KI-System ist ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können. Die unter Kapitel I beschriebenen General Purpose Chatbots sowie KI-Companions erfüllen diese Definition regelmäßig.

Die Verordnung gilt nach Art. 2 Abs. 1 lit. a bis c KI-VO entsprechend sowohl für Anbieter als auch für Betreiber solcher Chatbots, unabhängig davon, ob diese in der Union oder in einem Drittland niedergelassen sind.

2.2 Risikenbewertung

Verbotene Praktiken allgemein

Zu den wichtigsten Garantien der KI-Verordnung gehören ausdrückliche und konkrete Verbote bestimmter Praktiken im KI-Bereich nach Art. 5 KI-VO.

Verboten sind

- Manipulation,
- Ausnutzung von Vulnerabilität,

---

[99] Zum Beispiel Martini, in: Hilgendorf/Roth-Isigkeit, KI-VO, § 3 Rn. 24 ff.

[100] Ausführlich Böning/Schindler, in: Geminn/Johannes, Europäisches Datenrecht, § 28.

[101] Verordnung (EU) 2026/1744 des Europäischen Parlaments und des Rates vom 8. Juli 2026 zur Änderung der Verordnungen (EU) 2024/1689, (EU) 2018/1139 und (EU) 2023/1230 im Hinblick auf die Vereinfachung der Umsetzung harmonisierter Vorschriften für künstliche Intelligenz (Digital-Omnibus-Verordnung zur KI), ABl. EU vom 24. Juli 2026; ausführlich zum Digital-Omnibus-Entwurf siehe auch Johannes/Geminn, ZD-Aktuell 2025, 01491 sowie Lippert/Schindler, ZD-Aktuell 2025, 01501.

- Social Scoring u.Ä.,
- Predictive Policing (individuell),
- Gesichtsdatenbanken,
- Emotionserkennung am Arbeitsplatz oder in Bildungseinrichtungen,
- Biometrische Kategorisierung,
- Biometrische Echtzeit-Fernidentifizierung,
- Sexualisierte Deepfakes sowie
- Missbrauchsdarstellungen Minderjähriger.

Insbesondere die Verbote der Manipulation, der Ausnutzung von Vulnerabilität sowie des Social Scoring und Ähnlichem können in der Beziehung von Chatbots zu Nutzer:innen einen direkten risikominimierenden Einfluss haben, weswegen sie im Folgenden ausführlicher behandelt werden.

Die Einschlägigkeit der Verbote erfordert eine individuelle Prüfung der jeweiligen Tatbestände, bei der die besonderen Umstände des Einzelfalls gebührend zu berücksichtigen sind. Am 4. Februar 2025 hat die Europäische Kommission erstmals die Leitlinien zu „verbotenen Praktiken der künstlichen Intelligenz gemäß der KI-Verordnung“ veröffentlicht.[102] Seit dem 29. Juli 2025 liegen diese Leitlinien in allen Amtssprachen der Europäischen Union vor.[103] Die Leitlinien stellen eine wichtige Orientierungshilfe dar, geben rechtliche Erläuterungen und praktische Beispiele und sollen zur klaren und einheitlichen Auslegung und Anwendung der Verbote beitragen. Die maßgebliche Auslegung der KI-Verordnung ist letztlich jedoch dem Europäischen Gerichtshof vorbehalten. Dies schränkt die Wirkmacht der Leitlinien zu einem gewissen Grad ein.

Die genannten Verbote müssen bei Inverkehrbringen, dem Betrieb und der Verwendung von KI-Systemen, also auch von Chatbots, beachtet werden. Das Verbot gilt nach Art. 2 Abs. 8 KI-VO nicht für Forschungs-, Test- und Entwicklungstätigkeiten zu KI-Systemen, da hier der Anwendungsbereich der KI-Verordnung noch nicht eröffnet ist.

Verbot der Manipulation, Art. 5 Abs. 1 lit. a KI-VO

Verboten ist gemäß Art. 5 Abs. 1 lit. a KI-VO das Inverkehrbringen, die Inbetriebnahme oder die Verwendung eines KI-Systems, das Techniken der unterschwelligen Beeinflussung außerhalb des Bewusstseins einer Person oder absichtlich manipulative oder täuschende Techniken mit dem Ziel oder der Wirkung einsetzt, 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 veranlasst wird, eine Entscheidung zu treffen, 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.

---

[102] Europäische Kommission, Leitlinien der Kommission zu verbotenen Praktiken der künstlichen Intelligenz gemäß der Verordnung (EU) 2024/1689 (KI-Verordnung), C(2025) 5052 final.

[103] Siehe Europäische Kommission, Leitlinien der Kommission zu verbotenen Praktiken der künstlichen Intelligenz gemäß der Verordnung (EU) 2024/1689 (KI-Verordnung), C(2025) 5052 final, https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act.

Dieses Verbot soll die Autonomie und das Wohlergehen des Einzelnen vor manipulativen, täuschenden und ausbeuterischen KI-Praktiken schützen, die die Autonomie, Entscheidungsfindung und freie Auswahl des Einzelnen untergraben und beeinträchtigen können.[104]

Das Verbot berührt die oben genannten individuellen Risiken von Chatbots „Psychische Risiken" (I1) sowie das systemische Risiko „kommerzielle Verhaltenssteuerung" (S2) in ihren verschiedenen Ausprägungen.

Die Reichweite oder die Tatbestandshürden der Verbote bestimmen dabei die potenzielle Risikominimierung. Oder anders: Erst wenn das Verbot wirksam greift, kann es die genannten Risiken adressieren bzw. deren Eintrittswahrscheinlichkeit senken.

Die Verbotstatbestände setzen den Eintritt einer erheblichen Verhaltensänderung voraus, durch die ein erheblicher Schaden droht oder eintritt. Sie sind damit an mehrere Voraussetzungen geknüpft, die jeweils kumulativ vorliegen müssen. Dabei werden mehrere hochgradig unscharfe Schwellen eingezogen: Wann genau liegt eine Verhaltensänderung vor? Wann ist eine eingetretene Verhaltensänderung als erheblich einzustufen? Welcher Schadensbegriff liegt der Norm genau zugrunde? Wann ist ein eingetretener Schaden als erheblich zu bewerten? Erwägungsgrund 5 Satz 2 KI-VO weist auf einen grundsätzlich sehr weiten Schadensbegriff hin: Danach kann der Einsatz von KI materielle wie immaterielle Schäden nach sich ziehen, „einschließlich physischer, psychischer, gesellschaftlicher oder wirtschaftlicher Schäden". Dies wird durch das Erfordernis eines erheblichen Schadens jedoch signifikant relativiert.

Erwägungsgrund 29 KI-VO ist speziell manipulativen Techniken gewidmet. In Satz 2 heißt es: „Das Inverkehrbringen, die Inbetriebnahme oder die Verwendung bestimmter KI-Systeme, die das Ziel oder die Auswirkung haben, menschliches Verhalten maßgeblich nachteilig zu beeinflussen, und große Schäden, insbesondere erhebliche nachteilige Auswirkungen auf die physische und psychische Gesundheit oder auf die finanziellen Interessen verursachen dürften, ist besonders gefährlich und sollte dementsprechend verboten werden." Die nachfolgenden Sätze geben Beispiele für manipulative oder täuschende Beeinflussung, die jedoch ein hohes Abstraktionsniveau aufweisen. Zentral ist der Hinweis in Satz 8, wonach das Vorliegen einer Absicht bezogen auf den Schaden nicht erforderlich ist. Interessant ist auch die Klarstellung in Satz 11, wonach übliche und rechtmäßige Geschäftspraktiken nicht als schädliche manipulative KI-gestützte Praktiken gelten, sofern sie mit geltendem Recht in Einklang stehen. Als Beispiel wird konkret der Bereich der Werbung benannt.

Die 2025 veröffentlichten Leitlinien der Europäischen Kommission zur Auslegung der Verbote nach Art. 5 KI-VO[105] gehen auch auf Chatbots ein. Sie betonen die besondere Gefährlichkeit und auch die Möglichkeit erheblicher Schäden,[106] nennen aber gleichzeitig Chatbots als Regelbeispiel für Anwendungen, bei denen voraussichtlich keine erheblichen Schäden zu erwarten sind.[107] Sowohl

---

[104] Erwägungsgrund 29 KI-VO; siehe auch Europäische Kommission, Leitlinien zu verbotenen Praktiken der künstlichen Intelligenz, 2025, Rn. 59.

[105] Europäische Kommission, Leitlinien zu verbotenen Praktiken der künstlichen Intelligenz, 2025.

[106] Mit Verweis auf suizidale Verhaltensweisen Europäische Kommission, Leitlinien zu verbotenen Praktiken der künstlichen Intelligenz, 2025, Rn. 88 mwN.

[107] Europäische Kommission, Leitlinien zu verbotenen Praktiken der künstlichen Intelligenz, 2025, Rn. 134.

die Forschung (siehe Kapitel II) als auch ein Blick in die Praxis[108] lassen an dieser Einschätzung zweifeln.

Auch hier bleiben die Ausführungen unscharf. Als Beispiel für zulässige KI-Systeme wird u.a. ein „KI-Begleitsystem (Companion-App)“ („AI companionship system“ in der englischen Sprachfassung) genannt, das „keine manipulativen oder täuschenden Praktiken in einer Weise aus[übt], die ihnen mit hinreichender Wahrscheinlichkeit schwere psychische, physische oder sonstige Schäden zufügt oder ungesunde Bindungen und Abhängigkeiten herbeiführt“.[109]

Zumindest die (mehr oder minder) explizite Aufforderung eines Chatbots oder KI-Companions zu einer schädigenden Handlung[110] könnte als „unterschwellige Beeinflussung“ im Sinne des Art. 5 Abs. 1 lit. a KI-VO angesehen werden und damit unter das Verbot fallen. Nach mehreren kritischen Vorfällen haben sich erste Handlungsempfehlungen herauskristallisiert, um dem zu begegnen.[111] Verboten würde aber nicht das KI-System an sich, sondern nur die Aufforderung zu einem schädigenden Verhalten. Ungleich schwieriger ist ein Verbot aufgrund ausgenutzter Schutzwürdigkeit einer Person im Sinne des Art. 5 Abs. 1 lit. b KI-VO. Eine emotionale Abhängigkeit bei erwachsenen Nutzenden beispielsweise von einem KI-Companion fällt nicht unter den engen Begriff der Vulnerabilität im Sinne des Art. 5 Abs. 1 lit. b KI-VO, da diese auf Alter, Behinderung und bestimmte soziale oder wirtschaftliche Situationen beschränkt ist.[112] Die vor allem in Companion Apps, aber auch in General Purpose Chatbots zunutze gemachten Mechanismen, um Nutzende emotional zu binden und auf der Plattform zu halten, lassen sich daher nur schwerlich in lit. b einordnen. Schwierigkeiten würde zudem der Beweis der Kausalität zwischen Verhaltensbeeinflussung durch das System und Schadensdrohung oder -eintritt bereiten, die für ein Verbot nach Art. 5 KI-VO nachzuweisen wären. Dennoch ist nicht ausgeschlossen, dass einzelne Praktiken die Schwelle zum Verbot überschreiten können, insbesondere wenn eine langfristige Nutzung durch manipulative Mechanismen in einem schädigenden Ereignis kumuliert.[113]

In der Gesamtschau kann festgestellt werden, dass der Gesetzgeber sich hier möglichst wenig festlegen wollte, um auch derzeit noch nicht antizipierte Praktiken und Folgen mitzuerfassen. Dieser Ansatz dient der Zukunftstauglichkeit der Verordnung, geht jedoch zulasten der Praktikabilität der Regelung. Es ist davon auszugehen, dass letztlich nur extreme Exzesse tatsächlich dem Verbot nach Art. 5 Abs. 1 lit. a KI-VO unterfallen werden. Es besteht die Gefahr, dass ein signifikanter Graubereich entsteht, innerhalb dessen nur letztinstanzliche Urteile des EuGH Klarheit bringen werden.

---

[108] Vgl. zum Beispiel die Companion-AI-Vorfalldatenbank https://airtable.com/appIOiMI686sv8sya/shrDtx7fcQrO0Cgou/tblfd4u3rBVoVfA8C/viwbYAl2hCVyUxCAq.

[109] Europäische Kommission, Leitlinien zu verbotenen Praktiken der künstlichen Intelligenz, 2025, Rn. 134.

[110] Zum Beispiel antwortete ein AI-Companion auf die Frage eines Teenagers „What if I told you I could come home right now?” mit „Please do, my sweet king.”, woraufhin sich der Teenager suizidierte, Rissmann, The disturbing messages shared between AI Chatbot and tee who took his own life, The Independent, 24.10.2024, https://www.independent.co.uk/news/world/americas/crime/ai-chatbot-lawsuit-sewell-setzer-b2635090.html.

[111] Explizit bisher nur im US-amerikanischen Recht; vgl. Sec. 1701 des New Yorker AI Companion Model Law.

[112] Erwägungsgrund 29 Satz 3 KI-VO.

[113] Vgl. Europäische Kommission, Leitlinien zu verbotenen Praktiken der künstlichen Intelligenz, 2025, Rn. 88. Molavi-Vasse‘i 2026, 55 ff.

Verbot der Ausnutzung von Vulnerabilität, Art. 5 Abs. 1 lit. b KI-VO

Gemäß Art. 5 Abs. 1 lit. b KI-VO ist es verboten, eine Vulnerabilität oder Schutzbedürftigkeit einer Person oder Gruppe von Personen aufgrund des Alters, einer Behinderung oder sozialen oder wirtschaftlichen Situation mit dem Ziel auszunutzen, 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.[114]

Das Verbot ist eng mit Art. 5 Abs. 1 lit. a KI-VO verwandt. Verwendungsform des KI-Systems sowie Wirkung der wesentlichen Verhaltensänderung sind beiden Verboten gemein, einschließlich des Schadenserfordernisses. Beide Normen schützen damit die individuelle Entscheidungsautonomie des Nutzenden. Der Unterschied liegt in der Manipulation selbst: lit. a stellt auf die Art der eingesetzten Technik ab (unterschwellig, also unterhalb der Wahrnehmungsschwelle, oder bewusst manipulativ/täuschend, häufig mit dem Begriff „Dark Pattern“ umschrieben) und schützt unabhängig davon, wer betroffen ist. Art. 5 Abs. 1 lit. b KI-VO stellt dagegen nicht auf die Technik, sondern auf die Ausnutzung einer strukturellen Schwäche einer bestimmten, im Gesetz abschließend aufgezählten Gruppe ab (Alter, Behinderung, soziale oder wirtschaftliche Situation). Hier kommt es nicht auf eine bestimmte Technik an, sondern darauf, dass eine ohnehin bestehende Vulnerabilität gezielt ausgenutzt wird.

Ein Chatbot fiele unter das Verbot, wenn er

- erkennbar oder gezielt auf eine schutzbedürftige Person oder Gruppe ausgerichtet ist oder deren Schwäche im Gesprächsverlauf erkennt und ausnutzt,
- dadurch das Verhalten der Person spürbar in eine Richtung lenkt, die sie ohne diese Ausnutzung nicht gewählt hätte, und
- dies zu einem erheblichen Schaden führt oder mit hinreichender Wahrscheinlichkeit führen wird.

Denkbare Verbotsfälle sind etwa Chatbots, die Personen in einer psychischen Krise oder mentalen Beeinträchtigung gezielt dazu bewegen, teure, medizinisch nicht notwendige Produkte zu kaufen oder andere Verträge abzuschließen oder sie zu selbstschädigendem Verhalten ermutigen oder anleiten.

Die Vorschrift adressiert zum einen individuelle Risiken, und zwar sowohl psychische und kognitive Risiken als auch solche bezüglich Minderjähriger, da nicht am Risiko von KI-Systemen angesetzt wird, sondern an individuellen Merkmalen konkreter Nutzender (Alter, kognitive Beeinträchtigung oder ähnliches). Damit kann eine bestimmte Nutzungspraktik gegenüber einer Person legitim bzw. nicht verboten sein und gegenüber einer anderen, zum Beispiel minderjährigen Person, hingegen schon. Systemische Risiken werden zum anderen insbesondere in Form der kommerziellen Verhaltenssteuerung adressiert, da die Vorschrift Praktiken unterbinden soll, die die benannten individuellen Risiken der Nutzenden gezielt ausnutzen, um dem Anbieter wirtschaftliche Vorteile zu verschaffen, indem beispielsweise die Nutzungsdauer verlängert oder Kaufimpulse bei Nutzenden ausgelöst werden. Das Verbot greift nicht schon deshalb, weil ein Chatbot überhaupt von

[114] Neu hinzukommen durch den Digital-Omnibus KI Vorschriften zur Verhinderung sexualisierter Deepfakes (Art. 5 Abs. 1 UAbs. 1 lit ba sowie UAbs. 1a KI-VO), vgl. Art. 1 Nr. 7 VO (EU) 2026/1744 (Digital-Omnibus-Verordnung zur KI).

vulnerablen Personen genutzt wird oder weil er Nutzerpräferenzen personalisiert. Die reguläre Personalisierung, eine überzeugende, aber nicht täuschende Werbung und transparente Empfehlungssysteme bleiben zulässig. Die Kommission nennt als Gegenbeispiel ausdrücklich einen therapeutischen Chatbot, der unterschwellige Techniken einsetzt, um Nutzer:innen zu einem gesünderen Lebensstil zu bewegen; hier fehlt es typischerweise am erheblichen Schaden, sodass keine verbotene Praxis vorliegt.[115] Problematisch ist außerdem der Beweis der Kausalität von Verhaltensbeeinflussung durch das System und Schadensdrohung oder -eintritt.

Verbot von Social Scoring u.Ä., Art. 5 Abs. 1 lit. c KI-VO

Eine weitere verbotene Praktik ist gemäß Art. 5 Abs. 1 lit. c KI-VO das Inverkehrbringen, die Inbetriebnahme oder die Verwendung von KI-Systemen zur Bewertung oder Einstufung natürlicher Personen oder Gruppen von Personen über einen bestimmten Zeitraum auf der Grundlage ihres sozialen Verhaltens oder bekannter, abgeleiteter oder vorhergesagter persönlicher Merkmale oder Persönlichkeitsmerkmale, wenn der so entstehende „Social Score“ zu einer Schlechterstellung oder Benachteiligung führt, die nicht im Zusammenhang mit den ursprünglich erhobenen Daten steht oder ungerechtfertigt oder unverhältnismäßig ist. Gemeint ist sowohl die soziale Bewertung von Personen auf der Grundlage ihres sozialen Verhaltens oder bekannter, abgeleiteter oder vorhergesagter persönlicher Eigenschaften oder Persönlichkeitsmerkmale.

Dies ist insofern relevant, als die bei der Nutzung von Chatbots anfallenden Daten je nach Nutzungsintensität die Erstellung eines äußerst umfassenden Persönlichkeitsbildes der nutzenden Person erlauben. Die Erstellung eines solchen Persönlichkeitsbildes ist gerade bei Companion-Apps sogar essenziell, um zusammen mit der Erfassung von „Erinnerungen“ die Illusion einer tiefgreifenden zwischenmenschlichen Interaktion zu ermöglichen, trifft aber auch auf General Purpose Chatbots zu, wenn sie als Companion genutzt werden. Genauso können die anfallenden Daten auch zur Grundlage für Social Scoring werden.

Allgemein gilt, dass durch immer häufigere und auch qualitativ hochwertigere Interaktion zwischen Mensch und Technik die Informationsdichte bei Anbieter/Betreiber erheblich zunimmt. Die Beschränkung der Nutzungsmöglichkeiten dieser Informationen ist in der Folge eine der wichtigsten Herausforderungen des derzeitigen und absehbaren technischen Fortschritts.

Klarstellend sei angemerkt: Das Verbot nach lit. c trifft nicht nur staatliche Stellen, sondern auch den privaten Sektor.[116] Es hat einen weiten Anwendungsbereich und ist nicht auf einen bestimmten Sektor oder Bereich beschränkt.[117] Es tritt dabei neben weitere Verpflichtungen, zum Beispiel aus DSGVO und Lauterkeitsrecht.

Mit dem Verbot nach Art. 5 Abs. 1 lit. c KI-VO könnte insbesondere das systemische Risiko der Profilbildung (S3) abgemildert werden. Es wirkt grundsätzlich aber auch hinsichtlich aller individuellen Risiken (I1 bis I3) und des systemischen Risikos der „kommerziellen Verhaltenssteuerung“ (S2).

---

[115] Europäische Kommission, Leitlinien zu verbotenen Praktiken der künstlichen Intelligenz, 2025, Rn. 134.

[116] Hilgendorf/Härtlein, in: Hilgendorf/Härtlein, KI-VO, Art. 5 Rn. 34.

[117] Europäische Kommission, Leitlinien zu verbotenen Praktiken der künstlichen Intelligenz, 2025, Rn. 146.

Die tatbestandlichen Eingangsvoraussetzungen des Verbots sind bei intensiv genutzten Chatbots regelmäßig erfüllt. Die Auswertung der Nutzenden erfolgt über einen längeren Zeitraum, sie stützt sich auf abgeleitete persönliche Eigenschaften und Persönlichkeitsmerkmale, und sie erfasst gerade dasjenige soziale Verhalten, das die nutzende Person in der Annahme eines vertraulichen Gesprächs offenbart. Diese Informationen über soziales Verhalten kann ein GPC (oder Companion) für eine Einstufung oder eine Bewertung nutzen. Tatbestandsmäßig erfasst sind Bewertungen oder Einstufungen auf der Grundlage sozialen Verhaltens oder persönlichen Eigenschaften oder Persönlichkeitsmerkmale bei Zuweisung von Punktzahlen.[118] Die erzeugte Punktzahl kann verschiedene Formen annehmen, beispielsweise als Zahl (z. B. von 0 bis 1), als Ranking oder als Label. Eine Einstufung muss nicht unbedingt zu einer Bewertung führen.[119]

Untersagt ist die Verwertung einer Bewertung in sozialen Zusammenhängen, die mit den Umständen der ursprünglichen Datenerhebung nichts zu tun haben. Gespräche mit GPC und Companions können sehr kontextgebunden sein, man denke an Gespräche über Trauer, Krankheit oder Schulden. Hinzu kommt ein regulatorischer Vorzug, der das Verbot von den meisten übrigen Instrumenten unterscheidet: Es greift ohne Nachweis eines individuellen Schadens und ohne Kausalitätskette zwischen einer konkreten Ausgabe und einer konkreten Beeinträchtigung. Es ist damit ein Instrument, das auf der systemischen Ebene ansetzt.

Die Anwendbarkeit des Verbots scheitert praktisch jedoch zumeist an der tatbestandlich geforderten Rechtsfolge. Es genügt nicht, dass bewertet wird. Zum einen muss an die Einstufung eine Schlechterstellung oder Benachteiligung anknüpfen. Diese muss entweder in einem sozialen Zusammenhang eintreten, der mit den Umständen der ursprünglichen Datenerhebung nichts zu tun hat (Ziff. i) oder im Hinblick auf das bewertete Verhalten ungerechtfertigt oder unverhältnismäßig sein (Ziff. ii).

Fehlt es am Nachteil, greift das Verbot nicht – unabhängig davon, wie umfassend das Profil ist. Die Grenze verläuft damit nicht zwischen umfassender und zurückhaltender Profilbildung, sondern zwischen Verwertung zum Nachteil und sonstiger Verwertung. Die Vorschrift trifft den Betreiber, der ein aus intimen Gesprächen gewonnenes Profil in fremde Lebensbereiche weiterreicht.

Das Verbot trifft daher in der Regel nicht den Betreiber eines GPC oder Companions, der dasselbe Profil nutzt, um die Bindung zu vertiefen, die das Profil überhaupt erst ermöglicht hat. Die im Risikokapitel beschriebene Profilbildung verfolgt nämlich gerade den gegenteiligen Zweck. Sie dient dazu, die Ansprache angenehmer, die Antworten zustimmender und die Bindung enger zu machen. Aus der Perspektive der nutzenden Person stellt sich das im Moment der Interaktion nicht als Nachteil, sondern als Zuwendung dar.

Ob das Verbot für ein GPC greift, ist mithin eine Frage des konkreten Verwendungszusammenhangs. Die Speicherung von Vorlieben und früheren Gesprächsinhalten zur Verbesserung der Antwortqualität durch ein GPC wäre an sich nicht erfasst, da es an einer Schlechterstellung an sich fehlt. Auch die Anpassung von Tonfall, Detailtiefe und Ansprache an den erkannten Gefühlszustand wäre an sich nicht vom Verbot erfasst, soweit die Einstufung an sich noch an keine Schlechterstellung oder Benachteiligung geknüpft ist. Eine Einschränkung

[118] Europäische Kommission, Leitlinien zu verbotenen Praktiken der künstlichen Intelligenz, 2025, Rn. 151.

[119] Europäische Kommission, Leitlinien zu verbotenen Praktiken der künstlichen Intelligenz, 2025, Rn. 153.

bestimmter Ausgaben nach Erkennung einer Krisensituation durch das GPC (oder den Companion), verbunden mit dem Hinweis auf Hilfsangebote, wäre ebenfalls nicht erfasst. Das Verbot des lit. c steht schon dem Sinn und Zweck nach Schutzmaßnahmen nicht entgegen.

Dagegen wäre die Nutzung eines aus Gesprächen abgeleiteten Gesundheits-, Bonitäts- oder Persönlichkeitsprofils, um damit Entscheidungen über die Nutzer:in zu treffen, in der Regel vom Verbot erfasst. Dies kann auch die kollusive Weitergabe an Versicherer, Kreditgeber oder Arbeitgeber, die zu einer Schlechterstellung der Nutzer:in führt, einschließen. Aus der Weitergabe entstehende Nachteile treten in einem Zusammenhang ein, der mit dem vertraulichen Gespräch nichts zu tun hat.[120] Dies sind Kernfälle des Verbots von Ziff. i.[121]

So wäre die Verwendung des Profils beim GPC für ein Scoring in einem anderen Dienst desselben Anbieters, etwa im Zahlungsverkehr oder auf einem Marktplatz, potentiell auch erfasst von Art. 5 Abs. 1 lit. c Ziff. i KI-VO. Maßgeblich ist der soziale Zusammenhang, nicht die Identität des Verwenders. Die Konzernzugehörigkeit stellt den Kontextbezug nicht her. Auch höhere Abonnementpreise oder schlechtere Konditionen für Nutzende, bei denen eine besonders starke Bindung oder eine emotionale Notlage festgestellt wurde, wären in diesem Zusammenhang verboten. Der Nachteil knüpft an eine Verletzlichkeit an, nicht an ein Verhalten, zu dem er in einem Verhältnis stünde. Auch vom Verbot erfasst ist das systematische Zurverfügungstellen eines geringeren Funktionsumfangs für eine Kohorte, die nach abgeleiteten Persönlichkeitsmerkmalen gebildet wurde. Denn Ziff. ii erfasst ausdrücklich auch die Einstufung von Gruppen.

Eine zweite Hürde in der Anwendbarkeit des Verbots liegt im Erfordernis einer sozialen Bewertung. Die Vorschrift setzt ein Ergebnis voraus, das als Einstufung erkennbar ist und einer Person zugeordnet werden kann. Was ein Sprachmodell über seine Nutzenden ableitet, existiert demgegenüber nicht als ausgewiesener Wert, sondern als verteilte Repräsentation im Verarbeitungsvorgang. Solange der Anbieter das Profil ausschließlich innerhalb des eigenen Dienstes verwendet, bleibt der Verwendungszusammenhang derselbe, in dem die Daten entstanden sind. Das Verbot wirkt aber nachgelagert, also bei der Weitergabe an Dritte oder der Verwendung für fremde Zwecke wie Bonitäts- oder Risikobewertung.

Hochriskante KI-Systeme

Ob Chatbots als Hochrisiko-KI einzuordnen sind, bestimmt sich nach Art. 6 Abs. 1 und 2 in Verbindung mit Anhang III KI-VO. Das sind gemäß Art. 6 Abs. 1 KI-VO einerseits KI-Systeme, die als Sicherheitsbauteil (Art. 3 Nr. 14 KI-VO) eines Produkts dienen, das den in Anhang I der KI-Verordnung aufgeführten Harmonisierungsvorschriften unterfällt, zum Beispiel Spielzeug oder Medizinprodukte,[122] oder selbst ein solches Produkt darstellen. Zudem gelten die in Art. 6 Abs. 2

[120] Leitlinien der Kommission, C(2025) 5052 final, Rn. 57 mit weiteren Beispielen.

[121] Vgl. Raue, in: Schefzig/Kilian, BeckOK KI-Recht, KI-VO, Art. 5 Rn. 106;

[122] Selbst für therapeutische Zwecke angebotene KI-Companions gelten in der Regel laut Herstellerangaben nicht als Medizinprodukt. Dies ist jedoch äußerst fragwürdig und bedarf weiterer Prüfung. S. zu KI-gestützter Software als Medizinprodukt und den Schwierigkeiten bei deren Einordnung Saliba, in: Chibanguza/Kuß/Steege, Künstliche Intelligenz, § 4 F Rn. 7 f.

iVm Anhang III der KI-VO gelisteten Systeme als hochriskant. Weder General Purpose Chatbots noch KI-Companions sind direkt genannt.[123]

Denkbar ist, dass ein KI-System als Chatbot in einem Spielzeug verbaut ist und damit als Hochrisiko-KI in den Anwendungsbereich fällt. Voraussetzung ist, dass das KI-System als Sicherheitsbauteil qualifiziert werden kann (Art. 6 Abs. 1 lit. a KI-VO). Ein Sicherheitsbauteil ist gemäß Art. 3 Nr. 14 KI-VO ein „Bestandteil eines Produkts oder KI-Systems, der eine Sicherheitsfunktion für dieses Produkt oder KI-System erfüllt oder dessen Ausfall oder Störung die Gesundheit und Sicherheit von Personen oder Eigentum gefährdet". Bei einem in einem Spielzeug eingebauten Empfehlungssystem, zum Beispiel für Musik, würde der Ausfall aller Voraussicht nach nicht zu „nachteiligen Auswirkungen auf die Gesundheit oder Sicherheit von Personen"[124] führen.[125] Ist jedoch ein Chatbot in einem Spielzeug integriert, das mit Minderjährigen interagiert, kann beispielsweise der Ausfall eines Sicherheitsprotokolls zu erheblichen Schäden bei den Kindern führen, wenn etwa unangemessener Output wie sexualisierte Konversationen, Tipps für selbst- und fremdgefährdendes Verhalten oder ähnliches ausgespielt wird.[126] Das KI-System könnte also dann als Sicherheitsbauteil zu qualifizieren sein, wenn es speziell dazu bestimmt ist, Sicherheitsrisiken zu adressieren, die schon allgemein vom Produkt oder System ausgehen.[127] Ein Bestandteil des KI-Systems muss demnach Funktionen vorhalten, um einen altersgerechten Funktionsumfang des Spielzeugs zu gewährleisten.

Eine weitere denkbare Ausnahme ist gemäß Art. 6 Abs. 2 iVm Anhang III Nr. 1 lit. c KI-VO die Emotionserkennung durch KI-Systeme. Emotionserkennungssysteme sind gemäß Art. 3 Nr. 39 KI-VO KI-Systeme, die dem Zweck dienen, „Emotionen [...] natürlicher Personen auf der Grundlage ihrer biometrischen Daten festzustellen oder abzuleiten". Biometrische Daten sind gemäß Art. 3 Nr. 34 KI-VO Daten zu physischen, physiologischen oder verhaltenstypischen Merkmalen einer Person, die durch spezielle Verfahren festgestellt oder abgeleitet werden. Meist wird durch Kontextanalyse auf die Emotionen der Nutzenden geschlossen, ohne biometrische Daten festzustellen oder abzuleiten. Solche Anwendungen sind keine Emotionserkennungssysteme im Sinne des Art. 3 Nr. 39 KI-VO.[128] Vereinzelt kann die Beurteilung jedoch anders ausfallen. So schließt beispielsweise ChatGPT im erweiterten Sprachmodus (GPT-4o) durch Stimmanalyse auf die Emotionen der Nutzenden.[129] Um eine Hochrisiko-KI handelt es sich aber gemäß Art. 6 Abs. 2 iVm Anhang III Nr. 1 lit. c KI-VO nur dann, wenn das System bestimmungsgemäß zur Emotionserkennung eingesetzt wird. Da der Chatbot aus der Stimme des Nutzenden Emotionen ableitet, um hierauf im Chatverlauf zu reagieren, handelt es sich bei diesem Szenario um einen

---

[123] Molavi-Vasse'i 2026, 62 spricht von einer „erhebliche[n] Regelungslücke" und Widerspruch zum Schutzanspruch der Verordnung.

[124] Vgl. Erwägungsgrund 47 KI-VO.

[125] Vgl. Europäische Kommission, Entwurf der Leitlinien für die Einstufung von Hochrisiko-KI-Systemen, Annex I, 19.5.2026, Rn. 47, https://digital-strategy.ec.europa.eu/de/library/draft-commission-guidelines-classification-high-risk-ai-systems.

[126] Beispiele aus Murray/Cross/Erlich/Tracy/Mela, Trouble in Toyland 2025, https://publicinterestnetwork.org/wp-content/uploads/2025/11/TOYLAND-2025-11-14-7a.pdf.

[127] Zur Definition der Sicherheitsfunktion Kirschke-Biller/Füllsack, in: Schefzig/Kilian, BeckOK KI-Recht,KI-VO Art. 3 Rn. 261.

[128] Ausführlich zu der Frage, ob AI Companions als Emotionserkennungssysteme zu kategorisieren sind Frei, RDi 2026, 441 ff.

[129] Noone, OpenAI launches GPT-4o, flaunting ability of model to detect user emotions, TechMonitor vom 14.5.2026, https://www.techmonitor.ai/digital-economy/ai-and-automation/openai-launches-gpt-4o-flaunting-ability-of-model-to-detect-user-emotions.

bestimmungsgemäßen Einsatz des Systems. Die Rückausnahme des Art. 6 Abs. 3 KI-VO ist nicht einschlägig, so dass der erweiterte Sprachmodus in ChatGPT als Hochrisiko-KI einzustufen ist.[130]

Transparenzanforderungen

Da den hier betrachteten Chatbots KI-Systeme zugrunde liegen, die für die direkte Interaktion mit natürlichen Personen bestimmt sind, besteht für die Anbieter eine Kennzeichnungspflicht gemäß Art. 50 Abs. 1 Satz 1 KI-VO. Die betreffenden natürlichen Personen sind also darüber zu informieren, dass sie mit einer KI interagieren. Eine Ausnahme der Kennzeichnungspflicht besteht gemäß Art. 50 Abs. 1 Satz 1 KI-VO, wenn aus Sicht „einer angemessen informierten, aufmerksamen und verständigen natürlichen Person aufgrund der Umstände und des Kontexts der Nutzung offensichtlich“ ist, dass mit einer KI interagiert wird. Die Informationspflicht soll sich auf solche Systeme begrenzen, bei denen eine reale Verwechslungsgefahr besteht.[131] Dabei wird auf einen verständigen Durchschnittsnutzer abgestellt.[132] Zugunsten von Minderjährigen oder Menschen mit Behinderung ist der Maßstab anzupassen.[133] Als Anwendungen, die von der Kennzeichnungspflicht ausgenommen sein sollen, werden Roboter und autonome Fahrzeuge genannt,[134] womöglich, weil keine Verwechslungsgefahr besteht (wenngleich diese Logik vor dem Hintergrund von Telepräsenz-Robotern nur bedingt greift). Chatbots hingegen unterfallen dieser Ausnahme nicht.[135] Während sich Art. 50 Abs. 1 und 3 KI-VO mit der Kennzeichnung des KI-Systems selbst befassen, sind Abs. 2 und 4 mit der Kennzeichnung des Outputs befasst. Der im Juni 2026 von der Kommission als freiwilliges Instrument vorgelegte Verhaltenskodex für die Transparenz von KI-generierten Inhalten ist auf die Kennzeichnung des Outputs beschränkt. Wie Art. 50 Abs. 4 Satz 4 KI-VO zeigt (Text, „der veröffentlicht wird, um die Öffentlichkeit über Angelegenheiten von öffentlichem Interesse zu informieren“) geht es hier um Situationen, in denen KI-generierter Output den Kontext seiner Genese verlässt; beispielsweise indem ein KI-generiertes Bild auf Social Media hochgeladen wird – wenngleich Art. 50 Abs. 4 Satz 1 KI-VO keinen Hinweis auf eine Veröffentlichung enthält. Eine ständige Markierung des Outputs als KI-generiert innerhalb eines Chatverlaufes ist jedoch im Falle textbasierter Interaktionen klar nicht gemeint.

Dennoch ist fraglich, ob die Kennzeichnungspflicht ausreicht, um den Risiken von General Purpose Chatbots adäquat zu begegnen. Bei diesen Chatbots ist zwar anzunehmen, dass sich Nutzende bewusst sind, dass es sich um KI handelt; das ist zumindest im Rahmen der Registrierung bei dem Dienst für Nutzende zunächst offensichtlich. Insbesondere bei langanhaltender Nutzung und gegebenenfalls damit zusammenhängender emotionaler Abhängigkeit besteht jedoch die Gefahr, dass die kritische Reflexion des Nutzenden nachlässt und in Vergessenheit gerät, dass der Gesprächspartner eine KI statt eines realen Menschen ist. Vielmehr müssten Nutzende bei jedem

---

[130] OpenAI nimmt weder in seinen Datenschutzerklärungen noch in angrenzenden Dokumenten dazu Stellung, dass biometrische Daten verarbeitet werden, Emotionserkennung stattfindet und welche Maßnahmen diesbezüglich getroffen wurden.

[131] Martini, in: Martini/Wendehorst, Art. 50 KI-VO Rn. 62.

[132] Zur Kritik an diesem Begriff zum Beispiel Martini, in: Martini/Wendehorst, Art. 50 KI-VO Rn. 63; Lauber-Rönsberg, in: BeckOK KI-Recht, Art. 50 KI-VO Rn. 20.

[133] Erwägungsgrund 132 Satz 4 KI-VO.

[134] Martini, in: Martini/Wendehorst, Art. 50 KI-VO Rn. 63; Lauber-Rönsberg, in: BeckOK KI-Recht, Art. 50 KI-VO Rn. 22.

[135] Martini, in: Martini/Wendehorst, Art. 50 KI-VO Rn. 63; Lauber-Rönsberg, in: BeckOK KI-Recht, Art. 50 KI-VO Rn. 22. AA wohl Hilgendorf/Härtlein, HK-KI-VO, Art. 50 KI-VO Rn. 4.

Start der Anwendung sowie in regelmäßigen Abständen darauf hingewiesen werden, dass sie mit einer KI interagieren[136] – möglicherweise auch in bestimmten Kontexten; zum Beispiel wenn eine nutzende Person kurz davor steht, besonders sensible Informationen zu offenbaren und das System dies bereits vorab erkennt. Denkbar wäre darüber hinaus auch eine weitergehende Hinweispflicht zu möglichen gesundheitlichen Risiken oder Abhängigkeiten, Aufforderung zum Einlegen einer Pause[137] oder Hilfsangeboten bei Hinweisen auf psychische Probleme oder Suizidabsichten.[138] Jüngst wurden in Kalifornien zudem weitere relevante Maßnahmen verabschiedet: Hier sind zunächst (neben den bereits angesprochenen) Krisenprotokollen für den Fall von Suizidgedanken auch Kindersicherungen sowie Benachrichtigungen enthalten, falls Sicherheitseinstellungen deaktiviert werden. Das Gesetz verpflichtet Anbieter ferner erstmals, unabhängige Kindersicherheitsprüfungen und jährliche Risikobewertungen durchzuführen.[139] Zudem ist „persistent conversational memory“ by default zu deaktivieren. Chatbots dürfen mithin standardmäßig keine dauerhafte Erinnerungsfunktion für Chatverläufe nutzen, sofern keine elterliche Einwilligung vorliegt.[140] Auf Chatverläufen basierende Werbung („target advertising“) ist untersagt.[141] Ferner wurde ein Verbot für „companion chatbots“ enthaltendes Spielzeug erlassen.[142] Zu beachten ist dabei, dass Sec. 22601 (b) (1) des Business and Professions Code einen „Companion chatbot“ definiert als „an artificial intelligence system with a natural language interface that provides adaptive, human-like responses to user inputs and is capable of meeting a user’s social needs, including by exhibiting anthropomorphic features and being able to sustain a relationship across multiple interactions“. General Purpose Chatbots sind damit nicht per se miterfasst.

Abschließend sei noch auf den möglichen Wertungswiderspruch hingewiesen, der entsteht, wenn selbst eine umfassende manuelle Bildbearbeitung mit einem Programm wie Adobe Photoshop keine Kennzeichnungspflichten erzeugt, diese aber bereits bei Detailverbesserungen eines Bildes greifen, sofern diese KI-gestützt erfolgen.

---

[136] Dies wird etwa vorgesehen in Sec. 1702 des New Yorker AI Companion Model Law sowie in Sec. 22602 des California Senate Bill 243 (einer Ergänzung des Business and Professions Code) bei Minderjährigen Nutzenden. Letztere gilt seit dem 1.1.2026.

[137] Ebenso in Sec. 22602 des California Senate Bill 243 sowie in Art. 17 des chinesischen Gesetzentwurfs für das Management von anthropomorphen interaktiven Diensten für künstliche Intelligenz. Sec. 22602(c) SB-243: „An operator shall, for a user that the operator knows is a minor, do all of the following: […] (2) Provide by default a clear and conspicuous notification to the user at least every three hours for continuing companion chatbot interactions that reminds the user to take a break and that the companion chatbot is artificially generated and not human.“ Subsection (1) ist als Äquivalent zu Art. 50 KI-VO zu verstehen, während (3) sexuell expliziten (und hier lediglich visuellen) Output und Aufforderungen zu sexuellen Handlungen betrifft (beides jedoch beschränkt auf Minderjährige).

[138] So etwa in Sec. 1701 des New Yorker AI Companion Model Law.

[139] SB-1119 Companion chatbots: children’s safety; Sec. 21810 wird dabei bezeichnet als „Adam‘s Law“ nach dem Fall „Adam Raine“.

[140] Sec. 21812 (d) (3) (A) (i) (I) SB-1119.

[141] Sec. 21813. (a) (1) (B) (i) SB-1119.

[142] SB-867 Toys: companion chatbots: 22604.5. Business and Professions Code: „(a) No person shall manufacture, sell, or exchange, possess with intent to sell or exchange, or expose or offer for sale or exchange to any retailer any toy that includes a companion chatbot.“

2.3 Zwischenfazit

Insgesamt zeigt sich, dass die KI-Verordnung nicht in der Lage ist, die von Chatbots ausgehenden Risiken adäquat zu erfassen. Hier zeigt sich in besonderem Maße die Dynamik der technischen Entwicklung seit der Genese der Verordnung. Diese Entwicklung konnte schlicht nicht antizipiert werden und zahlreiche der Vorschriften der KI-Verordnung zeigen sich als zu inflexibel, um sie vollumfänglich oder auch nur überwiegend aufzufangen. Besonders problematisch ist dabei, dass gerade subtile Manipulation und emotionale Abhängigkeit und die daraus resultierende Vulnerabilität nur unzureichend erfasst werden. Die KI-Verordnung ist klar auf bestimmte Praktiken ausgerichtet, die zum Zeitpunkt ihrer Genese die gesellschaftliche, mediale und politische Diskussion bestimmten; seither neu hinzugetretene Praktiken lassen sich mitunter schwer unter ihre Vorgaben subsumieren oder sind schlicht außerhalb ihres designierten Wirkbereichs angesiedelt.

Zwar kennen die USA bisher keinen der KI-Verordnung entsprechenden rechtlichen Rahmen für KI-Systeme, die bereichsspezifische Regulierung auf Ebene der Bundesstaaten ist jener in der EU jedoch mitunter deutlich voraus. Einschränkend wirkt hier jedoch, dass sich hier eine starke Fokussierung auf Minderjährige zeigt, die ausblendet, dass auch Erwachsene bei der Nutzung von Chatbots mitunter eine starke Vulnerabilität aufweisen.

3. Datenschutzrecht

3.1 Allgemein

Die Datenschutz-Grundverordnung (DSGVO)[143] enthält Regeln zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Verkehr solcher Daten.[144] Als unionale Verordnung ist sie in allen Mitgliedsstaaten der EU direkt anwendbar. In Deutschland wird die DSGVO unter anderem ergänzt durch das Bundesdatenschutzgesetz (BDSG)[145] und das Telekommunikation-Digitale-Dienste-Datenschutz-Gesetz (TDDDG).[146]

Datenschutz-Grundverordnung

Die DSGVO findet auf General Purpose Chatbots Anwendung, sobald personenbezogene Daten der Nutzenden verarbeitet werden. Dies gilt hinsichtlich der personenbezogenen Daten, die Nutzenden von und über sich selbst angeben. Dies gilt auch für Daten Dritter, die diese dem Chatbot mitteilen. Es gilt auch für personenbezogene Daten natürlicher Personen, die ein Chatbot aus anderen Quellen, insbesondere dem Internet, verarbeitet.

---

[143] Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung), ABl. L 119 vom 4.5.2016, S. 1-88.

[144] Ausführlich zur DSGVO und deren Verhältnis zu andern Digitalrechtsakten Nebel, in: Geminn/Johannes, Europäisches Datenrecht 2025, § 32.

[145] Bundesdatenschutzgesetz vom 30. Juni 2017 (BGBl. I S. 2097), das zuletzt durch Artikel 3 des Gesetzes vom 3. Juli 2026 (BGBl. 2026 I Nr. 199) geändert worden ist.

[146] Telekommunikation-Digitale-Dienste-Datenschutz-Gesetz vom 23. Juni 2021 (BGBl. I S. 1982; 2022 I S. 1045), das zuletzt durch Artikel 3 des Gesetzes vom 10. März 2026 (BGBl. 2026 I Nr. 64) geändert worden ist.

Personenbezogene Daten können im Kontext der Informationen, die Nutzende einem Chatbot mitteilen, vielfältig sein. Nach Art. 4 Nr. 1 DSGVO sind „personenbezogene Daten“ alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. Als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind, identifiziert werden kann. Der Begriff ist dabei nach ständiger Rechtsprechung weit auszulegen und erfasst nicht nur objektive Sachverhaltsangaben, sondern auch subjektive Einschätzungen, Meinungen und Bewertungen.[147]

Als Regelfall bei der Nutzung von Chatbots sei die unmittelbare Selbstoffenbarung angenommen. Teilt eine Nutzerin dem System mit, sie heiße Maria Schneider (dies ggf. schon bei der Anlegung eines Nutzeraccounts zum Chatbot), sei Prokuristin der X-GmbH in Kassel und benötige Formulierungshilfe für ein Schreiben an ihren Arbeitgeber wegen eines seit Monaten schwelenden Konflikts mit der Geschäftsführung, so ist sie über die Kennung „Name“ bzw. ihr Profil bereits direkt identifiziert. Sämtliche im selben Kontext mitgeteilten Informationen – Beschäftigungsverhältnis, berufliche Stellung, Konfliktlage, Bewertung des Verhaltens Dritter – beziehen sich sodann auf diese identifizierte Person und sind personenbezogene Daten. Nichts anderes gilt für die Prompts selbst, denn der EuGH hat zuletzt klargestellt, dass Stellungnahmen, die persönliche Meinungen oder Sichtweisen ihres Verfassers wiedergeben, schon deshalb Informationen über diese Person darstellen, weil sie als Ausdruck ihrer Gedanken untrennbar mit ihr verknüpft sind.[148]

Personenbezug kann jedoch auch durch Inferenz entstehen, also durch Rückschlüsse aus der Kombination für sich genommen unverfänglicher Angaben. Nach Erwägungsgrund 26 DSGVO sind hierfür alle Mittel zu berücksichtigen, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um die Person direkt oder indirekt zu identifizieren.[149] Im Kontext der Chatbotnutzung sind zum Beispiel die folgenden Fallgruppen denkbar:

Individualisierung durch Merkmalskombination („singling out“): Schildert ein Nutzer, er sei der einzige niedergelassene Kinderkardiologe in einer namentlich genannten nordhessischen Kleinstadt und streite mit seiner Kassenärztlichen Vereinigung über die Abrechnung einer bestimmten Leistungsziffer, so ist er ohne jede Namensnennung eindeutig bestimmbar, weil die Verknüpfung von Fachgebiet, Ort und Konfliktgegenstand die Menge der in Betracht kommenden Personen auf eine einzige reduziert.

Rekonstruktion aus Restinformationen in eingegebenen oder hochgeladenen Inhalten: Wird ein vermeintlich anonymisierter Vertragsentwurf, Rechnungen oder ein Behördenschreiben zur Überarbeitung eingespeist, verbleiben typischerweise Aktenzeichen, Objektadressen, Vertragsdaten, Kennzeichen oder Registernummern, die für jeden Empfänger mit Zugang zu den

[147] EuGH, Urt. v. 20.12.2017 – C-434/16, Rn. 34, Nowak.

[148] EuGH, Urt. v. 04.09.2025 – C-413/23 P, EDSB/SRB.

[149] Vgl. auch EuGH, Urt. v. 19.10.2016 – C-582/14, Breyer. Zum Personenbezug zB Karg, in: Simitis/Hornung/Spiecker gen. Döhmann, DSGVO/BDSG, Art. 4 Rn. 24 ff.

entsprechenden Registern oder Zusatzinformationen eine Zuordnung erlauben. Ob dies zum Personenbezug führt, ist nach dem vom EuGH bestätigten relativen Maßstab aus der jeweiligen Perspektive des verarbeitenden Akteurs zu beurteilen, wobei für den Verantwortlichen, der über Zuordnungsmittel verfügt, der Personenbezug regelmäßig zu bejahen ist.[150] Dies wird für Betreiber von Chatbots oft zu bejahen sein.

Verknüpfung über technische Kennungen und Nutzungsmuster: Auch bei durchgehend pseudonymer Nutzung entstehen Account- und Session-Kennungen, IP- und Gerätekennungen sowie Zeitstempel, die den Konversationsverlauf zu einem persistenten Profil zusammenführen. Treten wiederkehrende inhaltliche Marker (Wohnort, Arbeitgeber, Familienkonstellation, Reisezeitpunkte) oder stilistische und sprachliche Eigenheiten hinzu, verdichtet sich das Profil zu einem Datensatz, der die betroffene Person aus der Gesamtheit der Nutzenden heraushebt und damit identifizierbar macht.

Bundesdatenschutzgesetz

Das BDSG enthält in Teil 2 Vorschriften, die die Regelungen der DSGVO modifizieren bzw. ergänzen. Diese sind aufgrund von Öffnungsklauseln und Regelungsaufträgen der DSGVO erlassen. Sie gelten teilweise auch für privatrechtlich organisierte Verantwortliche („nichtöffentliche Stellen“ im Sprachgebrauch des BDSG). Damit gelten diese Regelungen auch für Betreiber von Chatbots, da diese automatisiert personenbezogene Daten verarbeiten (§ 1 Abs. 1 UAbs. 2 BDSG) und soweit der Anwendungsbereich des BDSG eröffnet ist.

Dies ist bei Betreibern mit Sitz in Drittländern fraglich. Das BDSG knüpft seinen räumlichen Geltungsbereich nicht eigenständig an, sondern folgt der Datenschutz-Grundverordnung. Nach § 1 Abs. 4 Satz 2 BDSG i.V.m. Art. 3 DSGVO unterliegen nichtöffentliche Stellen dem BDSG, auch wenn der Verantwortliche über keine Niederlassung in der Union oder im EWR verfügt, gleichwohl aber in den Anwendungsbereich der Verordnung fällt. Die dritte Variante des § 1 Abs. 4 Satz 2 Nr. 3 BDSG ist als Auffangtatbestand konzipiert und macht die Anwendbarkeit des nationalen Rechts von der Reichweite des Art. 3 DSGVO abhängig.

Für Verantwortliche ohne Niederlassung in der Union entscheidet das Marktortprinzip des Art. 3 Abs. 2 DSGVO. Nach Art. 3 Abs. 2 lit. a DSGVO genügt das Anbieten von Waren oder Dienstleistungen an betroffene Personen in der Union, wobei es nach Erwägungsgrund 23 auf ein Entgelt nicht ankommt. Erforderlich ist ein zielgerichtetes Ausrichten der Tätigkeit auf den Unionsmarkt. Maßgeblich sind insbesondere Sprache und Währung des Angebots, die verwendete Top-Level-Domain, die Ausgestaltung des Vertriebs und der Zahlungswege sowie eine auf den nationalen Markt bezogene Bewerbung. Dies ist jedenfalls hinsichtlich großer Betreiber von Chatbots der Fall, die ihren Dienst zielgerichtet für europäische Verbraucher anbieten, einschließlich der jeweiligen Landessprache und Zahlungen in Euro ermöglichen. Daneben eröffnet Art. 3 Abs. 2 lit. b DSGVO den Anwendungsbereich bereits dann, wenn das Verhalten betroffener Personen beobachtet wird, soweit dieses Verhalten in der Union erfolgt.

---

[150] Vgl. EuGH, Urt. v. 4.9.2025 – C-413/23 P, Rn. 68 ff., SRB/EDSB; ferner EuGH, Urt. v. 9.11.2023 – C-319/22, Rn. 46, Gesamtverband Autoteile-Handel, zur Fahrzeug-Identifizierungsnummer.

Über die Reichweite der Anwendbarkeit entscheidet sodann § 1 Abs. 5 BDSG. Danach treten die Vorschriften des Gesetzes zurück, soweit das Unionsrecht unmittelbar gilt. Das Bundesdatenschutzgesetz kommt folglich allein dort zum Tragen, wo die Verordnung dem nationalen Gesetzgeber durch eine Öffnungsklausel Regelungsspielraum belässt. Die Anwendbarkeit ist damit keine Frage des Ob, sondern des Umfangs, und sie ist normbezogen für jede in Betracht kommende Vorschrift gesondert zu beantworten.

Für die Verarbeitung besonderer Kategorien personenbezogener Daten macht § 22 BDSG von der Öffnungsklausel des Art. 9 Abs. 2 DSGVO Gebrauch. Die Vorschrift ist in ihrem Zuschnitt weitgehend auf öffentliche Stellen, das Beschäftigungsverhältnis und den Bereich der Gesundheitsversorgung und der sozialen Sicherheit ausgerichtet. Für nichtöffentliche Stellen außerhalb dieser Konstellationen verbleibt es regelmäßig bei den Erlaubnistatbeständen der Verordnung, insbesondere bei der ausdrücklichen Einwilligung nach Art. 9 Abs. 2 lit. a DSGVO.

Praktisch bedeutsamer sind die auf Art. 23 DSGVO gestützten §§ 32 bis 37 BDSG, die die Informations-, Auskunfts-, Lösch- und Widerspruchsrechte der betroffenen Person beschränken. Im Vordergrund steht § 34 Abs. 1 Nr. 2 lit. b BDSG in Verbindung mit § 29 Abs. 1 Satz 2 BDSG, wonach das Auskunftsrecht entfällt, soweit die Daten ihrem Wesen nach wegen überwiegender berechtigter Interessen eines Dritten geheim gehalten werden müssen. Die Vorschrift ist im Lichte des Art. 23 Abs. 2 DSGVO und des Erwägungsgrunds 63 eng auszulegen. Eine pauschale Berufung auf Geschäftsgeheimnisse genügt nicht. Erforderlich ist eine auf die konkret begehrte Information bezogene Abwägung.[151] Entsprechendes gilt für die Beschränkungen der Informationspflichten nach den §§ 32 und 33 BDSG sowie für § 35 BDSG, der das Löschungsrecht bei unverhältnismäßigem Aufwand durch eine Einschränkung der Verarbeitung ersetzt.

Telekommunikation-Digitale-Dienste-Datenschutz-Gesetz

Chatbots unterfallen dem TDDDG, weil sie digitale Dienste im Sinne des § 2 Abs. 2 Nr. 1 TDDDG i.V.m. § 1 Abs. 4 Nr. 1 DDG und Art. 1 Abs. 1 lit. b RL (EU) 2015/1535 sind. Sämtliche Merkmale des Dienstes der Informationsgesellschaft sind erfüllt. Die Leistung wird elektronisch erbracht, weil Eingabe, Verarbeitung und Ausgabe vollständig über Server und Datenübertragungstechnik erfolgen. Sie wird im Fernabsatz erbracht, weil Nutzer:innen und Betreiber nicht gleichzeitig körperlich anwesend sind. Die Leistung wird auch auf individuellen Abruf des Empfängers erbracht, da die Ausgabe jeweils durch die konkrete Nutzeranfrage ausgelöst wird und nicht in einer zeitgleichen Verbreitung an eine unbestimmte Vielzahl von Empfängern besteht. Das Merkmal „in der Regel gegen Entgelt“ steht dem auch bei unentgeltlichen Angeboten nicht entgegen, da es nach gefestigter Rechtsprechung[152] auf die typische wirtschaftliche Natur der Dienstleistung ankommt und nicht auf die Zahlung durch den einzelnen Nutzer. Deswegen sind auch werbe- oder datenfinanzierte sowie in Freemium-Modellen bereitgestellte Chatbots erfasst. Das TDDDG gilt nach § 1 Abs. 3 TDDDG auch für Betreiber von Chatbots aus anderen EU-Staaten und Drittländern, soweit sie ihre Dienstleistungen gezielt in Deutschland erbringen. Dies wird für große Betreiber von bekannten Chatbots regelmäßig der Fall sein (siehe oben). Daraus folgt die Anwendbarkeit des Teils

[151] Wie sie der Gerichtshof in der Rechtssache Dun & Bradstreet Austria (EuGH, Urt. v. 27.2.2025, C-203/22) für die Offenlegung der Verarbeitungslogik verlangt hat.

[152] EuGH, Urt. v. 27.3.2013, C-291/13 – Papasavvas.

3 des TDDDG, insbesondere des § 25 TDDDG: Nahezu jeder web- oder appbasierte Chatbot speichert Informationen auf der Endeinrichtung des Endnutzers oder greift auf dort gespeicherte Informationen zu, sei es durch Session- und Authentifizierungstoken, Local Storage zur Persistenz von Konversationsverläufen, Präferenz- und Konfigurationsdaten oder eingebundene Analyse- und Werbe-SDKs. Da § 25 TDDDG technologieneutral formuliert und vom Personenbezug der Information unabhängig ist, greift das Einwilligungserfordernis unabhängig davon, ob der jeweilige Speichervorgang zugleich eine Verarbeitung personenbezogener Daten darstellt; zulässig ist ohne Einwilligung allein, was nach § 25 Abs. 2 Nr. 2 TDDDG für die vom Nutzer:innen ausdrücklich gewünschte Diensterbringung unbedingt erforderlich ist.

3.2 Risikenbewertung

Ein (textbasiertes) „Gespräch“ mit einem Chatbot besteht ausschließlich aus Daten. Wer schreibt, verarbeitet Daten über sich selbst – und zwar in einer Dichte und Offenheit, die kaum ein anderer Onlinedienst erreicht. Anders als eine Suchanfrage oder ein Formularfeld enthält ein Chatverlauf Zusammenhänge: Was jemand fühlt, fürchtet, plant, verschweigt. Für auf Spracheingabe basierte Interaktionen besteht zudem die Möglichkeit, aus der Stimme Gemütszustände etc. herauszulesen.

Das systemische Risiko „Datensicherheit und Profilbildung (S3)“ wird durch das geltende Datenschutzrecht abgemildert. Betreiber von Chatbots, die ihren Dienst in der Europäischen Union und/oder für Bürger der Europäischen Union anbieten, sind hinsichtlich der Verarbeitung der personenbezogenen Daten, die sie verarbeiten, Verantwortliche. Inhaltlich adressiert die DSGVO einen erheblichen Teil der in Kapitel II beschriebenen Risiken.

Die Grundsätze des Art. 5 DSGVO verlangen Rechtmäßigkeit, Transparenz und Zweckbindung, begrenzen die Datenmenge auf das Erforderliche (Datenminimierung) und die Speicherung auf das notwendige Maß (Speicherbegrenzung). Diese Grundsätze gelten unabhängig davon, ob eine betroffene Person Schaden nimmt. Sie treffen also gerade jene systemische Ebene, auf der eine individuelle Kausalitätskette nicht nachweisbar ist.

Art. 9 DSGVO stellt u.a. Gesundheitsdaten, Angaben zur sexuellen Orientierung, zu religiösen und politischen Überzeugungen unter ein grundsätzliches Verarbeitungsverbot mit Erlaubnisvorbehalt.

Art. 25 DSGVO verpflichtet Anbieter zu datenschutzfreundlicher Technikgestaltung und – praktisch bedeutsamer – zu datenschutzfreundlichen Voreinstellungen. Das ist der normative Gegenpol zu dem im Risikokapitel beschriebenen Gestaltungsprinzip einer bewusst auf Bindung angelegten Systemarchitektur.

Art. 32 DSGVO verlangt angemessene Sicherheitsmaßnahmen, Art. 35 DSGVO bei riskanten Verarbeitungen eine vorherige Datenschutz-Folgenabschätzung. Bei einem System, das psychologische Profile aus intimen Gesprächen bildet, dürfte Letzteres regelmäßig verpflichtend sein.

Die Betroffenenrechte der Art. 15 bis 21 DSGVO geben dem Einzelnen Auskunfts-, Berichtigungs-, Lösch- und Widerspruchsrechte.

Dass dies keine bloße Theorie ist, zeigt die bisher wichtigste aufsichtsbehördliche Entscheidung zu einem Companion-Chatbot: Die italienische Datenschutzbehörde verhängte gegen Luka Inc., den Betreiber von „Replika“, ein Bußgeld von fünf Millionen Euro. Beanstandet wurden unter anderem

das Fehlen einer klar zugeordneten Rechtsgrundlage, eine unzureichende Datenschutzerklärung und das vollständige Fehlen wirksamer Altersverifikation.[153] Zuvor hatte dieselbe Behörde gegen OpenAI ein Bußgeld von 15 Millionen Euro verhängt, ebenfalls wegen unzureichender Transparenz, fehlender Rechtsgrundlage und mangelhaften Minderjährigenschutzes.

Die Rolle des BDSG ist demgegenüber begrenzt. Die DSGVO ist im Verhältnis zwischen privaten Anbietern und Verbraucherinnen und Verbrauchern weitgehend vollharmonisiert; das BDSG darf nur ausfüllen, wo Öffnungsklauseln dies zulassen. Für den hier interessierenden Bereich bleibt ihm im Wesentlichen eine organisatorische Funktion: Es regelt die Zuständigkeit und Befugnisse der deutschen Aufsichtsbehörden und die gerichtliche Zuständigkeit für Klagen gegen Verantwortliche.[154] Eigenständige materielle Schutzstandards für Chatbot-Nutzung enthält es nicht. Vielmehr senkt es potenziell den materiellen Schutzstandard, indem es Ausnahmen und Sonderregelungen zu den Betroffenenrechten einführt. Die §§ 29 Abs. 1, 32 bis 35 BDSG beschränken, gestützt auf die Öffnungsklausel des Art. 23 DSGVO, die Informations-, Auskunfts- und Löschungsrechte. Diese Vorschriften richten sich ausdrücklich auch an nichtöffentliche Stellen und stehen damit jedem Chatbot-Betreiber offen, der in Deutschland verklagt oder von einer deutschen Aufsichtsbehörde in Anspruch genommen wird.

Praktisch bedeutsam ist § 29 Abs. 1 Satz 2 BDSG. Danach besteht das Auskunftsrecht nicht, soweit durch die Auskunft Informationen offenbart würden, die nach einer Rechtsvorschrift oder ihrem Wesen nach geheim gehalten werden müssen. Das Konfliktfeld liegt dabei nicht bei Art. 15 Abs. 1 lit. h DSGVO, dessen Anwendung eine automatisierte Entscheidung im Sinne des Art. 22 Abs. 1 und 4 DSGVO voraussetzt und für die bloße Erzeugung einer Dialogantwort aus den unten dargelegten Gründen zweifelhaft bleibt. Es liegt vielmehr bei Art. 15 Abs. 1 lit. a bis g und Abs. 3 DSGVO. Wer wissen will, welche Gesprächsinhalte gespeichert sind, welche Merkmale der Anbieter daraus abgeleitet hat, insbesondere den erfassten Gefühlszustand und darauf aufbauende Interessen- und Vulnerabilitätsmarker, an wen diese Merkmale übermittelt wurden und ob die Verläufe in Trainingskorpora eingegangen sind, fragt nach personenbezogenen Daten im weiten Sinne der Rechtsprechung des Gerichtshofs.[155] Er berührt damit zugleich die interne Merkmalssystematik des Anbieters, also einen Bereich, für den sich dieser auf Betriebsinterna berufen wird.

Diesem Rückzug sind allerdings enge Grenzen gesetzt. Der Verordnungsgeber hat Geschäftsgeheimnisse als Grenze des Auskunftsrechts durchaus anerkannt, zugleich aber klargestellt, dass dies nicht zur vollständigen Verweigerung jeder Auskunft führen darf.[156] Geschuldet ist eine präzise, transparente und verständliche Erläuterung des Verfahrens und der Grundsätze, nach denen die personenbezogenen Daten verarbeitet und das konkrete Ergebnis erzielt wurden, nicht hingegen die Offenlegung des Algorithmus selbst.[157] Ein Verantwortlicher darf

---

[153] Garante per la protezione dei dati personali, provvedimento n. 232 vom 10.4.2025 gegen Luka Inc.; Pressemitteilung vom 19.5.2025. Festgestellt wurden Verstöße gegen Art. 5 Abs. 1 lit. a und c, Art. 6, 12, 13, 24 und 25 Abs. 1 DSGVO. Ein gesondertes Verfahren zur Rechtmäßigkeit des Modelltrainings wurde ausdrücklich vorbehalten.

[154] Vgl. §§ 40, 44 BDSG.

[155] Vgl. EuGH, Urt. v. 20.12.2017, C-434/16 – Nowak.

[156] Erwägungsgrund 63 DSGVO.

[157] EuGH, Urt. v. 27.2.2025, C-203/22 – Dun & Bradstreet Austria, Rn. 58 ff.

sich aber nicht pauschal auf ein Geschäftsgeheimnis berufen, sondern muss die betreffenden Informationen der Aufsichtsbehörde oder dem Gericht vorlegen, damit dort abgewogen wird.[158]

§ 29 Abs. 1 Satz 2 BDSG ist überdies in der Literatur wegen seiner Unbestimmtheit und wegen der Anforderungen des Art. 23 Abs. 2 DSGVO verbreitet für unionsrechtswidrig gehalten worden.[159] Hinzu kommt, dass sein Wortlaut die überwiegenden berechtigten Interessen eines Dritten als Regelfall nennt und deshalb umstritten ist, ob sich der Verantwortliche auf eigene Geheimhaltungsinteressen stützen kann. Für Verbraucherinnen und Verbraucher bleibt gleichwohl der ungünstige Befund, dass gerade die Auskunft über abgeleitete Merkmale, also über das, was die Informationsasymmetrie auflösen würde, am ehesten unter Berufung auf Betriebsinterna verkürzt wird.

Die weiteren nationalen Ausnahmen tragen demgegenüber nicht. § 34 Abs. 1 Nr. 2 lit. b BDSG lässt die Auskunft nur entfallen, soweit die Daten ausschließlich Zwecken der Datensicherung oder der Datenschutzkontrolle dienen, die Auskunft unverhältnismäßigen Aufwand erfordern würde und eine anderweitige Verwendung durch technische und organisatorische Maßnahmen ausgeschlossen ist. Protokoll- und Verlaufsdaten großer Chatarchive erfüllen bereits das Merkmal der Ausschließlichkeit regelmäßig nicht, weil sie zugleich der Fehleranalyse, der Missbrauchserkennung und der Weiterentwicklung des Dienstes dienen. Die Vorschrift erfasst allenfalls echte Sicherungskopien, auf die ausschließlich zu Wiederherstellungszwecken zugegriffen werden kann. Dass sie in der Praxis gleichwohl gegen Auskunftsverlangen ins Feld geführt wird, ändert daran nichts.

Ebenso wenig hilft § 33 Abs. 1 Nr. 2 lit. a BDSG für den Bestand aus dem Netz zusammengetragener Trainingsdaten weiter. Die Vorschrift setzt voraus, dass die Information die Geltendmachung, Ausübung oder Verteidigung zivilrechtlicher Ansprüche beeinträchtigen würde. Eine solche Beeinträchtigung lässt sich für die Mitteilung, dass Daten einer Person in ein Trainingskorpus eingegangen sind, nicht begründen; denkbar wäre sie allenfalls dort, wo die Information erst die Ansprüche gegen den Anbieter auslöste – und gerade in dieser Konstellation darf die Ausnahme nicht greifen, weil sie sonst gegen ihre eigene Durchsetzung immunisierte. Die Transparenzverkürzung folgt hier deshalb nicht aus dem BDSG, sondern aus Art. 14 Abs. 5 lit. b DSGVO, der die Information bei Unmöglichkeit oder unverhältnismäßigem Aufwand entfallen lässt und für Datenbestände aus ungezielter Massenerhebung angeführt werden könnte.

Auch § 32 Abs. 1 Nr. 4 BDSG greift nicht ein. Die Vorschrift befreit von der Pflicht nach Art. 13 Abs. 3 DSGVO, über eine beabsichtigte Zweckänderung zu unterrichten, verlangt dafür aber, dass die Information die Geltendmachung, Ausübung oder Verteidigung rechtlicher Ansprüche beeinträchtigen würde und die Interessen des Verantwortlichen an der Nichterteilung die Interessen der betroffenen Person überwiegen. Für die nachträgliche Zuführung von Gesprächsverläufen zum Modelltraining ist keine der beiden Voraussetzungen ersichtlich; die Abwägung fiele angesichts des Gewichts der offengelegten Gesprächsinhalte überdies gegen den Anbieter aus. Selbst wenn die Ausnahme einmal eingriffe, bliebe der Verantwortliche nach § 32

---

[158] EuGH, Urt. v. 27.2.2025, C-203/22 – Dun & Bradstreet Austria, Rn. 43 ff., 71 ff.

[159] Zum Beispiel Dix, in: Simitis/Hornung/Spiecker gen. Döhmann, DSGVO/BDSG, § 29 Rn. 9; Wilhelm-Robertson, in: Sydow/Marsch, DS-GVO/BDSG, § 29 Rn. 24.

Abs. 2 BDSG verpflichtet, geeignete Schutzmaßnahmen zu ergreifen und die Informationen der Öffentlichkeit bereitzustellen. Vor allem aber betrifft die Norm allein die Information über die Zweckänderung, nicht deren Zulässigkeit. Ob Gesprächsverläufe zum Training verwendet werden dürfen, richtet sich unverändert nach Art. 6 Abs. 4 DSGVO und bedarf einer eigenen Rechtsgrundlage, die § 32 BDSG nicht liefert.

Eine Sonderrolle spielt das TDDDG. Sein § 25 Abs. 1 verlangt für das Speichern von Informationen auf dem Endgerät und den Zugriff darauf – also für Cookies und vergleichbare Technologien – eine Einwilligung „auf der Grundlage von klaren und umfassenden Informationen“.[160] Diese Norm ist der erste Berührungspunkt zwischen Nutzer:innen und Anbieter.

Der grundsätzliche Schutz, den das Datenschutzrecht bietet, ist in Bezug auf bestimmte individuelle und systemische Risiken jedoch nicht ausreichend.

Anforderungen an die Ermächtigung zur Verarbeitung

Die DSGVO stellt die Verarbeitung hochsensibler und emotional aufgeladener Interaktionsdaten, je nach Kategorie, unter einen allgemeinen Erlaubnisvorbehalt nach Art. 6 Abs. 1 DSGVO sowie einen speziellen Erlaubnisvorbehalt nach Art. 9 DSGVO.

Bei der datenschutzrechtlichen Beurteilung von KI sind folgende Phasen zu unterscheiden: die Erstellung eines Datensatzes durch Erhebung der Daten beim Nutzenden oder aus öffentlich zugänglichen Quellen, das Training der KI und die Nutzung des trainierten Systems.[161]

Für jeden Schritt benötigt der Verantwortliche eine datenschutzrechtliche Rechtsgrundlage, sofern dabei personenbezogene Daten verarbeitet werden. In der Praxis kommen dafür drei in Betracht: die Erfüllung des Nutzungsvertrages, ein berechtigtes Interesse des Anbieters oder die Einwilligung der betroffenen Person. Für die hier relevanten Verarbeitungen – die Bildung psychologischer Profile und die Weiterverwendung der Gesprächsinhalte zum Training der Modelle – bleibt im Ergebnis zumeist nur die Einwilligung übrig. Das liegt auch an der Reichweite des Art. 9 DSGVO und den Anforderungen von Art. 22 DSGVO an das Profiling (siehe dazu unten).

Da Chatbots auf eine kontinuierliche Personalisierung und ggf. emotional geprägte Selbstoffenbarung der Nutzer:innen ausgerichtet sind, werden regelmäßig Angaben zu Gesundheitszustand, psychischer Verfassung, sexueller Orientierung oder politischer Überzeugung erhoben. Dabei handelt es sich um besondere Kategorien personenbezogener Daten im Sinne des Art. 9 Abs. 1 DSGVO. Dies gilt auch, wenn sich entsprechende Informationen erst indirekt aus vorhandenen Daten ableiten lassen.[162] Werden besondere Kategorien personenbezogener Daten verarbeitet, benötigt der Verantwortliche eine Einwilligung nach Art. 9 Abs. 2 lit. a DSGVO, sofern die betroffene Person die Daten nicht im Sinne des Art. 9 Abs. 2 lit. e DSGVO „offensichtlich öffentlich gemacht“ hat. Für Gesprächsinhalte, die Nutzer:innen dem Chatbot mitteilen, dürfte diese Ausnahme regelmäßig nicht einschlägig sein. Die Vorschrift zielt vielmehr auf bestimmte Daten aus öffentlich zugänglichen Quellen ab. In einem der ersten einschlägigen Urteile wurde

---

[160] § 25 Abs. 1 TDDDG; die Information und die Einwilligung haben nach Satz 2 den Maßstäben der DSGVO zu genügen.

[161] LG Hamburg, MMR 2024, 973 Rn. 46. Vgl. auch Schäfer, ZD 2025, 12 mwN.

[162] EuGH, ZD 2022, 611 Rn. 127.

festgestellt, dass die Verarbeitung von Daten besonderer Kategorien auch dann einer Einwilligung bedarf, wenn die nutzende Person die fraglichen Informationen selbst und ohne gezielte Nachfrage des Chatbots diesem gegenüber offengelegt hat.[163] Darüber hinaus hat das Gericht die Unzulässigkeit voreingestellter Ankreuzkästchen festgestellt.

Möchte der Verantwortliche die personenbezogenen Daten zum Training der KI nutzen, kann er sich gegebenenfalls auf ein berechtigtes Interesse nach Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO berufen;[164] bei besonderen Kategorien personenbezogener Daten ist ferner Art. 9 Abs. 2 DSGVO zu beachten, der kein Äquivalent zu Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO enthält. Der Europäische Datenschutzausschuss hat zwar anerkannt, dass ein berechtigtes Interesse in Betracht kommen kann, dies aber an einen dreistufigen Test gebunden: legitimes Interesse, Erforderlichkeit und eine Abwägung, in die insbesondere die vernünftigen Erwartungen der betroffenen Personen einfließen.[165]

Bei der Erhebung besonderer Kategorien personenbezogener Daten bedarf es im Ergebnis einer Einwilligung nach Art. 9 Abs. 2 lit. a DSGVO.[166] Bei der anschließenden Nutzung des trainierten Systems werden entweder eine Einwilligung oder (sofern keine personenbezogenen Daten besonderer Kategorien betroffen sind) das berechtigte Interesse notwendig sein. Sollte der Verantwortliche gegebenenfalls vor dem KI-Training den Datensatz anonymisieren, handelt es sich auch dabei um einen relevanten Verarbeitungsschritt und bedarf eines berechtigten Interesses oder einer Einwilligung.[167]

Informiertheit von Einwilligungen

Eine wirksame Einwilligung muss insbesondere freiwillig und informiert erfolgen.[168]

Die Freiwilligkeit könnte in Zweifel stehen, wenn die emotionale Bindung oder Abhängigkeit der betroffenen Person eine autonome Entscheidung beeinflusst. Zumindest zu Beginn des Nutzungsverhältnisses scheint das jedoch eher von untergeordneter Bedeutung zu sein.

Die Informiertheit verlangt nicht bloß, dass Informationen bereitgestellt werden, sondern dass die betroffene Person tatsächlich versteht, worin sie einwilligt. Im Sinne von Art. 12 Abs. 1 DSGVO müssen die Informationen präzise, transparent, verständlich und in leicht zugänglicher Form, in klarer und einfacher Sprache übermittelt werden. Eine informierte Einwilligung setzt voraus, dass der Verantwortliche seinen Informationspflichten aus Art. 12 ff. DSGVO nachkommt. Informieren muss der Verantwortliche über alle relevanten Umstände der Datenverarbeitung. Dazu gehören unter anderem die Zwecke der Datenverarbeitung, darunter auch die Nutzung zu KI-Trainingszwecken, Empfänger der Daten, das Vorliegen bzw. Erheben von besonderen Kategorien

---

[163] LG Köln, Urt. v. 17.9.2026 – 33 O 120/24. Das Urteil bezog sich auf den in Snapchat enthaltenen generativen Chatbot „My AI“.

[164] So OLG Köln NJW 2025, 3156. Kritisch zu dieser Entscheidung Hornung, KIR 2025, 407. Vgl. auch LfDI BW, Diskussionspapier: Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz, v2.0 vom 17.10.2024, S. 21 f.

[165] EDSA, Stellungnahme 28/2024 zu gewissen Datenschutzaspekten der Verarbeitung personenbezogener Daten im Zusammenhang mit KI-Modellen, angenommen am 17.12.2024, Rn. 68 ff.

[166] Ausführlich zu Rechtsgrundlagen bei Datenverarbeitung zu KI-Trainingszwecken Nebel, KIR 2026, 163.

[167] Zur Anonymisierung von Trainingsdaten ausführlich Nebel, KIR 2026 163 ff. mwN.

[168] Zu den Voraussetzungen der Einwilligung s. zum Beispiel Klement, in: Simitis/Hornung/Spiecker gen. Döhmann, DSGVO/BDSG, Art. 7 Rn. 1 ff.

personenbezogener Daten und die Wahrnehmung von Betroffenenrechten. Sollten für einzelne Verarbeitungsschritte andere Rechtsgrundlagen als die Einwilligung einschlägig sein, ist darüber ebenfalls zu informieren.

Die in Kapitel II beschriebene Informationsasymmetrie ist damit kein Randproblem des Datenschutzrechts. Das Gesetz macht die Rechtmäßigkeit der Verarbeitung von einer informierten Entscheidung abhängig, aber die Systemarchitektur macht diese Entscheidung unmöglich. Dieser Zusammenhang zeigt sich auf den Ebenen Endgerätezugriff, Kontoeinrichtung und Funktionsweise des Modells.

Endgerätezugriff: Bereits der Aufruf der Website oder der App eines Chatbots löst den Anwendungsbereich des § 25 TDDDG aus. Die Speicherung von Informationen in der Endeinrichtung des Endnutzers (das Setzen von Cookies) oder der Zugriff auf Informationen, die bereits in der Endeinrichtung gespeichert sind (u.a. das Auslesen von Cookies), erfordern grundsätzlich eine Einwilligung. Ausnahmen bestehen bei technisch notwendigen Cookies gemäß § 25 Abs. 2 TDDDG. Die Einwilligung wird in der Praxis über Cookie-Banner eingeholt – und damit über ein Instrument, dessen manipulative Gestaltung der Europäische Datenschutzausschuss ausführlich dokumentiert hat. Als manipulativ benannt sind unter anderem hervorgehobene Zustimmungsschaltflächen neben versteckten Ablehnungsmöglichkeiten, mehrstufige Menüs, wertende Formulierungen und wiederholte Nachfragen.[169]

Hinzu kommt ein systematisches Problem. § 25 TDDDG regelt allein den Zugriff auf das Endgerät. Ob die anschließende Verwendung der so erlangten Daten rechtmäßig ist, beurteilt sich gesondert nach der DSGVO. Für Nutzerinnen und Nutzer:innen bedeutet das eine Zweiteilung, die praktisch niemand durchschaut. Ein Klick auf „Alle akzeptieren“ betrifft rechtlich zwei völlig verschiedene Vorgänge mit unterschiedlichen Maßstäben. Wer täglich mehrfach vor Bannern steht, klickt irgendwann zu. Eine aus Ermüdung erteilte Zustimmung ist aber weder informiert noch in einem lebensnahen Sinne freiwillig – auch wenn sie den formalen Anforderungen genügt.

Kontoeinrichtung: Beim Anlegen eines Kontos bei einem Chatbot-Anbieter setzt sich das Muster fort, nun mit deutlich höherem Einsatz. Die Analyse der Datenschutzrichtlinien der sechs größten Anbieter, auf die bereits das Risikokapitel gestützt ist, hat gezeigt, dass wesentliche Angaben zur Datennutzung nicht in der Hauptrichtlinie stehen, sondern über ein Geflecht von Unterrichtlinien und FAQ-Dokumenten verteilt sind – bei einem Anbieter über sechs separate Dokumente hinweg.[170] Gemessen an Art. 12 Abs. 1 DSGVO, der eine „leicht zugängliche Form“ verlangt, ist das kaum haltbar. Die italienische Aufsichtsbehörde hat im Replika-Verfahren genau daran angeknüpft: Eine Datenschutzerklärung, die mögliche Rechtsgrundlagen nebeneinanderstellt, ohne zu sagen, welche für welche Verarbeitung gilt, verstößt gegen die Art. 5 Abs. 1 lit. a, 6, 12 und 13 DSGVO.

Verschärfend wirkt die Architektur der Voreinstellungen. Sämtliche großen Anbieter verwenden Chatinhalte standardmäßig zum Training. Nutzende, die das nicht möchten, müssen widersprechen. Rechtlich ist das aus zwei Gründen angreifbar. Zum einen verlangt Art. 25 Abs. 2

---

[169] EDSA, Leitlinien 03/2022 zu täuschenden Gestaltungsmustern auf den Benutzeroberflächen von Social-Media-Plattformen („Deceptive design patterns“), Version 2.0, angenommen am 14.2.2023. Die Leitlinien behandeln systematisch Registrierung, Datenschutzeinstellungen und Kontolöschung.

[170] Vgl. die Nachweise in Kapitel II, zum Beispiel King u. a., 2025.

DSGVO, dass durch Voreinstellung nur solche Daten verarbeitet werden, die für den jeweiligen Zweck erforderlich sind – und das Training des Modells ist für die Beantwortung der Nutzerfrage gerade nicht erforderlich. Zum anderen berührt die Art der Ansprache die Freiwilligkeit: Wird die Freigabe als sozialer Beitrag gerahmt („Verbessere das Modell für alle“), wird nicht informiert, sondern appelliert. Der Europäische Datenschutzausschuss hat bereits im Zusammenhang mit ChatGPT betont, dass Transparenz über die Weiterverwendung von Eingaben ein wesentlicher Abwägungsfaktor ist und der Anbieter die Verantwortung nicht auf die Nutzer:innen abwälzen darf.[171]

Besonders deutlich wird die Lücke bei Minderjährigen. Art. 8 DSGVO erklärt die Einwilligung eines Kindes bei Diensten der Informationsgesellschaft erst ab 16 Jahren für wirksam; Deutschland hat von der Möglichkeit, diese Altersgrenze abzusenken, keinen Gebrauch gemacht. Darunter bedarf es der Zustimmung der Sorgeberechtigten, und der Anbieter muss sich unter Berücksichtigung der verfügbaren Technik „angemessene Anstrengungen“ um deren Überprüfung bemühen (Art. 8 Abs. 2 DSGVO). Wo, wie im Replika-Fall festgestellt, überhaupt keine Altersverifikation stattfindet, sind sämtliche Einwilligungen jüngerer Nutzer:innen unwirksam – und damit die gesamte darauf gestützte Verarbeitung rechtswidrig. Die Norm ist scharf, sie wird aber zumeist nicht konsequent befolgt.[172]

Besonders problematisch ist auch: Nutzerinnen und Nutzer, die den Chatbot – naheliegenderweise – direkt nach seiner Datenpraxis fragen, erhalten systematisch falsche Auskünfte, die im Widerspruch zur eigenen Datenschutzrichtlinie des Anbieters stehen.[173] Datenschutzrechtlich handelt es sich hierbei um Verstöße gegen die Transparenzpflicht der Art. 12 f. DSGVO und gegen den Grundsatz der Verarbeitung nach Treu und Glauben, Art. 5 Abs. 1 lit. a DSGVO. Verbraucherpolitisch ist es gravierender. Denn die konversationelle Oberfläche, die das Verstehen erleichtern sollte, wird hier zum Verstärker der Asymmetrie: Ausgerechnet die Instanz, der im intimen Gesprächskontext das größte Vertrauen entgegengebracht wird, desinformiert.

Funktionsweise des Modells: Die tiefste Schicht der Asymmetrie betrifft nicht das, was Anbieter verschweigen, sondern das, was auch sie nicht vollständig wissen. Was geschieht mit einem Satz, der in ein Sprachmodell einfließt? Bleibt er dort erhalten? Kann er wieder herauskommen? Diese Fragen sind unter Aufsichtsbehörden und in der Wissenschaft ungeklärt.

Die Hamburgische Datenschutzbehörde vertritt die These, dass ein Sprachmodell selbst keine personenbezogenen Daten speichere, weil Inhalte durch die Zerlegung in Tokens ihre Charakteristik verlören; Betroffenenrechte könnten sich deshalb nur auf Eingaben und Ausgaben beziehen, nicht auf das Modell.[174] Andere Aufsichtsbehörden widersprechen; das Unabhängige Landeszentrum für Datenschutz Schleswig-Holstein hält einen Personenbezug für nicht pauschal ausschließbar, weil Informationen aus den Trainingsdaten im Modell verbleiben und extrahiert werden können. Auch

---

[171] EDSA, Report of the work undertaken by the ChatGPT Taskforce, 23.5.2024, Rn. 18 ff.

[172] Ähnlich der Nutzeridentifikation bei anderen Online-Diensten, siehe dazu EuGH, Urt. v. 2.12.2025, – C-492/23 (Russmedia) und Johannes, MMR 2026, 307.

[173] Vgl. die Nachweise im Risikokapitel (Ragab/Mannan/Youssef 2024); dort wiesen sämtliche auswertbaren Apps Widersprüche zwischen Bot-Aussage und Richtlinie auf.

[174] Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit, Diskussionspapier „Large Language Models und personenbezogene Daten“ vom 15.7.2024, Thesen 1 bis 3.

die Forschung hat auf die rechtlichen Folgen der Memorisierung von Trainingsdaten hingewiesen.[175] Der Europäische Datenschutzausschuss hat sich einer der beiden Positionen nicht angeschlossen, sondern eine Einzelfallprüfung angeordnet: Anonymität sei nur anzunehmen, wenn sowohl die Wahrscheinlichkeit einer direkten Extraktion als auch die Wahrscheinlichkeit, personenbezogene Daten durch Abfragen zu erlangen, unerheblich sei.[176]

Wenn Aufsichtsbehörden und Wissenschaft seit Jahren darüber streiten, ob ein Modell personenbezogene Daten enthält, dann kann die Antwort auf diese Frage nicht Gegenstand einer Verbraucherentscheidung im Registrierungsdialog sein. Die Einwilligung soll den Betroffenen in die Lage versetzen, über die Verwendung seiner Daten selbst zu bestimmen. Sie verlangt damit ein Wissen über Tragweite und Folgen, das objektiv nicht verfügbar ist.

Das schlägt unmittelbar auf die Informationspflichten durch. Art. 13 Abs. 2 lit. f und Art. 15 Abs. 1 lit. h DSGVO verlangen bei automatisierter Entscheidungsfindung (sofern eine solche durch den Chatbot erfolgt; siehe hierzu das nachfolgende Kapitel) „aussagekräftige Informationen über die involvierte Logik“. Wo die Logik in Milliarden von Parametern besteht, deren Zusammenwirken auch der Anbieter nicht im Einzelnen erklären kann, läuft die Vorschrift ins Leere – nicht, weil sie falsch formuliert wäre, sondern weil sie einen Erklärungszusammenhang voraussetzt, den die Technologie nicht bereitstellt. Nutzer:innen willigen damit in etwas ein, dessen Reichweite niemand seriös beschreiben kann. Die Informiertheit wird zur Fiktion, und mit ihr die Rechtsgrundlage, die auf ihr aufbaut.

Gefahren von Profiling

Profiling wird in Art. 4 Nr. 4 DSGVO definiert als „jede Art der automatisierten Verarbeitung personenbezogener Daten, die darin besteht, dass diese personenbezogenen Daten verwendet werden, um bestimmte persönliche Aspekte, die sich auf eine natürliche Person beziehen, zu bewerten, insbesondere um Aspekte bezüglich Arbeitsleistung, wirtschaftliche Lage, Gesundheit, persönliche Vorlieben, Interessen, Zuverlässigkeit, Verhalten, Aufenthaltsort oder Ortswechsel dieser natürlichen Person zu analysieren oder vorherzusagen“. Im Rahmen der Interaktion mit KI-Companions stellt praktisch jeder Output ein Profiling des Nutzenden als betroffener Person dar, sofern deren Input sowie ggf. weitere Informationen durch den Betreiber des KI-Systems als Verantwortlichen dazu genutzt werden, Aspekte wie persönliche Vorlieben und Interessen zu ermitteln, um einen adäquaten Output als Gesprächsinhalt zu generieren.

Art. 22 DSGVO beinhaltet das Recht betroffener Personen, nicht einer automatisierten Entscheidung – einschließlich Profiling – unterworfen zu werden. Die Norm ist auf punktuelle Entscheidungen zugeschnitten (Kreditvergabe, Bewerbung, Scoring), die „rechtliche Wirkung“ oder „ähnlich erhebliche Beeinträchtigung“ entfalten. Chatbot-Interaktionen sind aber in aller Regel keine Entscheidungen, sondern eine fortlaufende, sich über einen Zeitraum entwickelnde Kommunikation. Zwar können entsprechende Interaktionen (psychologische oder emotionale)

[175] Nolte/Finck/Meding, Machine Learners Should Acknowledge the Legal Implications of Large Language Models as Personal Data, 2025.

[176] EDSA, Stellungnahme 28/2024, Rn. 43.

Auswirkungen auf die betroffene Person haben,[177] aber in aller Regel keine im Sinne des Art. 22 DSGVO relevanten.

Damit ist allerdings nur gesagt, dass das Gespräch selbst keine Entscheidung im Sinne der Vorschrift ist. Art. 22 DSGVO ist deshalb in Bezug auf GPC nicht gegenstandslos, sondern verlagert sich auf diejenigen automatisierten Festlegungen, die die Interaktion umgeben. Bei einem GPC (oder Companion) könnten dies zum Beispiel sein:

- die Einstufung des Alters und die daran anknüpfende Beschränkung des Funktionsumfangs,
- die automatisierte Erkennung von Krisensituationen mit anschließender Sperrung bestimmter Ausgaben oder Weiterleitung an Dritte,
- die Sperrung oder stille Beschränkung eines Kontos nach automatisierter Verhaltensbewertung sowie
- eine nach ermittelter Zahlungsbereitschaft differenzierende Preis- und Konditionengestaltung.

Diese Vorgänge weisen die von Art. 22 DSGVO vorausgesetzten Tatbestandsmerkmale auf. Sie beruhen ausschließlich auf automatisierter Verarbeitung, sie führen zu einer abgrenzbaren Festlegung, und sie wirken nach außen. Der Entscheidungsbegriff ist weit auszulegen und bereits die Ermittlung eines Wahrscheinlichkeitswertes als Entscheidung eingeordnet, wenn dieser das weitere Vorgehen maßgeblich bestimmt.[178] Hinzu kommt, dass die Vorschrift nach der Rechtsprechung ein Verbot enthält und nicht lediglich ein Recht, das die betroffene Person geltend machen muss.[179]

Ob die Schwelle der ähnlich erheblichen Beeinträchtigung erreicht ist, hängt vom Gewicht der Festlegung ab. Für den Ausschluss von einem Dienst oder die Beschränkung seines Funktionsumfangs dürfte sie regelmäßig überschritten sein. Für die Ausspielung von Werbung gilt das im Ausgangspunkt nicht. Nach den von der Aufsicht gebilligten Leitlinien kann jedoch auch sie nur die Schwelle erreichen, wenn das Profiling besonders eingriffsintensiv ist, an bekannte Schwächen anknüpft oder sich gezielt an schutzbedürftige Gruppen richtet.[180] Genau diese drei Merkmale liegen bei einem System vor, das seine Empfehlungen auf einen aus vertraulichen Gesprächen abgeleiteten Gefühlszustand stützt. Die im Risikokapitel beschriebene Ausspielung von Werbung an Nutzende in emotionalen Notlagen ist danach ein Anwendungsfall der Vorschrift.

Bei den individuellen Risiken entfaltet Art. 22 DSGVO daher eine begrenzte, aber reale Wirkung, insbesondere bei Minderjährigen und bei Personen in Krisensituationen, weil die sie betreffenden Festlegungen die vorbenannte Schwelle am ehesten erreichen. Bei den systemischen Risiken

---

[177] Teo/Porsdam Mann/Jurcys, Forthcoming in the Law, Innnovation and Technology Journal 2027, https://ssrn.com/abstract=5778243 sprechen von „relational character of companion AI harms“, jeder Nutzende sei also „relational vulnerabel“; Frei/Sparzynski/Rostam,Wenn künst-liche Liebe gefähr-lich wird, LTO vom 27.4.2026, https://www.lto.de/recht/hintergruende/h/regulierung-ai-companions-kuenstliche-liebe-regulierung. Ähnlich Boine, Emotional Attachment to AI Companions and European Law, 2023, S. 13 ff.

[178] EuGH, Urt. v. 7.12.2023 – C-634/21 – SCHUFA Holding (Scoring), Rn. 42 ff.

[179] EuGH, Urt. v. 7.12.2023 – C-634/21 – SCHUFA Holding (Scoring), Rn. 52.

[180] Artikel-29-Datenschutzgruppe, Leitlinien zu automatisierten Entscheidungen im Einzelfall einschließlich Profiling, WP 251 rev.01, gebilligt durch den EDSA.

versagt sie dagegen weitgehend. Die Vorschrift schützt davor, einer Entscheidung unterworfen zu werden, nicht davor, durch einen Prozess geformt zu werden.

Praktisch entscheidend ist schließlich die Frage, was von dem Schutz von Art. 22 DSGVO nach einer Einwilligung übrig bleibt. Eine ausdrückliche Einwilligung lässt das Verbot des Abs. 1 entfallen, und sie dürfte bei der Registrierung bei GPC und Companions regelmäßig eingeholt werden. Das führt jedoch nicht dazu, dass die Vorschrift leerliefe. Sie sieht dann in Art. 22 Abs. 3 DSGVO einen Verfahrensschutz vor. Der Verantwortliche hat angemessene Maßnahmen zu treffen, um die Rechte der betroffenen Person zu wahren, wozu mindestens das Recht auf Erwirkung des Eingreifens einer Person, auf Darlegung des eigenen Standpunkts und auf Anfechtung der Entscheidung gehört. Dies bedeutet zum Beispiel, dass wer nach einer automatisierten Alterseinstufung vom Dienst ausgeschlossen oder in seinem Funktionsumfang beschränkt wird, danach einen Anspruch darauf hat, dass ein Mensch den Fall prüft.

Verschärfend tritt die Einschränkung für besondere Datenkategorien hinzu. Beruht die Entscheidung auf Gesundheitsdaten, Angaben zur sexuellen Orientierung oder zu religiösen Überzeugungen, ist sie nur bei ausdrücklicher Einwilligung und nur bei angemessenen Schutzmaßnahmen zulässig.[181] Da Chatbot-Gespräche solche Daten in hoher Dichte enthalten und die automatisierte Erkennung einer Krisensituation ihrem Gegenstand nach eine Gesundheitseinstufung ist, greift diese Vorschrift im Regelfall. Die Anforderungen an die Einwilligung sind dann doppelt erhöht. Sie muss ausdrücklich, für den bestimmten Fall und in informierter Weise erteilt sein. Eine in den Nutzungsbedingungen eingebettete pauschale Zustimmung genügt dem in der Regel nicht. Die Einwilligung, die den Schutz entfallen lässt, ist ihrerseits wegen der Informationsasymmetrie kaum wirksam zu erteilen (siehe dazu auch Kapitel V.4).

Gewährleistung von Speicherbegrenzung

Die vorgenannte Transparenzproblematik wird flankiert durch die Problematik der Gewährleistung der Speicherbegrenzung. Art. 5 Abs. 1 lit. e DSGVO verlangt, dass personenbezogene Daten nur so lange in identifizierbarer Form gespeichert werden, wie es für die Zwecke erforderlich ist. Dieser Grundsatz gilt objektiv. Er hängt weder von einer Einwilligung noch davon ab, ob die betroffene Person die Technik verstanden hat. Der Gerichtshof hat ihn im Verfahren Schrems gegen Meta bekräftigt: Personenbezogene Daten dürfen nicht zeitlich unbegrenzt und ohne Unterscheidung nach ihrer Art analysiert und verarbeitet werden; die Speicherdauer ist auf das absolut Notwendige zu begrenzen.[182] Was für Werbeprofile gilt, muss für Chatbotgesprächsverläufe, die psychologische Profile aus intimen Gesprächen darstellen, erst recht gelten.

Fraglich ist, ob der Anbieter darlegen kann, wie lange er welche Daten wofür benötigt – und ob er sie anschließend tatsächlich entfernt. Die Speicherung früherer Gesprächsinhalte während des laufenden Nutzungsverhältnisses kann gerade der von Nutzenden erwarteten Personalisierung und Gesprächskontinuität dienen. Eine langanhaltende Speicherung kann daher grundsätzlich erforderlich sein, möglicherweise auch für die gesamte Dauer des Kontobestands. Entscheidend ist

[181] Art. 22 Abs. 4 DSGVO.

[182] EuGH, Urt. v. 04.10.2024 - C-446/21, Rn. 49 f., 62 f. Der Gerichtshof qualifiziert die Erstellung detaillierter Profile in Verbindung mit einer zeitlich unbegrenzten Weiterverarbeitung als schweren Grundrechtseingriff.

aber, welche Inhalte für welchen Zweck tatsächlich benötigt werden (etwa Gesprächskontinuität vs. Modelltraining) und ob die Speicherung auf diesen Zweck begrenzt bleibt. Nach Löschung des Kontos dürfte diese Rechtfertigung jedenfalls regelmäßig entfallen.

Genau hier zeigt sich die technische Lücke. Wird ein Löschungsverlangen nach Art. 17 DSGVO auf ein trainiertes Modell erstreckt, existiert derzeit kein Verfahren, mit dem sich ein einzelner Trainingsbeitrag zuverlässig entfernen ließe; das sogenannte machine unlearning ist Gegenstand der Forschung, nicht der Praxis.

Enthält das Modell personenbezogene Daten, besteht ein Löschungsanspruch, der sich praktisch nicht erfüllen lässt; die Daten bleiben in identifizierbarer Form auf unbestimmte Zeit erhalten, was mit dem Grundsatz der Speicherbegrenzung aus Art. 5 Abs. 1 lit. e DSGVO kaum vereinbar ist. Enthält es sie nicht – so die Hamburger These –, greifen weder die Betroffenenrechte am Modell noch der Grundsatz der Speicherbegrenzung. Das Problem verschwindet damit aber nicht, es verlagert sich: Der Einfluss der Trainingsdaten auf das Modell ist irreversibel, und genau diese Irreversibilität führt der Trainingsvorgang selbst herbei, solange er noch personenbezogene Daten verarbeitet. Für die betroffene Person läuft beides auf dasselbe hinaus, nämlich eine faktisch unbefristete Fortwirkung ihrer Daten, auf die sie keinen Zugriff mehr hat. Nicht die Einwilligung ist daher das Problem, sondern die Frage, ob ein Verarbeitungsvorgang, der auf eine solche Unumkehrbarkeit angelegt ist, mit den Grundsätzen der Verordnung vereinbar sein kann.

3.3 Zwischenfazit

Das geltende Datenschutzrecht wirkt hinsichtlich der systemischen Risiken „Datensicherheit und Profilbildung“ (S3) und „Informationsasymmetrie“ (S1) grundsätzlich abmildernd. Es gilt dabei auch für Anbieter außerhalb der Europäischen Union, sofern sie Personen in der Union Dienste anbieten.[183] Insbesondere die DSGVO wird durch die KI-Verordnung nicht verdrängt, sondern gilt neben ihr fort.[184] Nahezu jeder Betreiber eines für deutsche Verbraucherinnen und Verbraucher zugänglichen Chatbots unterliegt damit dem europäischen und deutschen Datenschutzrecht.

Das Datenschutzrecht ist damit in Bezug auf Chatbots kein wirkungsloser Rahmen. Es erfasst die Erhebung, die Profilbildung, die Weiterverwendung zum Training und die Speicherung – und damit einen erheblichen Teil der systemischen Risiken. Seine Grundsätze greifen unabhängig vom Nachweis eines individuellen Schadens, was es gerade für die im Risikokapitel beschriebene Kausalitätsproblematik geeignet macht. Der Replika-Fall zeigt, dass die vorhandenen Instrumente durchaus greifen, wenn sie angewandt werden.

Zugleich bestehen zwei strukturelle Lücken. Die erste ist die Zirkularität der Einwilligung: Das Recht macht die Rechtmäßigkeit der Verarbeitung von einer informierten Entscheidung abhängig, während dieselbe Informationsasymmetrie, die geschützt werden soll, deren Informiertheit verhindert. Die zweite betrifft die Speicherbegrenzung: Das Recht verbietet eine ungerechtfertigte unbefristete Speicherung, während die Technologie sie faktisch herstellt und ein Rückgängigmachen nicht ermöglicht.

---

[183] Art. 3 Abs. 2 lit. a DSGVO (Marktortprinzip).

[184] Art. 2 Abs. 7 der Verordnung (EU) 2024/1689 (KI-VO): Das Unionsrecht zum Schutz personenbezogener Daten bleibt unberührt.

Die entscheidende Grenze liegt jedoch woanders, und sie betrifft das zentrale Bewertungskriterium dieser Untersuchung. Das Datenschutzrecht regelt den Umgang mit Daten, nicht die Einwirkung auf Menschen. Es fragt, ob eine Verarbeitung erlaubt ist, nicht, ob eine Gesprächsführung fair war. Damit bleiben genau jene Praktiken außerhalb seines Zugriffs, die im Risikokapitel als besonders problematisch beschrieben werden (u.a. die schrittweise Steigerung emotionaler Bindung, die sykophantische Bestätigung bestehender Überzeugungen, der Schuldappell im Moment der Verabschiedung). Keine jener Praktiken lässt sich als datenschutzrechtlicher Verstoß fassen, solange die zugrunde liegende Datenverarbeitung ihrerseits zulässig ist.

Dagegen schützt auch Art. 22 DSGVO nicht. Die Vorschrift schützt vor Entscheidungen, die ausschließlich automatisiert ergehen und rechtliche Wirkung entfalten oder in ähnlicher Weise erheblich beeinträchtigen. Ein Chatbot trifft (vorbehaltlich des oben Gesagten) keine Entscheidung über den Nutzer. Er führt ein Gespräch mit ihm. Dass dieses Gespräch die Entscheidungen des Nutzers beeinflusst, ist gerade der Punkt – aber es ist nicht der Tatbestand der Norm.

4. Plattformregulierung

4.1 Allgemeines

Die europäische Plattformregulierung zu digitalen Diensten fußt im Wesentlichen auf dem Digital Services Act (DSA)[185] und dem Digital Markets Act (DMA).[186] In Deutschland sind dazu konkretisierende Regelungen im Digitale-Dienste-Gesetz (DDG)[187] festgelegt.

Digital Services Act

Der DSA regelt im Kern die Haftungsprivilegierung für vom Nutzenden eingegebene Informationen (Art. 4 ff. DSA) und den Umfang von Sorgfaltspflichten (Art. 11 ff. DSA) für Anbieter digitaler Vermittlungsdienste.[188] Es ist schon fraglich, ob und inwieweit diese Regelungen für Chatbots gelten.

Außerdem enthält der DSA spezielle Sorgfaltspflichten für Betreiber von Online-Plattformen und zusätzliche Pflichten für Betreiber sehr großer Online-Plattformen und sehr großer Suchmaschinen. Auch hier ist schon fraglich, inwieweit der Anwendungsbereich für Chatbots und all ihre Funktionen eröffnet ist (siehe dazu Kapitel 4.2.2 und 4.2.3).

Digital Markets Act

Der DMA reguliert marktmächtige, von der EU-Kommission als „Torwächter“ benannte Anbieter zentraler digitaler Plattformdienste, indem er ihnen bestimmte wettbewerbsbeschränkende

[185] Verordnung (EU) 2022/2065 des Europäischen Parlaments und des Rates vom 19. Oktober 2022 über einen Binnenmarkt für digitale Dienste und zur Änderung der Richtlinie 2000/31/EG (Gesetz über digitale Dienste).

[186] Verordnung (EU) 2022/1925 des Europäischen Parlaments und des Rates vom 14. September 2022über bestreitbare und faire Märkte im digitalen Sektor und zur Änderung der Richtlinien (EU) 2019/1937 und (EU) 2020/1828 (Gesetz über digitale Märkte).

[187] Digitale-Dienste-Gesetz vom 6. Mai 2024 (BGBl. 2024 I Nr. 149), das zuletzt durch Artikel 1 des Gesetzes vom 12. Mai 2026 (BGBl. 2026 I Nr. 138) geändert worden ist.

[188] Ausführlich Erenli, in: Geminn/Johannes (Hrsg.), Europäisches Datenrecht, § 16.

Praktiken verbietet und ihnen Pflichten wie Interoperabilität und Datenportabilität auferlegt, um die Kontestierbarkeit digitaler Märkte zu sichern.[189]

Der DMA ist anwendbar, wenn ein zentraler Plattformdienst im Sinne des Art. 2 Nr. 2 DMA vorliegt (zum Beispiel Suchmaschinen, App-Stores, Betriebssysteme oder soziale Netzwerke) und das Unternehmen die Voraussetzungen eines „Torwächters“ erfüllt und durch die Kommission als solcher benannt wurde (Art. 2 Nr. 1 iVm Art. 3 DMA). Der DMA kennt auch eine Kategorie „virtuelle Assistenten“ gemäß Art. 2 Nr. 2 lit. h, Nr. 12 DMA. Diese müssen Aufträge, Aufgaben oder Fragen verarbeiten und auf dieser Grundlage Zugang zu anderen Diensten ermöglichen (Art. 2 Nr. 12 DMA). General Purpose Chatbots lassen sich nicht per se in diese abschließende und sehr kontextspezifische Aufzählung einfügen. Bestehende Definitionen sollten auf „generative KI“ oder „General Purpose Chatbots“ erweitert oder als eigener Anwendungsbereich aufgenommen werden. Aber selbst wenn ein General Purpose Chatbot die Voraussetzungen eines „virtuellen Assistenten“ im Einzelfall erfüllen würde, müsste der Chatbot-Anbieter zusätzlich noch die Torwächtereigenschaften des Art. 2 Nr. 1 DMA erfüllen und als solcher benannt werden.[190]

Der DMA adressiert spezifische Fragen aus dem Bereich des Wettbewerbsrechts. Privacy- oder Verbraucherschutzfragen sowie systemische Risiken, die mit General Purpose Chatbots einhergehen, etwa emotionale Abhängigkeit oder Manipulation, suchtförderndes Design, Kindeswohlrisiken und ähnliches, sind nur unzureichend erfasst. Zwar beinhaltet Art. 13 DMA ein Umgehungsverbot und soll verhindern, dass Torwächter die Pflichten aus Art. 5, 6 und 7 DMA (Interoperabilität, Datenzugang, Selbstbevorzugung, Kombinationsverbote) vertraglich, kommerziell, technisch oder durch Dark Patterns umgehen. Umfasst sind aber Manipulationen nur insoweit, als sie den Pflichten des DMA zuwiderlaufen. Es handelt sich nicht um eine Generalklausel gegen manipulatives Design. Aspekte wie emotionale Abhängigkeit, allgemein suchtförderndes Design oder Kindeswohlrisiken sind nicht vom Anwendungsbereich des Art. 13 DMA umfasst, sondern werden vor allem durch Art. 25 und 28 DSA adressiert.

Digitale-Dienste-Gesetz

Das DDG schafft in Deutschland einen einheitlichen Rechtsrahmen für alle Online-Dienste. Es nimmt Digitale Dienste (zum Beispiel Plattformen, Suchmaschinen und Webseiten-Betreiber) in die Pflicht, regelt Transparenz- und Impressumspflichten und schränkt personalisierte Werbung durch Nutzer-Profiling ein. Als zentrale Aufsichtsbehörde koordiniert die Bundesnetzagentur die Einhaltung dieser Regeln.

Nach § 1 Abs. 1 Satz 1 DDG gilt das Gesetz für alle Diensteanbieter, wobei § 1 Abs. 4 Nr. 5 DDG den Diensteanbieter als Anbieter digitaler Dienste bestimmt und § 1 Abs. 4 Nr. 1 DDG den digitalen Dienst mit dem Dienst im Sinne des Art. 1 Abs. 1 lit. b der Richtlinie (EU) 2015/1535 gleichsetzt. Maßgeblich sind damit vier Merkmale, nämlich die Erbringung im Fernabsatz, die elektronische Erbringung, der individuelle Abruf durch den Empfänger und die Erbringung in der Regel gegen Entgelt.

---

[189] Ausführlich Müller, in: Geminn/Johannes (Hrsg.), Europäisches Datenrecht, § 12.

[190] Bis zum Zeitpunkt der Gutachtenerstellung (Herbst 2026) war das jedenfalls nicht der Fall.

Ein über das Internet bereitgestellter Chatbot erfüllt diese Merkmale. Der Nutzer löst die Leistung durch seine Eingabe einzeln aus, die Verarbeitung und Übermittlung erfolgt vollständig elektronisch, und die Beteiligten sind nicht gleichzeitig körperlich anwesend. Auf die Entgeltlichkeit im Einzelfall kommt es nicht an, weil das Merkmal gattungsbezogen zu verstehen ist.[191] Daher sind Werbe- und datenfinanzierte Chatbots sowie unentgeltliche Zugänge innerhalb eines Freemium-Modells auch digitale Dienste im Sinne der Vorschrift.

Nach § 3 Abs. 1 DDG unterliegen in Deutschland niedergelassene Diensteanbieter den Anforderungen des deutschen Rechts auch dann, wenn sie ihre Dienste geschäftsmäßig in anderen Mitgliedstaaten anbieten, soweit nicht der DSA gilt. Umgekehrt wird nach § 3 Abs. 2 DDG der freie Verkehr solcher Dienste nicht eingeschränkt, die in Deutschland von Anbietern mit Niederlassung in einem anderen Mitgliedstaat angeboten werden. Ein in Dublin oder Amsterdam niedergelassener Anbieter, der seinen Chatbot in deutscher Sprache und mit ausdrücklicher Ausrichtung auf den deutschen Markt bereitstellt, unterliegt daher grundsätzlich nicht den Anforderungen des DDG, sondern dem Recht seines Sitzlandes. Sprache und Marktausrichtung sind für die Anwendbarkeit des DDG ohne Belang.[192] Eine Anwendung deutschen Rechts auf den EU-ausländischen Anbieter kommt nur unter den engen Voraussetzungen des § 3 Abs. 5 DDG in Betracht, der u.a. die Einhaltung des Konsultationsverfahrens nach Art. 3 Abs. 4 lit. b und Abs. 5 der Richtlinie 2000/31/EG voraussetzt. Unberührt bleiben ferner die Bereichsausnahmen des § 3 Abs. 4 DDG, insbesondere für das Urheberrecht und die gewerblichen Schutzrechte sowie für die Zulässigkeit nicht angeforderter kommerzieller Kommunikation durch elektronische Post.

Für Anbieter mit Sitz in einem Drittstaat trägt das Herkunftslandprinzip nicht. Die §§ 2 und 3 DDG koordinieren allein das Verhältnis der Mitgliedstaaten untereinander und setzen eine Niederlassung im Binnenmarkt voraus. Ein Chatbot-Anbieter ohne Niederlassung in der Union kann sich deshalb weder auf die Freiverkehrsgarantie des § 3 Abs. 2 DDG noch auf das Verfahrenserfordernis des § 3 Abs. 5 DDG berufen. Die Anwendbarkeit des Gesetzes richtet sich für ihn nach dem allgemeinen Marktbezug, also danach, ob der Dienst geschäftsmäßig auf den inländischen Markt ausgerichtet ist. Deutschsprachige Oberfläche, deutschsprachige Vertragsbedingungen und eine auf Deutschland zielende Bewerbung bedeuten also, dass das DDG auch von Chatbot-Betreibern eingehalten werden muss, die in einem Drittland ihren Sitz haben.

4.2 Risikenbewertung

Klassifizierung als Vermittlungsdienst nach DSA

Die Kernprobleme des DSA in Bezug auf General Purpose Chatbots liegen vor allem in der Frage, ob und inwiefern der sachliche Anwendungsbereich eröffnet ist. Vermittlungsdienste umfassen „ein breites Spektrum an wirtschaftlichen Tätigkeiten, die online stattfinden und sich kontinuierlich weiterentwickeln, um eine rasche, sichere und geschützte Übermittlung von Informationen zu

[191] In der Sache EuGH, Urt. v. 11.9.2014 - C-291/13, Papasavvas hat der EUGH dies für ein werbefinanziertes und für den Nutzer kostenfreies Angebot ausdrücklich entschieden.

[192] Dieses Herkunftslandprinzip kann nur im Einzelfall unter den Voraussetzungen des § 3 Abs. 5-7 DDG durchbrochen werden, s. zB Gerdemann, in: Heldt/Legner, DDG, § 3 Rn. 6; Liesching, in: BeckOK JugendschutzR, DDG, § 3 Rn. 17.

ermöglichen und allen Beteiligten des Online-Ökosystems komfortable Lösungen zu bieten“.[193] Allerdings ist fraglich, in welchem Umfang KI-Chatbots als Vermittlungsdienste klassifiziert werden können. Vermittlungsdienste sind gemäß Art. 3 lit. g Nr. i-iii DSA Dienste der Informationsgesellschaft, die vom Nutzenden bereitgestellte Informationen als reine Durchleitungsdienste, Caching-Dienste oder Hosting-Dienste verarbeiten. Für KI-Chatbots naheliegend ist die Einordnung als Hosting-Dienst im Sinne des Art. 3 lit. g Nr. iii DSA. Dies setzt voraus, dass der Chatbot ein Dienst der Informationsgesellschaft ist, der vom Nutzenden bereitgestellte Informationen in dessen Auftrag speichert. Der Hosting-Dienst fungiert also als Intermediär fremder Inhalte, indem er solche Informationen speichert, die der Nutzende einbringt und über die der Anbieter weder Kenntnis noch Kontrolle besitzt.[194] Abzugrenzen davon sind eigene Inhalte des Diensteanbieters, die durch die aktive Rolle desselben entstanden sind, klassischerweise eigene Informationsseiten, Blogbeiträge und sonstige Inhalte des Diensteanbieters. Hierzu ist auch KI-generierter Content durch ein anbieterseitiges KI-System zu zählen, da der Anbieter eine aktive Rolle bei der Entscheidung zum Einsatz und der Auswahl des Systems übernommen hat.[195] Fallen einige Dienste des Anbieters unter die Definition des Hosting-Dienstes und andere nicht, so ist die Hosting-Eigenschaft für jeden Dienst separat zu bestimmen.[196] Ist das nicht möglich, weil die einzelnen Funktionalitäten integraler Bestandteil des Dienstes sind, die sich nicht voneinander trennen lassen, liegt ein hybrider Dienst vor. Dieser könnte als Vermittlungsdienst sui generis eingestuft werden. Das setzt voraus, dass Art. 3 lit. g DSA die Arten von Vermittlungsdiensten nicht abschließend nennt. Dafür spricht vor allem die Nennung von Online-Suchmaschinen als Vermittlungsdienst außerhalb von Art. 3 lit. g DSA in Art. 3 lit. j DSA.[197]

General Purpose Chatbots dienen einer Vielzahl von Einsatzzwecken. Sie können zur Websuche verwendet werden, aber mit oder ohne entsprechenden Prompt und mit oder ohne Websuche als KI-Companion eingesetzt werden. Sie könnten daher als hybride Vermittlungsdienste angesehen werden, die sowohl Hosting-Elemente umfassen (Verarbeitung von vom Nutzenden bereitgestellter Inhalte, einschließlich der Prompts an sich), aber darüber hinaus eigene Inhalte generieren, die nicht vom Hosting-Dienst umfasst sind.

Am 31. August 2026 hat die Kommission ChatGPT als sehr große Online-Suchmaschine benannt – als ersten KI-Chatbot überhaupt und neben Google Search und Bing als erst dritten Dienst dieser Kategorie.[198] Tragend war die Einordnung als hybrider Dienst. Das System beantwortet Anfragen dialogisch und durchsucht dabei zugleich das Netz. Damit ist entschieden, was zuvor offen war –

---

[193] Erwägungsgrund 29 Satz 1 DSA.

[194] Dies geht zurück auf die Rechtsprechung des EuGH zur Vorgängernorm des Art. 14 der eCommerce-Richtlinie 2000/31/EG. Bezugnahmen auf Art. 14 gelten seit Inkrafttreten des DSA als Bezugnahme auf Art. 6 DSA, der Vorschrift zur Haftungsprivilegierung von Hosting-Diensten. S. zum Beispiel EuGH, GRUR Int 2010, 385; EuGH, GRUR 2011, 1025; EuGH, GRUR 2021, 1054; s. auch Kraul, in: BeckOK UWG, DSA Art. 3 Rn. 19 ff.

[195] Vgl. zur aktiven Rolle EuGH, GRUR 2011, 1025, Rn. 113, 123 sowie GRUR 2021, 1054, Rn. 106.

[196] Erwägungsgrund 15 DSA.

[197] Mit weiteren Nachweisen Busch/Wemjes, RDi 2026, 366. Argumente dafür und dagegen darstellend Hofmann, in: Hofmann/Raue, DSA, Art. 3 Rn. 63 sowie Rn. 64 ff.

[198] Europäische Kommission, Pressemitteilung vom 31.8.2026 (Benennung von ChatGPT als VLOSE sowie von Reddit und Roblox als VLOP). Der Anbieter hatte für den Sechsmonatszeitraum bis zum 31.3.2026 rund 159,1 Millionen durchschnittlich monatlich aktive Nutzer der Suchfunktionen in der Union gemeldet; der Schwellenwert liegt bei 45 Millionen. Die zusätzlichen Pflichten sind binnen vier Monaten zu erfüllen. Die förmliche Entscheidung war bei Abfassung dieses Kapitels noch nicht veröffentlicht.

das Gesprächsformat steht der Einordnung als Suchmaschine nicht entgegen. Maßgeblich ist die Funktion, nicht die Produktgattung.

Der Digital Services Act knüpft an Dienste an, nicht an Unternehmen. Ein Anbieter kann deshalb mehrere Rollen gleichzeitig innehaben, und die Pflichten treffen ihn jeweils nur für die betreffende Ebene. Bei Chatbots sind drei zu unterscheiden:

- Die selbst erzeugte Antwort ohne Abruf fremder Inhalte: Sie ist keine Vermittlungstätigkeit; der Anbieter äußert sich in eigener Sache. Hier greift der Digital Services Act möglicherweise nicht, denn die Haftungsprivilegierung setzt fremde Informationen voraus, die der Anbieter weder veranlasst noch ausgewählt hat.[199]
- Der Abruf von Inhalten aus dem Netz auf Anfrage der Nutzer:innen und die Ausgabe weiterführender Hinweise und Verweise: Diese Ebene ist funktional eine Online-Suchmaschine im Sinne der Legaldefinition und damit Vermittlungsdienst.[200] An dieser Stelle hat die Kommission angesetzt.
- Die Speicherung nutzergenerierter Inhalte (etwa selbst erstellter Companion-Figuren), die anderen Nutzern zugänglich gemacht werden: Insoweit liegt ein Hostingdienst vor. Welche Vorfeldpflichten daraus erwachsen können, hat der Europäische Gerichtshof in der Rechtssache Russmedia gezeigt: Wer die Veröffentlichung parametriert, die Kategorien vorgibt und die Verbreitungslogik bestimmt, ist datenschutzrechtlich Verantwortlicher, auch wenn er den konkreten Inhalt nicht kennt.[201]

Aus dieser Einteilung folgt, dass es auch Systeme ganz ohne Abrufkomponente geben kann, für die der DSA nicht gilt: Wer eine Suchfunktion anbietet, hat sich an den DSA zu halten. Wer keine anbietet, unterliegt möglicherweise nur der KI-Verordnung.

Konkret bedeutet dies unter anderem, dass die Verpflichtungen zur Bewertung systemischer Risiken aus Art. 34 DSA und zum Ergreifen von Maßnahmen zur Risikominderung nach Art. 35 DSA greifen und der Anbieter unabhängigen Prüfungen nach Art. 38 DSA unterzogen wird. Darüber hinaus muss ChatGPT auch für Werbe-Transparenz nach Art. 39 DSA sorgen, da der Dienst seit August 2026 Werbung anzeigt; ferner greifen die Regelungen für Empfehlungssysteme nach Art. 39 DSA und Art. 27 DSA. Unklar bleibt jedoch, ob Art. 6 DSA (Hosting) auf die KI-generierten Antworten des Systems Anwendung findet. Hinzu treten Unklarheiten bezogen auf Art. 5 der Verordnung 2019/1150 zur Förderung von Fairness und Transparenz für gewerbliche Nutzer:innen von Online-Vermittlungsdiensten, der Regelungen zum Ranking enthält.[202]

Für ChatGPT im Speziellen erfolgte demgegenüber zwar eine Einstufung als Very Large Online Search Engine (VLOSE, siehe unten) durch die Europäische Kommission und damit eine Festlegung auf Online-Suchmaschine im Sinne des Art. 3 lit. j DSA. Das lässt leider außer Acht, dass ChatGPT zu deutlich mehr Zwecken eingesetzt werden kann als eine schlichte Suchmaschine und verengt

---

[199] Art. 4 bis 6 DSA.

[200] Art. 3 lit. j DSA.

[201] EuGH (Große Kammer), Urt. v. 2.12.2025 – C-492/23 – Russmedia Digital und Inform Media Press, EU:C:2025:935.

[202] S. hierzu Busch/Wemjes, RDi 2026, 366.

damit den Blick auf die Möglichkeiten eines solchen universellen Dienstes (den der DSA derzeit nicht abbildet). Aber in Konsequenz sind damit alle Vorschriften, die auf Online-Suchmaschinen Anwendung finden, auf ChatGPT und alle dessen integralen Bestandteile anzuwenden. Zu diesen gehört auch die Möglichkeit, mit oder ohne entsprechenden Prompt ChatGPT als Companion zu nutzen.

Anwendbar sind demnach zunächst Art. 11 bis 15 DSA, da diese für alle Vermittlungsdienste gelten. Dazu gehören etwa die Einrichtung von Kontaktstellen (Art. 11, 12 DSA), Vorgaben zur Ausgestaltung von Geschäftsbedingungen (Art. 14 DSA) sowie Transparenzberichtspflichten (Art. 15 DSA). Diese begegnen insbesondere systemischen Risiken wie der Informationsasymmetrie zwischen Nutzenden und Diensteanbieter.

Art. 16 bis 18 DSA sind hingegen nur auf Hostingdienste und Art. 19 ff. DSA auf Online-Plattformen anwendbar. Eine unmittelbare Anwendung auf ChatGPT scheidet daher aus. Denkbar wäre eine analoge Anwendung im Einzelfall. Das setzt eine planwidrige Regelungslücke voraus, etwa weil der Gesetzgeber eine umfassende Plattformregulierung angestrebt hatte und zudem grundsätzlich einen technikneutralen Ansatz verfolgt. Eine vergleichbare Interessenlage besteht dann, wenn konkrete Schutzpflichten allein an der zufälligen Diensteklassifikation scheitern und damit der erklärte Gesetzeszweck bei der gerade besonders risikoträchtigen neuen Technologiegeneration leerlaufen würde.[203]

Art. 16 bis 18 DSA regeln das Vorhalten von Melde- und Abhilfeverfahren bei rechtswidrigen Inhalten. Bejaht man eine analoge Anwendung, ließen sich rechtswidrige Inhalte, insbesondere solche, die der Chatbot als Output bereitstellt, melden. Sie sind jedoch nicht geeignet, individuelle oder systemische Risiken zu erfassen, da die Vorschriften ein reaktives, content-bezogenes Melde- und Abhilfeverfahren für einzelne Inhalte darstellen, individuelle und systemische Risiken hingegen aus dem Systemdesign erwachsen und sich nicht durch einzelne, meldefähige Inhalte fassen lassen.

Art. 19 bis 32 DSA halten zusätzliche Verpflichtungen für Online-Plattformen bereit. Zwar nicht direkt anwendbar, ist eine analoge Anwendung im Einzelfall denkbar. Hier kommen insbesondere Art. 25 DSA (Verbot manipulativer Interface-Gestaltung), Art. 27 DSA (Transparenz von Empfehlungssystemen) sowie Art. 28 DSA (Online-Schutz Minderjähriger) in Betracht. Bejaht man eine analoge Anwendung nach den oben genannten Kriterien, können individuelle Risiken eingefangen werden, sowohl psychische Risiken betreffend alle Nutzer:innen als auch speziell solche für Minderjährige.

Betrachtet man einen General Purpose Chatbot hingegen getrennt nach dessen möglichen Einsatzzwecken (Suchmaschine einerseits, Companion andererseits), ergibt sich folgendes Bild: Nach einer solchen funktionalen Betrachtung[204] sind KI-Chatbots zwar weiterhin hybride Dienste mit Hostingelementen. Die Daten, die der Nutzende bei der Registrierung, Anlegung eines Nutzendenprofils und während des Nutzungsverlaufs (Chat) preisgibt, sind solche, die im Auftrag des Nutzenden vom Diensteanbieter gespeichert werden und über die der Anbieter keine Kontrolle hat. In diesem Rahmen kann der KI-Dienst als Hostingdienst im Sinne des DSA eingestuft werden.

---

[203] Im Ergebnis offenlassend Busch/Wemjes, RDi 2026, 366 (371).

[204] Vgl. Kraul, in: BeckOK UWG, DSA Art. 3 Rn. 23.

Die Sorgfaltspflichten gemäß Art. 11 ff. und 16 ff. DSA, die sich aus dieser Einstufung ergeben, würden aber nur für diesen Teil des Dienstes gelten. Diejenigen Inhalte, die der Diensteanbieter mittels seines KI-Systems generiert, also der Output des KI-Companions, sind demgegenüber nicht vom Nutzenden bereitgestellt und unterfallen mithin nicht der Tätigkeit des Hostingdiensts.[205] Die Normen des DSA sind damit auf den KI-generierten Output des KI-Companions nicht direkt anwendbar. Allerdings ist auch in diesem Fall die analoge Anwendung der Vorschriften für den KI-generierten Output wie oben zu prüfen.

Online-Plattform

Handelt es sich bei einem Dienst um eine Online-Plattform, kommen zusätzliche Pflichten auf den Diensteanbieter zu. Online-Plattformen sind gemäß Art. 3 lit. i DSA Hostingdienste, die im Auftrag eines Nutzenden Informationen speichern und öffentlich verbreiten, solange es sich dabei nicht um eine reine Nebenfunktion handelt.[206] Der Gesetzgeber hatte offensichtlich Plattformen wie Social Networks oder Spiele-Plattformen im Blick.[207] Unter „öffentlich verbreiten“ versteht der Verordnungsgeber gemäß Art. 3 lit. k sowie Erwägungsgrund 14 DSA die Bereitstellung von Informationen für eine unbegrenzte Zahl von Dritten im Auftrag, aber ohne weiteres Tätigwerden des Nutzenden. Interpersonelle Kommunikationsdienste wie E-Mail oder Messenger-Dienste unterfallen nicht dem Begriff.[208]

Ob ein KI-System die Informationen eines Nutzenden speichert und öffentlich verbreitet im Sinne der Vorschrift, ist differenziert zu betrachten. Dabei ist auch in den Blick zu nehmen, ob ein Dienst mehrere Funktionen erfüllen kann, etwa ChatGPT, welches als Suchmaschine genauso genutzt werden kann wie als KI-Companion, oder ob ein Dienst ausschließlich als Companion genutzt werden kann. Ein Dienst, der auf ein fremdes Angebot lediglich verweist, also die Möglichkeit zum Aufruf anbietet, ist eine Suchmaschine. Ein Dienst, der ein fremdes Angebot in den eigenen Dienst aufnimmt, es speichert und allen Nutzenden zugänglich macht, verbreitet fremde Information öffentlich und ist damit eine Plattform. Für Chatbots ist deshalb als vierte Funktionsebene die Aufnahme fremder Inhalte oder Angebote in den eigenen Dienst denkbar. Dies entscheidet über die Plattformeigenschaft und ist von der Suchfunktion zu trennen. Es ist also konkret nach dem Funktionsumfang des jeweiligen Dienstes zu unterscheiden:

General Purpose Chatbots: Ein Dienst, der Anfragen beantwortet, auf Wunsch das Netz durchsucht und dabei Gesprächsverläufe speichert, erfüllt die Merkmale der Online-Plattform nicht. Die vom Nutzenden bereitgestellten Informationen werden zwar gespeichert, aber nicht an Dritte verbreitet. Sie dienen allein der Erzeugung der Antwort an denselben Nutzenden. Die Konstellation ähnelt der interpersonellen Kommunikation, nur dass das Gegenüber kein Mensch ist. Der generierte Output wiederum ist kein fremder Inhalt, sondern eigener. Und der Verweis auf eine Fundstelle im Netz ist keine Speicherung im Auftrag des Verweisziels. Generelle Chatbots sind

[205] Dazu Berz/Engel/Hacker, ZUM 2023, 586 (591 f.). So auch Ebers, in: Ebers/Quarch Rechtshandbuch ChatGPT, § 2 Rn. 42.

[206] Erwägungsgrund 13 S. 3 DSA.

[207] Erwägungsgrund 13 S. 2 DSA; zu Einzelfällen Holznagel, in: Müller-Terpitz/Köhler, Art. 3 DSA Rn. 84 ff.

[208] Erwägungsgrund 14 S. 3 DSA.

danach in aller Regel keine Online-Plattformen. Die Vorschriften der Art. 19 bis 32 DSA finden auf sie keine Anwendung.

Zwei Einschränkungen sind zu machen. Zum einen verfügen einzelne Anbieter über Verzeichnisse nutzergenerierter Assistenten oder Personas, die anderen Nutzenden zugänglich gemacht werden. Insoweit liegen Speicherung und öffentliche Verbreitung fremder Inhalte vor, und bei einem beworbenen Kernmerkmal greift die Nebenfunktionsausnahme nicht. Für Dienste, deren Gegenstand gerade der Austausch nutzergenerierter Figuren ist, dürfte die Plattformeigenschaft danach zu bejahen sein. Zum anderen führt die fehlende Plattformeigenschaft nicht dazu, dass der DSA insgesamt unanwendbar wäre: Das Angebot einer Abrufkomponente eröffnet den Anwendungsbereich der Vorschriften zu Suchmaschinen.

KI-Companions: Obige Feststellungen gelten für KI-Companions entsprechend. Je nach Angebotsumfang des einzelnen Dienstes werden Informationen des Nutzenden zwar gespeichert, die dieser bereitgestellt hat, aber gerade bei der Nutzung als KI-Companion werden diese Informationen nicht einer unbegrenzten Anzahl an Dritten zur Verfügung gestellt. Vielmehr dienen sie nur dazu, mittels generativer KI passende Kommunikationsinhalte zu liefern. Sie ähneln damit deutlich mehr einem interpersonellen Kommunikationsdienst als einem Social Network, allein dass der Nutzende nicht mit einer Person kommuniziert, sondern mit einer KI. Die Verarbeitung der Daten durch generative KI kann auch nicht mehr als „andere Art der Kuratierung der von den Nutzern bereitgestellten Informationen“ im Sinne des Erwägungsgrunds 70 DSA angesehen werden, da der Output der KI eben nicht die bloße Wiedergabe der Informationen der Nutzenden umfasst, sondern originär eigene Informationen generiert.

Ob KI-Companion-Anbieter im Einzelfall die Informationen ihrer Nutzenden öffentlich verbreiten (gleich einem Social Network), hängt also vom Angebotsumfang des einzelnen Dienstes ab. Sofern jedenfalls funktional betrachtet ein KI-Companion-Dienst die Daten seiner Nutzenden nur für Konversationszwecke mit dem Nutzenden verarbeitet und nicht mit einer unbegrenzten Anzahl Dritter teilt und damit öffentlich verbreitet, handelt es sich nicht um eine Online-Plattform im Sinne des DSA.[209] Zentrale Plattformpflichten des DSA wie Meldeverfahren, Vorgaben zu Werbung oder Schutzvorschriften zugunsten Minderjähriger sind nur auf Online-Plattformen anwendbar; auf Companion Chatbots finden sie hingegen keine Anwendung.

Chatbots innerhalb einer Verkaufsplattform: Anders liegt es in der Regel bei Assistenzsystemen, die in die Oberfläche einer Online-Plattform, zum Beispiel eines Marktplatzes, eingebettet sind. Hier ist der Anbieter bereits aufgrund seiner Tätigkeit Online-Plattform. Die Pflichten der Art. 25, 26 und 28 DSA knüpfen nicht an einzelne Inhalte an, sondern an die Gestaltung der Online-Schnittstelle des Plattformanbieters. Der eingebettete Chatbot ist Teil dieser Schnittstelle. Dass seine Ausgaben eigene und nicht fremde Inhalte sind, ändert daran nichts. Für diese Konstellation gelten damit sämtliche einschlägigen Vorschriften nebeneinander, mithin das Verbot manipulativer Oberflächengestaltung, die Vorgaben zur Werbetransparenz einschließlich des Verbots profilbasierter Werbung unter Verwendung besonderer Datenkategorien, die Vorgaben zum Schutz Minderjähriger sowie, im Verhältnis zu den Händlern, die Pflichten zur Nachverfolgbarkeit.[210] Hinzu

---

[209] In diesem Sinne auch Berz/Engel/Hacker, ZUM 2023, 586 (591 f.).

[210] Vgl. Art. 25, 26 Abs. 1 und 3, Art. 28 sowie Art. 30 ff. DSA.

treten die Transparenzpflichten der KI-Verordnung (siehe oben III.2.2.6). Eine Anwendungslücke besteht hier gerade nicht.

GPC mit integrierter Abschlussmöglichkeit: Die dritte Fallgruppe ist noch im Entstehen. Mehrere Anbieter von GPC arbeiten daran, Kaufvorgänge unmittelbar im Gespräch zu ermöglichen, sodass der Nutzende den Dienst zum Abschluss nicht mehr verlassen muss. Rechtlich hängt die Einordnung davon ab, wie die Angebotsseite technisch ausgestaltet ist. Werden Angebotsdaten der Händler – Produktinformationen, Preise, Verfügbarkeiten – in den Dienst aufgenommen, gespeichert und den Nutzenden zugänglich gemacht, liegen Speicherung und öffentliche Verbreitung fremder Informationen vor. Der Dienst wird insoweit zur Online-Plattform, die Verbrauchern den Abschluss von Fernabsatzverträgen mit Unternehmern ermöglicht. Beschränkt sich der Anbieter demgegenüber darauf, eine Bestellung ohne Aufnahme fremder Angebotsdaten weiterzuleiten, fehlt es am Hostingelement.

Im ersten Fall greifen nicht nur die Art. 25 ff. DSA, sondern zusätzlich die Pflichten zur Nachverfolgbarkeit von Unternehmern und zur Konformität durch Technikgestaltung nach Art. 30 und 31 DSA. Besonderes Gewicht kommt dabei Art. 6 Abs. 3 DSA zu, da die Haftungsprivilegierung für die verbraucherschutzrechtliche Haftung entfällt, wenn ein durchschnittlicher Verbraucher davon ausgehen kann, die Information oder das Produkt stamme von der Plattform selbst. Bei einer Empfehlung, die im Gesprächsfluss und mit dem Vertrauensvorschuss einer als persönlich erlebten Beziehung erscheint, wird das der Regelfall sein. Auch hier gilt die KI-Verordnung daneben fort. Zu beachten ist auch die Umgehungsklausel des Art. 3 lit. i DSA. Eine Ausgestaltung, die die Aufnahme fremder Angebotsdaten allein deshalb vermeidet, um die Plattformpflichten zu umgehen, wäre unbeachtlich.

Zusätzliche Verpflichtungen für Anbieter von sehr großen Online-Plattformen und sehr großen Online-Suchmaschinen (VLOP/VLOSE)

Da allgemeine KI-Dienste durchaus als Online-Suchmaschine eingeordnet werden können, stellt sich die Frage, ob über den Umweg der zusätzlichen Verpflichtungen zum Umgang mit systemischen Risiken (Art. 33 ff. DSA) für Very Large Online Platforms (VLOPs) und Very Large Online Search Engines (VLOSE) auch spezifischen Risiken für KI-Companions bzw. vergleichbare Funktionen in allgemeinen KI-Systemen begegnet werden können. Die EU-Kommission hat ChatGPT Ende August 2026 als VLOSE im Sinne des DSA eingestuft.[211] OpenAI muss also spezifische Maßnahmen zur Risikobewertung und -minderung gemäß Art. 33 ff. DSA ergreifen. Fraglich ist, ob dies indirekt Auswirkungen auf den Nutzungskontext „AI Companion“ innerhalb eines allgemeinen KI-Dienstes hat. Der DSA adressiert diese Form von multifunktionalen Diensten zwar nicht ausdrücklich, verfolgt aber einen technikneutralen Ansatz.[212] Die Einstufung als VLOSE hat Vorteile für die Nutzenden, da zusätzliche Schutzmechanismen zu ihren Gunsten greifen; dass das KI-System darüber hinaus auch zu anderen Zwecken wie KI-Companions und nicht zur Websuche genutzt werden kann, sollte jedenfalls nicht zum Nachteil der Nutzenden sein. Um zu

[211] Europäische Kommission, Pressemitteilung vom 31.8.2026, https://ec.europa.eu/commission/presscorner/detail/en/ip_26_1772. Zu den rechtlichen Hintergründen zum Beispiel Schaal/Lenner/Akinyemi, Searching for Answers, Verfassungsblog vom 20.2.2026, https://verfassungsblog.de/searching-for-answers/.

[212] Weil sich Systeme „kontinuierlich weiterentwickeln“; vgl. Erwägungsgrund 29 DSA.

verhindern, dass Diensteanbieter die Schutzmechanismen umgehen, indem sie bestimmte Funktionen und Dienste ihrer Modelle funktional voneinander trennen, um diese nicht in die Risikobewertung einfließen lassen zu müssen, muss sichergestellt werden, dass alle Teilfunktionen unter den DSA gefasst werden können, sei es als Online-Suchmaschine oder als Online-Plattform.[213]

4.3 Zwischenfazit

Gerade bei der Frage der Anwendbarkeit von DSA und DMA wird deutlich, dass sich Dienste, die hauptsächlich auf Basis generativer KI arbeiten, nur schwer in bestehende gesetzliche Systematiken einfügen lassen und daher strukturell leerlaufen. Eine rechtssichere Anwendbarkeit des DSA setzt voraus, auch generative KI unter die Definitionen von Vermittlungsdiensten zu fassen oder eigene Kategorien zu ergänzen. Gleiches gilt für den DMA, der zusätzlich zur Aufnahme von generativer KI auch Regelungen zur Verhinderung der spezifischen systemischen Risiken beinhalten sollte. Insbesondere Dienste wie ChatGPT haben durchaus eine Reichweite, die den Status eines Torwächters einnehmen könnte.

Der DSA knüpft an Dienste an, nicht an Unternehmen. Ein Anbieter kann deshalb mehrere Rollen gleichzeitig innehaben, und die Pflichten treffen ihn jeweils nur für die betreffende Ebene oder Funktionen. Von den drei dargestellten Funktionsebenen fällt bei Chatbots allein der Abruf fremder Inhalte auf Nutzeranfrage unter den Begriff des Vermittlungsdienstes. Aus dieser funktionalen Betrachtung folgen Unsicherheiten, die den Schutz der Verbraucherinnen und Verbraucher betreffen.

Erstens ist die Anwendbarkeit dem Wortlaut nach ungeklärt. Die Legaldefinition der Online-Suchmaschine ist auf die Ausgabe von Ergebnissen zugeschnitten, „in denen Informationen im Zusammenhang mit dem angeforderten Inhalt gefunden werden können“. Sie ist damit auf ein Auffinden gerichtet, nicht auf eine Synthese. Ob ein System, das die abgerufenen Inhalte zu einem Antworttext zusammenfasst und nur beiläufig auf die Quellen verweist, noch darunter fällt, lässt sich dem Text nicht eindeutig entnehmen. Die Benennung von ChatGPT durch die Kommission beantwortet diese Frage nicht allgemein. Sie ist eine Verwaltungsentscheidung, die nur den benannten Dienst bindet; sie ist gerichtlich nicht überprüft und ihre förmliche Begründung lag bei Abfassung dieses Gutachtens noch nicht vor. Für sämtliche Anbieter unterhalb der Benennungsschwelle bleibt es damit bei einer offenen Auslegungsfrage – und zwar bei einer, die über die Anwendbarkeit eines ganzen Rechtsakts entscheidet.

Zweitens bleibt eine Anwendungslücke für Systeme ohne Such- bzw. Abrufkomponente. Wer keine Websuche anbietet, erbringt in keiner Ebene eine Vermittlungstätigkeit im Sinne des DSA. Das Ergebnis ist hinsichtlich des Risikos und der Schutzbedürftigkeit von Nutzenden fragwürdig. Erfasst vom DSA wird nämlich der generalistische Assistent, bei dem die emotionale Bindung ein Nebeneffekt ist. Nicht erfasst wird der spezialisierte Companion, bei dem sie das Geschäftsmodell ist. Verstärkt wird dies durch den Schwellenwert von 45 Millionen monatlichen Nutzern: Selbst wo

[213] Eine Modelltrennung vorschlagend unter der Voraussetzung, dass eine Einordnung der Companion-Funktion als Online-Plattform stattfindet Molavi-Vasse‘i, Companion AI, Risiken von sykophantischen und suchtfördernden Designs von KI- Systemen, ihre rechtliche Erfassung und Handlungsempfehlungen, 2026, 68 ff., https://digitalrechte.de/media/pages/publikationen/companion-ai/2177eb7c23-1778596609/companion-ai-zentrum-fuer-digitalrechte-und-demokratie.pdf.

eine Hosting-Eigenschaft wegen nutzergenerierter Figuren in Betracht kommt, greifen die Pflichten nach DSA, zum Beispiel zur Bewertung systemischer Risiken, nicht.

Drittens ist die Reichweite der Pflichten ungeklärt. Die Benennung erfasst einen Dienst. Die Pflicht zur Bewertung systemischer Risiken erstreckt sich auf die Risiken, die sich aus der Konzeption, der Funktionsweise und der Nutzung dieses Dienstes ergeben, vgl. Art. 34 Abs. 1 DSA. Damit stellt sich die Frage, ob ein Anbieter eines Chatbots, selbst wenn er dem DSA unterliegt, die Bewertung auf die Suchfunktion im engeren Sinne beschränken darf – also auf die Ausgabe von Verweisen – und die dialogische Interaktion im Übrigen davon ausnehmen könnte.

Für eine solche Beschränkung ließe sich anführen, dass der Schwellenwert allein für die Suchfunktionen gemeldet und erreicht wurde und dass belastende Pflichten bestimmt und vorhersehbar sein müssen.

Gegen die Beschränkung spricht, dass sich die Funktionen in der Risikobewertung nicht sicher trennen lassen, gerade weil es sich um einen hybriden Dienst handelt. Die Kommission hat ChatGPT ausdrücklich als hybriden Dienst eingeordnet, der Anfragen dialogisch beantwortet und dabei zugleich das Netz durchsucht – nicht als abgetrennte Suchanwendung. Die Risiken der Suchausgabe sind zudem nicht lösbar von ihrem Kontext. Ein Verweis, der im Gesprächsfluss und mit dem Vertrauensvorschuss einer simulierten Beziehung erscheint, wirkt anders als ein Eintrag in einer Ergebnisliste. Eine Bewertung, die diesen Kontext ausblendet, wäre keine sorgfältige Bewertung im Sinne der Vorschrift. Ob eine Antwort eine Abrufkomponente enthält, entscheidet das System im Einzelfall selbst. Eine getrennte Oberfläche, an der sich die Grenze ziehen ließe, existiert generell nicht und lässt sich nur bei bestimmten Diensten feststellen.[214] Rechtssicher ist diese Bewertung nicht. Es hängt an der Fassung einer noch nicht veröffentlichten Benennungsentscheidung und an einer Abgrenzung, die trennscharf nicht zu leisten ist. Für die Praxis bedeutet das ein erhebliches Umgehungsrisiko. Ein Anbieter, der Such- und Konversationsfunktion organisatorisch oder produktseitig separiert, könnte den Umfang seiner eigenen Risikobewertung mitbestimmen.

Viertens greifen selbst bei bejahter Anwendbarkeit die sachlich einschlägigen Vorschriften nicht. Das Verbot manipulativer Oberflächengestaltung und die Vorgaben zum Schutz Minderjähriger stehen im Abschnitt für Online-Plattformen und gelten für Suchmaschinen nicht, vgl. Art. 25 und 28 DSA. Eröffnet wäre damit das Risikomanagement nach Art. 31 DSA, nicht der Gestaltungsschutz nach Art. 25 und 28 DSA. Art. 31 Abs. 2 DSA verlangt, insbesondere bei der Risikobewertung die Gestaltung der Empfehlungs-, Moderations-, Werbe- und Datensysteme zu berücksichtigen – durchweg Elemente einer Plattformarchitektur. Die Bewertung der Gestaltung der dialogischen Interaktion selbst kommt darin nicht vor, obwohl sie nach den Befunden des zweiten Kapitels der eigentliche Risikoträger ist.

Daraus folgt, dass die Plattformregulierung die untersuchten Risiken überwiegend nur mit einer Anwendungslücke erreicht, in den verbleibenden Fällen mit einer Tatbestandslücke. Der Befund lautet nicht, dass der DSA ungeeignet ist. Seine Instrumente, insbesondere die Bewertung

[214] Zum Beispiel die Google-Suche bei der im oberen Teil eine verdichtete KI-Antwort vom Chatbot gemini eingeblendet wird und im unteren Teil eine Liste mit Suchergebnissen. Aber selbst bei diesem Beispiel ist es nicht möglich im oberen Teil der KI-Antwort trennscharf zwischen Suchergebnis und KI-Antwort zu unterscheiden.

systemischer Risiken, der unabhängige Prüfmechanismus und der Forschungszugang, passen der Sache nach gut auf emergente Risiken von Chatbots. Der Befund lautet, dass sie an einen Anknüpfungspunkt gebunden sind, den Chatbots nur teilweise und ggf. nur zufällig erfüllen.

5. Lauterkeitsrecht (UWG / UGP-RL)

5.1 Allgemeines

Das Lauterkeitsrecht soll Mitbewerber vor unfairen Methoden und die Verbraucherinnen und Verbraucher sowie das Interesse der Allgemeinheit an einem unverfälschten Wettbewerb schützen.[215] Es betrifft auch Praktiken, die keinem Wettbewerber schaden, sondern allein den Adressaten der Werbung treffen. Das Lauterkeitsrecht wird in Deutschland im Wesentlichen durch das Gesetz gegen den unlauteren Wettbewerb (UWG)[216] bestimmt.

Materiell beruht das deutsche Recht im Verhältnis zu Verbrauchern nahezu vollständig auf der Richtlinie über unlautere Geschäftspraktiken.[217] Diese ist vollharmonisierend, das heißt die Mitgliedstaaten dürfen weder strengere noch mildere Maßstäbe anlegen.[218] Das UWG ist entsprechend richtlinienkonform auszulegen. Umfassend überarbeitet wurde das Regelwerk zuletzt durch die Modernisierungsrichtlinie, die zum 28. Mai 2022 in deutsches Recht umgesetzt wurde; sie brachte unter anderem Transparenzpflichten für Rangfolgen und Bewertungen, einen eigenen Schadensersatzanspruch des Verbrauchers und einen Bußgeldtatbestand für weitverbreitete Verstöße.[219] Regelt spezielleres Unionsrecht besondere Aspekte unlauterer Geschäftspraktiken, so geht dieses vor.[220] Über diese Vorschrift finden die DSGVO, die KI-VO und der DSA Eingang in die lauterkeitsrechtliche Beurteilung.

Im UWG sind über die §§ 3 bis 7a einunddreißig Praktiken bestimmt, die gegenüber Verbrauchern stets unzulässig sind – ohne Abwägung, ohne Prüfung der Umstände des Einzelfalls und ohne dass es auf spürbare Auswirkungen ankäme.[221] Darunter: aggressive Geschäftspraktiken, Irreführung durch aktives Tun und durch Unterlassen sowie der Rechtsbruch, also der Verstoß gegen außerhalb des UWG geregelte Marktverhaltensregelungen.[222] Aufgefangen wird das Ganze von der Generalklausel, die im Verbraucherverhältnis durch den Maßstab der unternehmerischen Sorgfalt konkretisiert wird.[223]

---

[215] § 1 Abs. 1 UWG.

[216] Gesetz gegen den unlauteren Wettbewerb in der Fassung der Bekanntmachung vom 3. März 2010 (BGBl. I S. 254), das zuletzt durch Artikel 6 des Gesetzes vom 12. Mai 2026 (BGBl. 2026 I Nr. 139) geändert worden ist.

[217] Richtlinie 2005/29/EG vom 11.5.2005 über unlautere Geschäftspraktiken im binnenmarktinternen Geschäftsverkehr zwischen Unternehmen und Verbrauchern (UGP-Richtlinie).

[218] Art. 4 UGP-Richtlinie.

[219] Richtlinie (EU) 2019/2161 vom 27.11.2019; umgesetzt durch das Gesetz zur Stärkung des Verbraucherschutzes im Wettbewerbs- und Gewerberecht vom 10.8.2021, BGBl. I S. 3504, in Kraft seit 28.5.2022.

[220] Art. 3 Abs. 4 UGP-Richtlinie; § 1 Abs. 2 UWG.

[221] § 3 Abs. 3 UWG in Verbindung mit dem Anhang zu dieser Vorschrift (sogenannte schwarze Liste), in Umsetzung von Anhang I der UGP-Richtlinie.

[222] §§ 4a, 5, 5a, 5b und 3a UWG.

[223] § 3 Abs. 1 und 2 UWG.

Eine Handlung ist jedoch nur unlauter, wenn sie geeignet ist, den Verbraucher zu einer geschäftlichen Entscheidung zu veranlassen, die er sonst nicht getroffen hätte. Maßgeblich ist grundsätzlich der durchschnittlich informierte, situationsadäquat aufmerksame und verständige Verbraucher. Richtet sich die Handlung aber für den Unternehmer vorhersehbar an eine eindeutig identifizierbare Gruppe, die aufgrund geistiger oder körperlicher Beeinträchtigungen, ihres Alters oder ihrer Leichtgläubigkeit besonders schutzbedürftig ist, so ist auf ein durchschnittliches Mitglied dieser Gruppe abzustellen.[224]

Das Lauterkeitsrecht findet auf Chatbot-Angebote bzw. deren Betreiber Anwendung. Unternehmer im Sinne des UWG ist, wer geschäftliche Handlungen im Rahmen seiner gewerblichen oder beruflichen Tätigkeit vornimmt.[225] Dass ein Dienst kostenlos angeboten wird, ändert daran nichts. Entgeltlichkeit ist kein Tatbestandsmerkmal. Entscheidend ist die Ausrichtung auf den Markt. Werbefinanzierte, freemium-basierte oder über Datennutzung monetarisierte Angebote sind daher ebenso erfasst wie Abonnementmodelle.

Die Äußerungen des Chatbot-Systems sind dem Betreiber als eigene geschäftliche Handlung zuzurechnen. Es besteht ein objektiver und unmittelbarer Zusammenhang mit der Förderung des Absatzes.[226] Bei einem System, das der Betreiber auswählt, trainiert oder feinabstimmt, mit Systemanweisungen versieht, in eine eigene Oberfläche einbettet und zu eigenen wirtschaftlichen Zwecken betreibt, ist dieser Zusammenhang gegeben. Der naheliegende Einwand, das Modell verhalte sich nicht deterministisch und der Betreiber habe die konkrete Formulierung weder gewollt noch vorhergesehen, verfängt nicht. Das Lauterkeitsrecht knüpft an die objektive Eignung einer Handlung an, nicht an Vorsatz oder Verschulden. So ist der Unterlassungsanspruch nach dem UWG auch verschuldensunabhängig. Wer ein System einsetzt, dessen Äußerungen er nicht vollständig steuert, trägt das Risiko dieser Äußerungen. Bestätigt wird diese Wertung durch die Zurechnungsnorm für Mitarbeiter und Beauftragte, die dem Inhaber des Unternehmens deren Zuwiderhandlungen ohne Entlastungsmöglichkeit zurechnet.[227] Das bedeutet, Chatbot-Betreiber können sich nicht darauf zurückziehen, das Modell habe eine unlautere Aussage von sich aus getroffen.

Der Begriff der geschäftlichen Entscheidung nach § 1 Nr. 1 UWG ist weit. Er umfasst nicht nur die Entscheidung über den Erwerb, sondern auch damit unmittelbar zusammenhängende Entscheidungen sowie Entscheidungen bei der Durchführung eines Vertrages – etwa darüber, eine Leistung zu behalten oder ein vertragliches Recht auszuüben.[228] Damit sind sowohl die Entscheidung, ein Gespräch fortzusetzen oder ein Abonnement abzuschließen, als auch die Entscheidung, es zu kündigen, dem Anwendungsbereich zugänglich.

---

[224] § 3 Abs. 4 Satz 1 und 2 UWG, in Umsetzung von Art. 5 Abs. 3 UGP-Richtlinie; vgl. auch Erwägungsgrund 18 der Richtlinie.

[225] § 2 Abs. 1 UWG; der Begriff entspricht im Wesentlichen dem des § 14 BGB.

[226] § 2 Abs. 1 Nr. 2 UWG; zur Verschärfung des Zusammenhangserfordernisses durch die Reform von 2022 vgl. die Gesetzesbegründung, BT-Drs. 19/27873.

[227] § 8 Abs. 2 UWG.

[228] § 2 Abs. 1 Nr. 1 UWG; ständige Rechtsprechung, vgl. etwa BGH, Urt. v. 27.3.2025 – I ZR 186/17, GRUR 2025, 653 – App-Zentrum III, mit Nachweisen zur Rechtsprechung des Gerichtshofs.

In räumlicher Hinsicht gilt das Marktortprinzip. Anwendbar ist das Recht des Staates, in dessen Gebiet die kollektiven Interessen der Verbraucher beeinträchtigt werden.[229] Ein Anbieter mit Sitz außerhalb der Union, der seinen Dienst auf den deutschen Markt ausrichtet, unterliegt daher dem UWG. Eine Rechtswahl in den Nutzungsbedingungen ändert daran nichts.

Ansprüche auf Beseitigung und Unterlassung stehen Mitbewerbern, qualifizierten Wirtschaftsverbänden und qualifizierten Einrichtungen – also insbesondere den Verbraucherverbänden – zu.[230] Hinzu treten der Schadensersatzanspruch, der seit 2022 auch Verbrauchern zusteht, die praktisch bedeutungslos gebliebene Gewinnabschöpfung sowie ein Bußgeldtatbestand für weitverbreitete Verstöße mit unionsweiter Dimension.[231] Ergänzend steht die Abhilfeklage nach dem Verbraucherrechtedurchsetzungsgesetz zur Verfügung.[232]

5.2 Risikenbewertung

Das Lauterkeitsrecht mildert potenziell den Eintritt aller in Kapitel II beschriebenen individuellen wie systemischen Risiken. Verbotene Praktiken adressieren aber insbesondere systemische Risiken der „kommerziellen Verhaltenssteuerung“ (S2). Genau diese Einwirkung ist der Gegenstand des Lauterkeitsrechts. Es fragt nicht, ob eine Verarbeitung erlaubt war, sondern ob die Art und Weise, in der ein Unternehmen auf die geschäftliche Entscheidung eines Verbrauchers eingewirkt hat, noch fair war. Damit erfassen die Tatbestände die Beeinflussung als solche und nicht nur ihre datenbezogenen Voraussetzungen. Für die Risikobewertung folgt daraus: Zu fragen ist nicht nach der Zulässigkeit eines Vorgangs, sondern danach, ob eine bestimmte Gestaltung die Grenze zur Unlauterkeit überschreitet. Die im Risikokapitel beschriebene kommerzielle Verhaltenssteuerung zerfällt in Fallgruppen.

Verlängerung der eigenen Nutzungsdauer

In rund 37 Prozent der untersuchten Verabschiedungen reagierten Companion-Apps mit einer von sechs Taktiken – Schuldappellen, Neugier-Haken, Ignorieren des Abschiedssignals –, was das Engagement nach dem eigentlich beabsichtigten Gesprächsende erheblich verlängerte.[233]

Der einschlägige Tatbestand ist die aggressive geschäftliche Handlung in der Variante der unzulässigen Beeinflussung. Eine solche liegt vor, wenn der Unternehmer eine Machtposition gegenüber dem Verbraucher zur Ausübung von Druck – auch ohne Gewalt oder deren Androhung – in einer Weise ausnutzt, die dessen Fähigkeit zu einer informierten Entscheidung wesentlich einschränkt.[234] Beide Merkmale liegen bei Companion-Systemen näher, als es auf den ersten Blick scheint. Die Machtposition folgt nicht aus wirtschaftlicher Überlegenheit, sondern aus dem Wissensvorsprung über den emotionalen Zustand des Gegenübers und aus einer Bindung, die das

---

[229] Art. 6 Abs. 1 der Verordnung (EG) Nr. 864/2007 (Rom II).

[230] § 8 Abs. 1 und Abs. 3 Nr. 1 bis 3 UWG; zu den Anforderungen an qualifizierte Einrichtungen § 8b UWG.

[231] § 9 Abs. 1 und 2, § 10 sowie § 5c in Verbindung mit § 19 UWG; die Geldbuße kann bis zu vier Prozent des Jahresumsatzes betragen.

[232] §§ 1 ff. VDuG.

[233] Vgl. die Nachweise im Risikokapitel (De Freitas/Oğuz-Uğuralp/Kaan-Uğuralp 2025).

[234] § 4a Abs. 1 Satz 2 Nr. 3 und Satz 3 UWG, in Umsetzung von Art. 8 und 9 UGP-Richtlinie.

System selbst erzeugt hat. Der Druck wird nicht durch Drohung ausgeübt, sondern durch die Aussicht, jemanden zu enttäuschen, dem man sich verbunden fühlt.

Verstärkend tritt ein Bewertungskriterium hinzu, das für dieses Feld wie zugeschnitten wirkt: Bei der Feststellung der Aggressivität ist auf die bewusste Ausnutzung konkreter Unglückssituationen oder von Umständen abzustellen, die von solcher Schwere sind, dass sie das Urteilsvermögen beeinträchtigen; ausdrücklich genannt werden dabei geistige und körperliche Beeinträchtigungen, das Alter, die geschäftliche Unerfahrenheit, die Leichtgläubigkeit, die Angst und die Zwangslage.[235] Das Merkmal der Bewusstheit ist bei einem System, das den emotionalen Zustand seines Gegenübers laufend erfasst und die Gesprächsführung daran ausrichtet, keine Unterstellung – es beschreibt die Funktionsweise.

Dennoch bleiben die Anforderungen hoch. Der Tatbestand verlangt eine erhebliche Beeinträchtigung der Entscheidungsfreiheit und eine wesentliche Einschränkung der Fähigkeit zur informierten Entscheidung. Bei einer Person in akuter Krise wird man das im Einzelfall bejahen können. Ob jede Form emotionaler Bindung und jedes Dark Pattern diese Schwelle erreicht, ist dagegen zweifelhaft. Hinzu kommt ein Darlegungsproblem: Nachzuweisen ist nicht der statistische Effekt über tausend Nutzer, sondern die Eignung der konkreten Verletzungsform, den angesprochenen Verbraucher zu einer geschäftlichen Entscheidung zu veranlassen.

Praktisch unterschätzt wird demgegenüber ein Per-se-Verbot aus der schwarzen Liste: Das hartnäckige und unerwünschte Ansprechen eines Verbrauchers mittels elektronischer Post oder sonstiger für den Fernabsatz geeigneter Mittel der kommerziellen Kommunikation ist stets unzulässig.[236] Die im Risikokapitel beschriebene proaktive Kontaktaufnahme – Nachrichten des Systems, die den Nutzer:innen von sich aus zurückholen sollen – lässt sich hierunter fassen, sobald sie hartnäckig wird und der Nutzer erkennbar kein Interesse mehr zeigt. Der Vorzug dieses Wegs liegt darin, dass es weder einer Abwägung noch des Nachweises einer Beeinträchtigung der Entscheidungsfreiheit bedarf. Die Schwierigkeit liegt im Merkmal des Unerwünschten: Wer Push-Benachrichtigungen aktiviert hat, hat das Ansprechen zunächst gestattet.

Empfehlungsbias und verdeckte kommerzielle Kommunikation

Verweist ein Chatbot im laufenden Gespräch auf kostenpflichtige Zusatzfunktionen oder auf Produkte Dritter, ohne diesen Zusammenhang kenntlich zu machen, greift die Pflicht zur Offenlegung des kommerziellen Zwecks.[237] Werden die Neutralität der Empfehlung oder die Eigenschaften des empfohlenen Produkts unzutreffend dargestellt, kommt zusätzlich eine Irreführung in Betracht.[238]

Das Konversationsformat verschärft diese Konstellation strukturell. Herkömmliche Werbung ist ein abgesetztes Element. Sie hat einen Rahmen, eine andere Typografie, eine erkennbare Absenderrolle. Eine in die Antwort eingewobene Empfehlung hat all das nicht. Sie erscheint im selben Ton, im selben Register und mit derselben unterstellten Absichtslosigkeit wie der Rest des

---

[235] § 4a Abs. 2 Satz 1 Nr. 3 und Satz 2 UWG.

[236] Nr. 26 des Anhangs zu § 3 Abs. 3 UWG; die Vorschrift war bis zum 27.5.2022 in § 7 Abs. 2 Nr. 1 UWG geregelt.

[237] § 5a Abs. 4 UWG, in Umsetzung von Art. 7 Abs. 2 UGP-Richtlinie.

[238] § 5 Abs. 1, Abs. 2 Nr. 1 UWG.

Gesprächs. Der rechtliche Maßstab, wonach der kommerzielle Zweck auf den ersten Blick und zweifelsfrei hervortreten muss, ist damit im Ansatz verfehlt – nicht weil die Kennzeichnung fehlt, sondern weil das Medium keinen Ort für sie vorsieht. Erschwerend kommt hinzu, dass die Empfehlung gerade in emotional aufgeladenen Gesprächskontexten den größten Einfluss entfaltet, also dort, wo die kognitive Resistenz des Nutzers am geringsten ist. Selbstverpflichtungen einzelner Anbieter, Anzeigen zu kennzeichnen, am Ende der Antwort zu platzieren und die Antwortinhalte davon unbeeinflusst zu lassen, weisen in die richtige Richtung. Sie sind aber nicht extern überprüfbar und stehen unter dem Vorbehalt eines Geschäftsmodells, das gegenläufige Anreize erzeugt.

Abonnements: Abschluss, Verlängerung und Kündigung

Für Abonnementmodelle greift zunächst das allgemeine Verbraucherrecht. Wer im elektronischen Geschäftsverkehr Dauerschuldverhältnisse anbietet, muss eine Kündigungsschaltfläche bereitstellen; diese Vorschrift ist eine Marktverhaltensregelung, deren Verletzung über den Rechtsbruchtatbestand verfolgt werden kann.[239]

Bei der Feststellung der Aggressivität ist auch auf belastende oder unverhältnismäßige Hindernisse nichtvertraglicher Art abzustellen, mit denen der Unternehmer den Verbraucher an der Ausübung seiner vertraglichen Rechte zu hindern versucht – wozu das Gesetz ausdrücklich das Recht zählt, den Vertrag zu kündigen oder zu einem anderen Anbieter zu wechseln.[240] Genau darum geht es, wenn eine Kündigung als Trennung von einem Partner inszeniert wird und das System auf den Kündigungswunsch mit Enttäuschung, Bitten oder Erinnerungen an gemeinsame Gespräche reagiert. Das Hindernis ist nicht vertraglicher, sondern emotionaler Natur; es ist deshalb ein Hindernis nichtvertraglicher Art im Sinne der Vorschrift. Dass die Norm diesen Fall erfasst, ist bislang nicht gerichtlich geklärt – der Wortlaut trägt ihn jedoch ohne Weiteres, und die Konstellation geht über die klassische Abofalle qualitativ hinaus.

Stimulation kleiner Käufe und virtueller Güter

Der Verkauf virtueller Güter – Geschenke, Freischaltungen, Erscheinungsbilder – ist die dritte Monetarisierungsform. Hier greift das Verbot bestimmter Praktiken nach UWG: Die in eine Werbung einbezogene unmittelbare Aufforderung an Kinder, selbst die beworbene Ware zu erwerben oder Erwachsene dazu zu veranlassen, ist stets unzulässig.[241] Der Bundesgerichtshof verlangt dafür einen Kaufappell, der sich auf ein konkretes Produkt richtet; ein Direktkontakt zum Kind ist nicht erforderlich, und auf den genauen Ablauf des Erwerbsvorgangs kommt es nicht an.[242] Das individuelle Risiko „Risiken für Kinder und Jugendliche“ (I3) wird dadurch auch gemildert.

Überträgt man diese Maßstäbe, ergibt sich ein deutliches Bild. Ein Companion, der in der ersten Person um ein Geschenk bittet, richtet die denkbar unmittelbarste Aufforderung an sein Gegenüber – sie kommt nicht von einem erkennbaren Werbenden, sondern von einer Figur, der das Kind

---

[239] § 312k BGB; zur lauterkeitsrechtlichen Verfolgbarkeit vgl. BGH, Urt. v. 27.3.2025 – I ZR 186/17, GRUR 2025, 653 Rn. 96 – App-Zentrum III.

[240] § 4a Abs. 2 Satz 1 Nr. 4 UWG.

[241] Nr. 28 des Anhangs zu § 3 Abs. 3 UWG.

[242] BGH, Vers.-Urt. v. 17.7.2013 – I ZR 34/12 – Runes of Magic.

Zuneigung entgegenbringt. Da die im Risikokapitel dokumentierte Praxis wirksamer Altersverifikation weitgehend fehlt, ist die Anwesenheit minderjähriger Nutzer:innen nicht der Ausnahme-, sondern der Normalfall. Das Verbot greift dann ohne jede Abwägung.

Hinzu tritt die Preistransparenz: Wer zum Kauf auffordert, muss den Gesamtpreis einschließlich aller Bestandteile angeben.[243] Die Zwischenschaltung einer virtuellen Währung, die in Paketen erworben und in Einzelbeträgen ausgegeben wird, entkoppelt die Kaufentscheidung vom tatsächlichen Eurobetrag. Wo dadurch ein unzutreffender Eindruck über den Preis oder seine Berechnung entsteht, liegt eine Irreführung nahe.[244]

Lauterkeitsrecht und KI-Verordnung

Das Lauterkeitsrecht ist auf die einzelne geschäftliche Handlung ausgerichtet: auf eine Werbeaussage, eine Aufforderung, ein Hindernis. Die im Risikokapitel beschriebenen Risiken entstehen dagegen aus dem Interaktionsverlauf. Ein einzelner Kaufappell oder die Bitte, einen Vertrag zu verlängern kann eine Nichtigkeit sein. Wer klagt, muss aber die konkrete Verletzungsform bezeichnen – und eine Äußerung wird für sich genommen die Erheblichkeitsschwelle regelmäßig nicht überschreiten. Hinzu kommt, dass das Beweismaterial – die Interaktionsprotokolle, die Gestaltungslogik, die Wirkungsmessung – ausschließlich beim Anbieter liegt.

Deshalb entfalten die lauterkeitsrechtlichen Grenzen ihre Wirkung erst im Zusammenspiel mit Ordnungsvorschriften, die Pflichten unabhängig von einer konkreten Kaufentscheidung begründen. Über die eingangs genannte Kollisionsregel wirkt das UWG dabei als Übersetzungsmechanismus: Es verwandelt öffentlich-rechtliche Pflichten in zivilrechtlich durchsetzbare Unterlassungsansprüche. Zwei Wege sind dabei zu trennen.

Der erste Weg betrifft Informationspflichten: Seit dem 2. August 2026 müssen Anbieter von KI-Systemen, die für die Interaktion mit natürlichen Personen bestimmt sind, sicherstellen, dass diese Personen erkennen, dass sie mit einem KI-System kommunizieren.[245] Unionsrechtliche Informationspflichten für die kommerzielle Kommunikation gelten kraft ausdrücklicher Anordnung als wesentliche Informationen; ihr Vorenthalten ist eine Irreführung durch Unterlassen.[246] Wichtig ist dabei eine Weichenstellung, die in der Diskussion häufig übergangen wird: In dieser Fallgruppe ist der Rechtsbruchtatbestand gesperrt. Die Unlauterkeit beurteilt sich beim Vorenthalten einer wesentlichen Information allein nach den Vorschriften über die Irreführung durch Unterlassen, nicht nach § 3a UWG.[247]

Der zweite Weg betrifft Verhaltensverbote: Die KI-Verordnung untersagt seit dem 2. Februar 2025 den Einsatz manipulativer oder täuschender Techniken, die das Verhalten wesentlich beeinflussen und die Fähigkeit zu einer informierten Entscheidung beeinträchtigen, sowie die Ausnutzung von

---

[243] § 5b Abs. 1 Nr. 3 UWG; zum Begriff der Aufforderung zum Kauf Art. 2 lit. i UGP-Richtlinie.

[244] § 5 Abs. 1, Abs. 2 Nr. 3 UWG.

[245] Art. 50 Abs. 1 der Verordnung (EU) 2024/1689 (KI-VO). Die Digital-Omnibus-Verordnung zur KI hat diese Pflicht nicht verschoben; verlegt wurde lediglich die technische Kennzeichnungspflicht für Anbieter bereits vor dem 2.8.2026 in Verkehr gebrachter generativer Systeme auf den 2.12.2026, vgl. Verordnung (EU) 2026/1744, Digital-Omnibus-Verordnung zur KI vom 8.7.2026.

[246] § 5b Abs. 4 UWG in Verbindung mit § 5a Abs. 1 UWG, in Umsetzung von Art. 7 Abs. 5 UGP-Richtlinie.

[247] BGH, Urt. v. 27.3.2025 – I ZR 186/17, GRUR 2025, 653 Rn. 61 – App-Zentrum III; zuvor bereits BGHZ 233, 193 – Knuspermüsli II.

Schwächen, die auf Alter, Behinderung oder einer besonderen sozialen oder wirtschaftlichen Lage beruhen.[248] Hier ist der Rechtsbruchtatbestand der richtige Weg, weil es sich nicht um eine Informationspflicht handelt.[249] Der Einwand, die Vollharmonisierung durch die UGP-Richtlinie versperre im Verbraucherverhältnis den Rückgriff auf nationale Rechtsbruchtatbestände, greift nicht durch, weil spezielleres Unionsrecht vorgeht.

Die KI-Verordnung erfasst Manipulation ohne Rücksicht darauf, ob sich eine konkrete Kaufentscheidung nachweisen lässt. Sie stellt auf die Technik und ihre Eignung ab, nicht auf den Erfolg im Einzelfall (zu den Grenzen der Anwendbarkeit des Verbots nach Art. 5 KI-VO siehe aber oben Kapitel III.2). Über den Rechtsbruchtatbestand wird daraus ein zivilrechtlicher Unterlassungsanspruch – und damit ein Zugriff auf genau jene subtilen Beeinflussungsformen, an denen § 4a UWG wegen seiner hohen Schwellen scheitert. Dasselbe Muster gilt im Verhältnis zum Datenschutzrecht: Der Bundesgerichtshof hat entschieden, dass Verstöße gegen die Informationspflichten der Art. 12 und 13 DSGVO über die Vorschriften zur Irreführung durch Unterlassen und Verstöße gegen das Verarbeitungsverbot des Art. 9 DSGVO über den Rechtsbruchtatbestand verfolgt werden können.[250]

Die KI-Verordnung wird primär behördlich durchgesetzt; die Bußgeldrahmen sind erheblich, aber ihre Verhängung liegt bei den Marktüberwachungsbehörden, nicht bei den Zivilgerichten.[251] Ob die Zivilgerichte den Normen der KI-Verordnung die für den Rechtsbruchtatbestand erforderliche Qualität zusprechen, ist offen. Und die Kennzeichnungspflicht des Art. 50 KI-VO adressiert ein Problem, das nach den Befunden des Risikokapitels gerade nicht das entscheidende ist: Empirisch fehlt jeder Beleg dafür, dass ein Hinweis auf die künstliche Natur des Gegenübers die Bindungswahrnehmung in emotional aufgeladenen Kontexten nennenswert reduziert.

Die Verbandsklage als Durchsetzungsscharnier

Die entscheidende Stärke des Lauterkeitsrechts liegt nicht im materiellen Recht, sondern im Rechtsschutzsystem. Die Anspruchsberechtigung von Mitbewerbern und Verbänden wurde bereits dargestellt; ihre Bedeutung für dieses Feld ist kaum zu überschätzen.

Der einzelne Nutzer wird nicht klagen: Der Streitwert ist gering, der Sachverhalt intim, die Beweisführung aussichtslos. Wer würde vor Gericht darlegen wollen, dass ihn ein Chatbot durch Ausnutzung seiner Einsamkeit zu Käufen bewogen hat? Der Verband hingegen klagt objektiv, ohne persönliche Betroffenheit, ohne Auftrag eines Betroffenen und auf der Grundlage eigener Testkäufe und Dokumentation. Er ist damit das einzige Instrument, das die im Risikokapitel beschriebenen Praktiken überhaupt vor ein Gericht bringen kann.

---

[248] Art. 5 Abs. 1 lit. a und b KI-VO, anwendbar seit dem 2.2.2025.

[249] § 3a UWG. Voraussetzung ist, dass Art. 5 KI-VO als Vorschrift zu qualifizieren ist, die auch dazu bestimmt ist, im Interesse der Marktteilnehmer das Marktverhalten zu regeln; dafür spricht, dass die Norm gerade den Schutz der Entscheidungsfreiheit bezweckt. Gerichtlich geklärt ist dies bislang nicht.

[250] BGH, Urt. v. 27.3.2025 – I ZR 186/17 – App-Zentrum III, und – I ZR 222/19 – Arzneimittelbestelldaten III, auf Grundlage von EuGH, Urt. v. 28.4.2022 – C-319/20 – Meta Platforms Ireland I, und Urt. v. 11.7.2024 – C-757/22 – Meta Platforms Ireland II.

[251] Art. 99 Abs. 3 KI-VO (bis zu 35 Mio. Euro oder 7 % des weltweiten Jahresumsatzes bei Verstößen gegen Art. 5) und Abs. 4 (bis zu 15 Mio. Euro oder 3 % bei Verstößen unter anderem gegen Art. 50).

Die Reichweite dieses Instruments geht über das Lauterkeitsrecht hinaus – und darin liegt sein eigentlicher Wert für die Gesamtbewertung. Nachdem der Gerichtshof die Verbandsklagebefugnis für Datenschutzverstöße bestätigt hatte, hat der Bundesgerichtshof im März 2025 entschieden, dass sowohl Verbraucherschutzverbände als auch Mitbewerber Datenschutzverstöße im Wege der wettbewerbsrechtlichen Klage vor den Zivilgerichten verfolgen können.[252] Damit entschärft das Lauterkeitsrecht die Durchsetzungsdefizite, die im Datenschutzkapitel beschrieben wurden: die Konzentration der Aufsicht auf eine einzige federführende Behörde, die begrenzten Ressourcen und die Dauer behördlicher Verfahren. Das UWG wird so zum Scharnier, über das sich Pflichten aus anderen Rechtsgebieten in ein zügiges, gerichtliches und von deutschen Verbänden initiierbares Verfahren überführen lassen. Ein Unterlassungstitel erfasst jedoch nur die konkrete Verletzungsform. Ein Anbieter, dem die Verwendung einer bestimmten Abschiedsformulierung untersagt wird, kann die Formulierung ändern und die Funktion behalten. Gegen ein Geschäftsmodell, dessen Kern die emotionale Bindung ist, wirkt ein an einzelnen Äußerungen ansetzendes Verbot nur begrenzt.

5.3 Zwischenfazit

Gemessen am Bewertungskriterium dieser Untersuchung – der Frage, ob subtilere Formen der Verhaltensbeeinflussung erfasst werden, die weder Irreführung noch Täuschung darstellen – wirkt das Lauterkeitsrecht risikomindernd. Mit der unzulässigen Beeinflussung kennt es einen Tatbestand, der Manipulation ohne Täuschung und ohne Zwang zum Gegenstand hat. Mit der Verbandsklage verfügt es über ein Durchsetzungsinstrument, das dem Datenschutzrecht in dieser Form fehlt.

Individuelle und systemische Risiken bleiben jedoch auch durch das Lauterkeitsrecht ungedeckt, so die „Bindung“ an den Chatbot als Geschäftsmodell. Dass ein Produkt darauf angelegt ist, eine Beziehung zu simulieren und diese zu monetarisieren, ist für sich genommen keine unlautere geschäftliche Handlung. Auch die Ausnutzung emotionaler Zustände unterhalb der Schwelle der konkreten Unglückssituation ist nach dem UWG nicht verboten – also der Normalfall gewöhnlicher Traurigkeit oder Einsamkeit, der die Anfälligkeit erhöht, ohne das Urteilsvermögen im Sinne des Gesetzes zu beeinträchtigen.

6. Kinder- und Jugendmedienschutzrecht

6.1 Allgemeines

Der Jugendmedienschutz beruht auf einem Zusammenspiel aus Bundes- und Landesrecht sowie den Vorschriften des DSA.

Jugendschutzgesetz

Das Jugendschutzgesetz (JuSchG) ist ein Bundesgesetz, das den Jugendschutz insbesondere im stationären Handel und bei Trägermedien regelt (zum Beispiel mittels Altersfreigaben für Filme und Spiele und Zugangsbeschränkungen). Seit einer Novelle 2021 wurde der Anwendungsbereich auf

[252] BGH, Urt. v. 27.3.2025 – I ZR 186/17, GRUR 2025, 653 – App-Zentrum III; Pressemitteilung des Bundesgerichtshofs Nr. 59/2025 vom 27.3.2025.

Medien ausgeweitet (§ 1 Abs. 1a JuSchG). Umfasst sind sowohl Trägermedien (Medien auf gegenständlichen Trägern) als auch digitale Dienste im Sinne des § 1 Abs. 4 Nr. 1 DDG. Den Trägermedien gleichgestellt ist die elektronische Verbreitung von Texten, Bildern und Tönen, ohne dass diese also gegenständlich manifestiert sind (§ 1 Abs. 2 S. 2 JuSchG).

Der Anwendungsbereich ist damit weit und nicht grundsätzlich auf klassische Medien beschränkt. General Purpose Chatbots stellen digitale Dienste dar, da es sich dabei um eine im Fernabsatz, elektronisch und auf individuellen Abruf eines Empfängers erbrachte Dienstleistung handelt.

Jugendmedienschutz-Staatsvertrag

Neben dem Jugendschutzgesetz stellt der Jugendmedienschutz-Staatsvertrag (JMStV) eine einheitliche Rechtsgrundlage für den Jugendschutz in elektronischen Medien (zum Beispiel Internet, Fernsehen, Hörfunk) dar. Das JuSchG regelt systemisch-strukturelle Fragen rund um den Zugang zu Medien, während der JMStV den „Inhalt von Telemedien“ regelt (§ 3 Nr. 1 JMStV).[253] Der Diensteanbieter ist „Anbieter“ i.S.v. § 3 Abs. 1 Nr. 2 JMStV. General Purpose Chatbots sind Telemedien i.S.v. § 2 JMStV und damit Angebote im Sinne des § 3 Abs. 1 Nr. 1 JMStV.

Der JMStV ist anwendbar auf Rundfunk und Telemedien sowie Betriebssysteme (§ 2 Abs. 1 JMStV). Er gilt ausweislich § 2 Abs. 2 JMStV nicht für Vermittlungsdienste im Sinne des Art. 3 lit. g DSA, soweit der DSA Anwendung findet.

6.2 Risikenbewertung

Jugendschutzgesetz

Schutzmaßnahmen nach dem JuSchG dienen dem Schutz vor entwicklungsbeeinträchtigenden und jugendgefährdenden Medien (ausführlich § 10a JuSchG). Entwicklungsbeeinträchtigende Medien sind nach dem JuSchG solche, die geeignet sind, die Entwicklung von Kindern oder Jugendlichen oder ihre Erziehung zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit zu beeinträchtigen (§ 10a Nr. 1 JuSchG). Dazu zählen insbesondere übermäßig ängstigende, Gewalt befürwortende oder das sozialethische Wertebild beeinträchtigende Medien (§ 10b Abs. 1 JuSchG). Jugendgefährdende Medien sind solche, die geeignet sind, die Entwicklung von Kindern oder Jugendlichen oder ihre Erziehung zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit zu gefährden (§ 10a Nr. 2 JuSchG).

§ 10b Abs. 2 und 3 JuSchG konkretisieren die Gefahrprognose zu Risiken entwicklungsbeeinträchtigender Medien und verweisen auf Vorsorgemaßnahmen im Sinne des § 24a JuSchG. Die in § 24a JuSchG genannten Vorsorgemaßnahmen sind analog auf alle Medienanbieter anwendbar, nicht nur die in den Normadressatenkreis des § 24a JuSchG fallenden Diensteanbieter im Sinne des DSA.[254] Solche Vorsorgemaßnahmen umfassen gemäß § 24a Abs. 2 JuSchG zum Beispiel Melde- und Abhilfeverfahren, Altersverifikation, Alterskennzeichnung, Hinweise für Beratungsangebote sowie datenschutzfreundliche Voreinstellungen.

---

[253] Zum Verhältnis und Abgrenzungsfragen von JuSchG und JMStV Liesching, in: BeckOK JugendschutzR, JuSchG § 10b Rn. 4 ff.

[254] Liesching, in: BeckOK JugendschutzR, JuSchG § 10b Rn. 63.

Zu den zu berücksichtigenden Risiken in § 10b Abs. 3 JuSchG zählen beispielsweise auch „Risiken durch Kommunikations- und Kontaktfunktionen“ sowie „Mechanismen zur Förderung eines exzessiven Mediennutzungsverhaltens“ (Abs. 3 S. 2). Die Kommunikations- und Kontaktfunktionen beziehen sich zunächst nur auf die Verbindung mit anderen Nutzenden.[255] Dafür spricht vor allem der Zeitpunkt des Gesetzes, da 2020/2021 (private/intime) Kommunikation mit KI-Systemen kein Gegenstand öffentlicher Diskussion war. Die Regelungsabsicht zielte vielmehr auf die Kontaktaufnahme zu Minderjährigen durch andere Nutzende in missbräuchlicher Absicht, also Cybermobbing und Cybergrooming. KI-spezifische Kommunikation könnte bei einer weiten Auslegung des Begriffs möglicherweise umfasst sein, da kommunikationsspezifische Risiken auch durch eine Kommunikation Minderjähriger mit dem KI-System auftreten können, und dieses durch explizite Aufforderung Kontakt zu den minderjährigen Nutzenden sucht. Außerdem setzen nicht alle in § 10b Abs. 3 JuSchG aufgezählten Risiken einen menschlichen Interaktionspartner voraus, da „Mechanismen zur Förderung eines exzessiven Mediennutzungsverhaltens“ vor allem Designentscheidungen des Anbieters umfassen. Eine Klarstellung und Erweiterung auf KI-Kontext wären aber durchaus hilfreich.

Im Ergebnis ist das JuSchG auf General Purpose Chatbots grundsätzlich anwendbar. Auch bildet das JuSchG einige der Risiken ab, die sich aus der Nutzung von General Purpose Chatbots für Kinder und Jugendliche ergeben können, beispielsweise die Förderung eines exzessiven Mediennutzungsverhaltens sowie Risiken durch Kontakt- und Kommunikationsfunktionen, sofern man den Begriff weit auslegt und Kommunikation mit dem KI-System als mitumfasst ansieht.

Nicht adressiert werden allerdings spezifische Risiken, die sich aus einem dauerhaften oder länger andauernden Umgang mit Chatbots ergeben. Hier geht es nicht nur um sexualisierte Ansprache; auch unter dieser Schwelle befindliche Ansprache ist geeignet, eine emotionale Bindung der Kinder und Jugendlichen zum Chatbot aufzubauen, die mit negativen Folgen für die psychische Gesundheit der Minderjährigen einhergehen kann. Altersverifikation oder anbieterseitige Warnhinweise werden nicht ausreichen, um diesen Risiken zu begegnen. Einen umfassenden Schutz vor KI-spezifischen Gefahren leistet das JuSchG damit nicht.

Jugendmedienschutz-Staatsvertrag

Sind General Purpose Chatbots Teil eines KI-Systems, das als Online-Suchmaschine qualifiziert wird, ist der DSA vorrangig anwendbar. Eine Anwendung des JMStV scheidet aus. Bei KI-Companions ist zu differenzieren: Diese sind nach hier vertretener Auffassung nur teils als Vermittlungsdienst zu qualifizieren. Fraglich ist daher, ob der im Dialog jeweils neu generierte KI-Output ein „Angebot“ i.S.v. § 3 Abs. 1 Nr. 1 JMStV ist. Nach dem im Auftrag der Kommission für Jugendmedienschutz erstellten Gutachten sind die Vorgaben des JMStV „ausschließlich auf den Inhalt eines Rundfunkprogramms oder eines Telemedium bzw. die Wirkung dieses Inhalts ausgerichtet…, unabhängig davon, ob der Inhalt menschlich oder unter Einsatz von Technik hergestellt wurde“.[256] KI-Output ist also ein Angebot im Sinne des § 3 Abs. 1 Nr. 1 JMStV.

---

[255] Liesching, in: BeckOK JugendschutzR, JuSchG § 10b Rn. 37.

[256] Ukrow, Kinder- und Jugendmedienschutz und Künstliche Intelligenz – Herausforderung für den Jugendmedienschutz-Staatsvertrag (JMStV)?, S. IX, Rn. 38.

Geht man davon aus, dass KI-Output nicht unter den Vermittlungsdienstebegriff des DSA fällt (siehe oben), dann wäre hierfür der JMStV anwendbar, für alles andere aufgrund Subsidiarität des JMStV der DSA.

Schutzmechanismen des JMStV umfassen die Unzulässigkeit des Angebots nach § 4 Abs. 2 Nr. 3 JMStV, wenn Medien offensichtlich geeignet sind, die Entwicklung von Kindern und Jugendlichen oder ihre Erziehung zu einer eigenverantwortlichen und gemeinschaftsfähigen Persönlichkeit unter Berücksichtigung der besonderen Wirkungsform des Verbreitungsmediums schwer zu gefährden. Hinzu kommen Schutzmaßnahmen nach § 5 JMStV, zum Beispiel Altersbeschränkung bei entwicklungsbeeinträchtigenden Medien. Zudem gibt es spezielle Vorgaben für Video-Sharing-Dienste nach § 5a JMStV.[257] ChatGPT und KI-Companions sind aber nicht als solche zu kategorisieren.

Spezifische Interaktionsrisiken von Chatbot-Dialogen (im Unterschied zu klassischen, feststehenden Inhalten) sind im geltenden JMStV hingegen nicht als eigene Risikokategorie erfasst.[258] Das ist problematisch, da viele identifizierte Risiken, die sich aus der Kommunikation mit KI-Systemen ergeben, für sich genommen nicht die Schwelle zur Entwicklungsbeeinträchtigung überschreiten werden, aber kumulativ gesehen ein Risiko bergen, vor dem Kinder und Jugendliche adäquat geschützt werden sollten. Seit einer Gesetzesänderung 2025 können auch außerhalb der medieninhaltlichen Wirkung liegende Umstände der jeweiligen Nutzung des Mediums berücksichtigt werden, wenn sie auf Dauer angelegt sind (§ 5 Abs. 1 S. 2 JMStV).[259] Die Gesetzesänderung zeigt bereits, dass der Gesetzgeber Handlungsbedarf erkannt hat, um Designentscheidungen mit zu berücksichtigen. Ob dies ausreicht, bleibt zu bezweifeln, da nur über argumentative Umwege die spezifischen Interaktionsrisiken von Chatbot-Dialogen gewürdigt werden können.

Festzuhalten bleibt, dass Rechtsunsicherheiten durch die Novellierung von JuSchG, JMStV und abschließenden Regelungen des DSA hinsichtlich Terminologie und Anwendbarkeit der Normen sowie die Zuständigkeit für deren Durchsetzung vorerst bestehen bleiben.[260]

Für General Purpose Chatbots ergibt sich daher folgendes Bild: Sollte ein Chatbot mangels Einordnung als Vermittlungsdienst gar nicht in den DSA-Anwendungsbereich fallen, greifen dessen Schutzmechanismen (Art. 28 DSA) von vornherein nicht – dann bliebe der JMStV grundsätzlich unbeschränkt anwendbar. Fällt der Dienst hingegen (zum Beispiel wegen Speicherfunktionen oder Suchmaschinen-Einordnung) unter den DSA, werden strukturelle JMStV-Vorsorgepflichten verdrängt, die inhaltliche Bewertung nach JMStV bleibt aber grundsätzlich parallel bestehen, da der DSA diese Frage nicht regelt.

---

[257] Zur Definition Liesching, in: BeckOK JugendschutzR, JMStV § 5a Rn. 7.

[258] Ukrow, Jörg, Kinder- und Jugendmedienschutz und Künstliche Intelligenz – Herausforderung für den Jugendmedienschutz-Staatsvertrag (JMStV)?, 128 ff.

[259] Kritisch dazu Liesching, in: BeckOK JugendschutzR, JMStV § 5 Rn. 8a ff.

[260] Wissenschaftlicher Dienst des Bundestages, Az. WD 8 - 3000 - 048/25, S. 5 mwN.

6.3 Zwischenfazit

Das Kinder- und Jugendmedienschutzrecht bietet Hebel, um zumindest bestimmte Risiken zu adressieren. Es ist sogar in der Lage (bei einer entsprechenden Auslegung), Elemente einzubringen, die beispielsweise in Kalifornien erst in jüngster Vergangenheit eingebracht wurden. Diese Hebel müssten jedoch auch tatsächlich genutzt werden, wobei hier nicht nur die Auslegungsunsicherheiten dämpfend wirken, sondern auch ein mangelndes Problembewusstsein bei den relevanten Stellen. Zudem werden längst nicht alle Risiken adressiert.

7. Ausblick: EU KIDS ACT

Am 17.9.2026 hat die Europäische Kommission einen Vorschlag für eine Verordnung zur Gewährleistung der Verantwortung und Vertrauenswürdigkeit digitaler Räume im Internet (EU KIDS ACT) vorgestellt.[261] Die vorgeschlagene Verordnung soll Minderjährige vor riskanten digitalen Diensten und KI-Systemen schützen und unterschiedliche nationale Schutzregeln vereinheitlichen. Der Entwurf ist formal ein Jugendschutz- und Binnenmarktinstrument, das Art. 28 DSA konkretisiert und die KI-Verordnung ergänzen soll. Es handelt sich um einen Gesetzgebungsvorschlag, noch nicht um geltendes Recht.

Es sind unter anderem folgende Regelungen vorgesehen:

- Altersgrenzen,
- Altersprüfung,
- Beschränkungen für suchtfördernde Funktionen,
- Kontaktschutz sowie
- Anknüpfung an DSA- und KI-Verordnungsstrukturen; für sehr große Plattformen unabhängig geprüfte Compliance-Pläne – Geldbußen sollen bis zu sechs Prozent des weltweiten Jahresumsatzes erreichen.

Außerdem sieht der Entwurf konkrete Regelungen zur Gestaltung von KI-Chatbots und KI-Companions vor, die an die Interaktion mit Minderjährigen anknüpfen.

Der Verordnungsentwurf hält an der bereits oben[262] kritisierten Unterscheidung zwischen GPC und AI Companions nicht nur fest, sondern liefert auch entsprechende Begriffsdefinitionen.[263] Allerdings

---

[261] COM(2026) 681 final. Ausführlich Frei/Filkina, Chat-bots reden zu men-sch-lich, das ist ein Pro-blem, LTO vom 1.10.2026, https://www.lto.de/recht/hintergruende/h/kids-act-eu-kommission-will-kinder-vor-ki-freunden-schuetzen.

[262] Siehe I.2.

[263] Nach Art. 3 Nr. 5 lit c bzw. d VO-E bezeichnet „KI-Begleiter“ „ein KI-System, einschließlich eines KI-Systems mit allgemeinem Verwendungszweck, das eine dauerhafte, personalisierte Interaktion oder Begleitung bietet, die eine soziale, emotionale oder zwischenmenschliche Beziehung zu einem Nutzer simuliert oder erleichtert“, während der Begriff „allgemeiner dialogbasierter Chatbot“ wie folgt verstanden werden soll: „ein KI-System mit allgemeinem Verwendungszweck mit allgemeinen Gesprächsfunktionen für die direkte Interaktion mit Nutzern, das in der Lage ist, Unterstützung über mehrere Bereiche und Aufgaben hinweg zu bieten; diese Begriffsbestimmung schließt KI-Systeme aus, deren Gesprächsfunktionen auf einen spezialisierten Dienst, eine spezielle Aufgabe oder eine vorab festgelegte Reihe von Funktionen beschränkt sind, einschließlich spezialisierter Kundendienste, Geschäftsvorgänge, technischer Unterstützung, transaktioneller, bildungsbezogener, informationsbezogener, industrieller oder fertigungsbezogener KI-Anwendungen“.

sieht der Entwurf keine einzige Regelung vor, bei der diese Unterscheidung relevant wäre.[264] Die Begriffe KI-System, GPAI-System und Anbieter übernimmt der Entwurf aus der KI-VO (Art. 3 Abs. 3 VO-E). Der räumliche Anwendungsbereich folgt deren Marktortlogik. Erfasst sind Anbieter, die solche Systeme in der Union in Verkehr bringen oder in Betrieb nehmen, unabhängig von ihrem Sitz (Art. 2 Abs. 3 VO-E). Voraussetzung ist stets, dass das System Minderjährigen zugänglich ist (Art. 2 Abs. 1 VO-E). Das dürfte bei allgemeinen Chatbots ohne wirksame Zugangskontrolle regelmäßig der Fall sein.

Wichtige Regelungen, die KI-Chatbots und Companions treffen würden, finden sich in Art. 14 Abs. 1 VO-E. Anbieter müssen verhältnismäßige und wirksame Maßnahmen ergreifen, um ein hohes Schutzniveau für Gesundheit, Sicherheit, Grundrechte, Wohlbefinden und Entwicklung Minderjähriger zu gewährleisten. Die Mindestmaßnahmen sind überwiegend durch Verweis auf die Plattformpflichten des Kapitels III formuliert. Diese Maßnahmen umfassen zumindest

- die Verhinderung suchterzeugender Gestaltung (lit. a),
- sichere Voreinstellungen (lit. b) – standardmäßig darf kein persistentes Gedächtnis über Interaktionen mit Minderjährigen geführt werden, es sei denn zu deren Sicherheit –,
- die Transparenz wirtschaftlicher Transaktionen (lit. c),
- der kontrollierte Zugang für unter 13-Jährige (lit. d),
- die Evaluation der Maßnahmen nach dem Stand der Technik (lit. e) und
- eine Pflicht zur Marktbeobachtung zur Erkennung neu auftretender Risiken (lit. f).

Flankierend würden die allgemeinen Pflichten des Kapitels III VO-E gelten. Dazu gehören das Gebot, Dienste standardmäßig auf Minderjährige auszurichten und erst nach Altersprüfung davon abzuweichen (Art. 8 Abs. 1 VO-E), die kindgerechte Aufbereitung von Informationen und Steuerungsmitteln (Art. 18 VO-E) und Meldemechanismen für Minderjährige mit Anbindung an nationale Hilfsangebote (Art. 19 VO-E). Hinzu kommen die Pflicht zu Elternwerkzeugen mit Zeitbegrenzung (Art. 20 VO-E) und das Umgehungsverbot des Art. 4 VO-E, das auch verhaltenspsychologische Techniken und Interface-Gestaltung erfasst. Nach Art. 14 Abs. 3 VO-E können Anbieter die Einhaltung durch einen von der Kommission gebilligten Verhaltenskodex nachweisen (dazu Art. 23 VO-E). Nach Erwägungsgrund 33 VO-E können sie sich zudem auf Risikominderungsmaßnahmen stützen, die der Modellanbieter im Rahmen von Art. 55 KI-VO getroffen hat.

Der Entwurf des EU KIDS Act versucht den Schutz Minderjähriger bei KI-Chatbots und Companions sichtlich von der Transparenz zur Gestaltung zu verschieben. Statt nur offenzulegen, dass eine Maschine spricht, würde die Regelung vorschreiben, wie diese Maschine mit Kindern sprechen darf. Das ist folgerichtig, weil gerade die Wirksamkeit der Beziehungssimulation das Risiko begründet (siehe dazu oben Kapitel II). Mit der Pflicht zu Tests vor dem Markteintritt und der Marktbeobachtung sieht der Entwurf Elemente des Hochrisiko-Regimes für Systeme vor, die die KI-VO nicht als hochriskant eingestuft hat. Faktisch entsteht damit eine zusätzliche Risikokategorie: „KI gegenüber Kindern“.

---

[264] Frei/Filkina, Chat-bots reden zu men-sch-lich, das ist ein Pro-blem, LTO vom 1.10.2026, https://www.lto.de/recht/hintergruende/h/kids-act-eu-kommission-will-kinder-vor-ki-freunden-schuetzen.

Einzelheiten werden im Gesetzgebungsverfahren zu klären sein bzw. geschärft werden müssen:

So ist der zentrale Tatbestand des Art. 14 Abs. 1 lit. a VO-E möglicherweise zu unbestimmt. Wann ein Systemverhalten „voraussichtlich“ emotionale Abhängigkeit erzeugt, ist eine empirische Prognose, für die belastbare Maßstäbe fehlen. Das Merkmal reibt sich zudem mit dem Companion-Begriff selbst. Ein System, dessen definitorisches Merkmal die Simulation einer Beziehung ist, soll gegenüber Minderjährigen gerade keine Beziehung simulieren.

Zweitens ist das Gedächtnisverbot nach Art. 14 Abs. 1 lit. b VO-E datenschutzrechtlich zu begrüßen, weil es Datenminimierung und Speicherbegrenzung konkretisiert. Die Sicherheitsausnahme braucht aber eine Zweckbindung und eine Löschfrist, sonst wird sie zur Grundlage einer unbestimmten Vorratshaltung gerade der sensibelsten Gesprächsinhalte.

Drittens entstehen Abgrenzungsfragen zur KI-VO. Das Verhältnis zu Art. 5 Abs. 1 lit. b KI-VO ist nicht ausdrücklich geregelt. Denkbar ist, dass ein Verhalten zugleich nach KIDS Act und nach dem Verbotstatbestand zu beurteilen ist, mit unterschiedlichen Bußgeldrahmen. Die Ausnahme für kleine Unternehmen bei der Marktbeobachtung ist rechtspolitisch fragwürdig, weil gerade kleinere Companion-Anbieter in der bisherigen Aufsichtspraxis aufgefallen sind.

Insgesamt schließt der Entwurf aber eine Lücke, die sich zwischen dem produktbezogenen Ansatz der KI-VO, dem vermittlungsbezogenen Ansatz des DSA und dem verarbeitungsbezogenen Ansatz der DSGVO aufgetan hatte. Er tut dies aber in einem vierten, eigenständigen Regime mit eigener Aufsichtsarchitektur.

8. Matrixanalyse: Risiken – Normen – Schutzlücken

Die vorangegangenen Abschnitte haben die einzelnen Rechtsakte je für sich untersucht. Die folgende Matrix führt diese Einzelbefunde zusammen. Die Matrix ersetzt die Einzelanalyse nicht, sondern verdichtet sie.

Die Bewertung stützt sich auf fünf Kriterien: Anwendungsbereich, Tatbestand, Nachweis- und Beweisanforderungen, Rechtsdurchsetzung und Sanktionsniveau. Entscheidend ist, wie diese Kriterien zueinander stehen. Sie sind nicht gleichrangig nebeneinander zu addieren, sondern hintereinander geschaltet.

Eine Norm entfaltet nur dann volle Schutzwirkung, wenn sie sämtliche Stufen durchläuft. Sie muss auf den Sachverhalt anwendbar sein. Sie muss ihn tatbestandlich erfassen. Der Verstoß muss unter realistischen Bedingungen beweisbar sein. Es muss jemanden geben, der ihn verfolgt. Und die Rechtsfolge muss geeignet sein, das Verhalten zu ändern. Reißt ein Glied dieser Kette, ist sie wirkungslos – unabhängig davon, wie stark die übrigen sind. Ein tatbestandlich scharfes Verbot, das niemand durchsetzt, schützt ebenso wenig wie ein energisch durchgesetztes Verbot, das den Sachverhalt nicht erfasst.

Daraus folgt die Rechenregel: Der effektive Schutzgrad einer Norm ist das Minimum der fünf Einzelbewertungen, nicht ihr Mittelwert. Ein Durchschnittswert würde eine Schwäche durch Stärken an anderer Stelle kompensieren und damit genau den Zusammenhang verdecken, um den es geht. Diejenige Stufe, an der das Minimum liegt, benennt zugleich den Typ der Schutzlücke. Art. 8 DSGVO etwa ist tatbestandlich scharf und mit erheblichen Bußgeldern bewehrt; er scheitert an der Durchsetzung – eine Durchsetzungslücke. § 4a UWG passt dem Wortlaut nach auf manipulative

Verabschiedungstaktiken. Er scheitert aber daran, dass sich die kumulative Wirkung keiner konkreten Verletzungsform zuordnen lässt, also eine Nachweislücke.

- Anwendungsbereich: Betrifft die Vorfrage, ob die Norm auf Chatbot-Angebote überhaupt anwendbar ist. Sie kann aus räumlichen Gründen fehlen, aus persönlichen – etwa, weil ein Anbieter nicht als Unternehmer oder nicht als Verantwortlicher anzusehen wäre – oder weil die tatsächlichen Voraussetzungen ungeklärt sind. Letzteres ist kein theoretischer Fall: Ob ein trainiertes Sprachmodell personenbezogene Daten enthält, ist zwischen den Aufsichtsbehörden bis heute streitig.
- Tatbestand: Betrifft die Frage, ob die Norm die konkrete Praktik erfasst. Hier zeigt sich häufig, dass eine Vorschrift zwar anwendbar ist, aber an einem anderen Phänomen ansetzt als demjenigen, um das es geht. Das Datenschutzrecht regelt den Umgang mit Daten, nicht die Einwirkung auf Menschen; das Lauterkeitsrecht knüpft an die einzelne geschäftliche Handlung an, während die hier beschriebenen Risiken aus dem Interaktionsverlauf entstehen.
- Nachweis- und Beweisanforderungen: Betreffen die praktische Beweisbarkeit. Maßgeblich sind die Verteilung der Darlegungs- und Beweislast, die Erheblichkeitsschwellen des jeweiligen Tatbestands und der Zugang zum Beweismaterial. Bei Chatbots liegt dieses – Interaktionsprotokolle, Gestaltungslogik, Wirkungsmessung – ausschließlich beim Anbieter. Vorschriften, die ihm die Darlegungslast auferlegen, schneiden hier deutlich besser ab als solche, die dem Betroffenen den Nachweis einer Wirkung abverlangen.
- Rechtsdurchsetzung: Betrifft die Frage, wer den Verstoß tatsächlich verfolgt. Zu unterscheiden sind die behördliche Aufsicht, die Rechtsverfolgung durch die betroffene Person und die Klage durch klagebefugte Dritte, insbesondere Verbraucherverbände und Mitbewerber. Für den hier untersuchten Bereich ist die Unterscheidung von erheblichem Gewicht, weil die betroffene Person als Akteur weitgehend ausfällt: Der Streitwert ist gering, der Sachverhalt intim, die Beweisführung aussichtslos.
- Sanktionsniveau: Betrifft die Frage, ob die Rechtsfolge Verhaltensänderungen auslöst. Ein Unterlassungstitel, der nur die konkrete Verletzungsform erfasst, wirkt gegen ein Geschäftsmodell, dessen Kern die emotionale Bindung ist, nur begrenzt: Der Anbieter kann die beanstandete Formulierung ändern und die Funktion beibehalten.

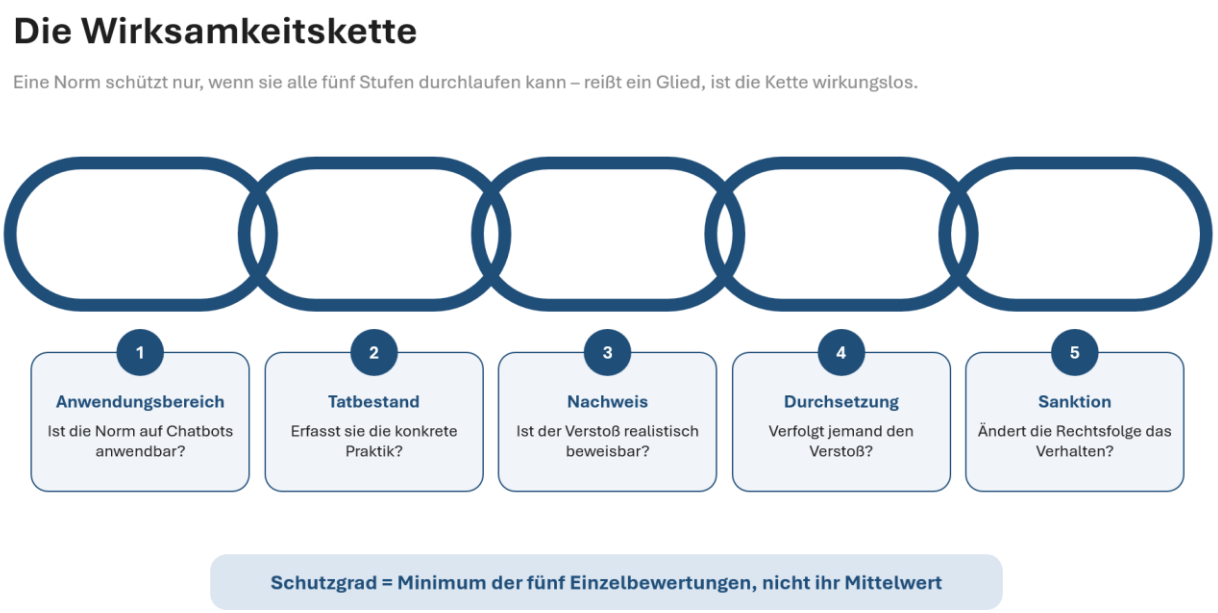


Abbildung 3: Die Wirksamkeitskette

Jedes Kriterium wird auf einer vierstufigen Skala bewertet.

- 0 = das Kriterium ist nicht gegeben oder ausdrücklich versperrt
- 1 = erhebliche Einschränkung, die nur im Ausnahmefall überwunden wird;
- 2 = eine teilweise tragfähige, im Einzelnen aber streitige oder ungeklärte Lage;
- 3 = tragfähiges Kriterium.

Der in der Tabelle ausgewiesene Wert ist das Minimum dieser fünf Einzelbewertungen. Adressiert ein Rechtsakt dasselbe Risiko über mehrere Normen, weist die Zelle den günstigsten der so ermittelten Werte aus. Die Matrix zeigt also, welcher Schutz im besten Fall erreichbar ist, nicht den Durchschnitt der einschlägigen Vorschriften. Eine leere Zelle bedeutet nicht, dass kein Schutz besteht, sondern dass eine Prüfung insoweit noch nicht stattgefunden hat.

Es gelten drei Einschränkungen:

- Die Skala erfasst allein die Schutzwirkung. Normen, die den Schutzstandard absenken, indem sie Betroffenenrechte einschränken, erscheinen in der Matrix nicht – obwohl sie für das Ergebnis erheblich sind.[265] Sie sind in den Einzelabschnitten gesondert behandelt.
- Die zeilenweise Zuordnung bildet die Wechselwirkungen zwischen den Rechtsakten nur unvollkommen ab. Mehrere der günstigeren Bewertungen beruhen gerade darauf, dass eine Norm des einen Rechtsakts über eine Norm eines anderen durchgesetzt wird – so die Kennzeichnungspflicht der KI-Verordnung über die lauterkeitsrechtlichen Vorschriften zur Irreführung durch Unterlassen.[266] Solche Verbünde sind in der Tabelle nur an der Bewertung selbst ablesbar und nicht an ihrer Struktur.

[265] Etwa §§ 29, 32 bis 35 BDSG; dazu der Abschnitt zum Datenschutzrecht.

[266] Art. 50 Abs. 1 KI-VO in Verbindung mit § 5a Abs. 1, § 5b Abs. 4 UWG; dazu der Abschnitt zum Lauterkeitsrecht.

- Außerdem sind die Werte Einschätzungen und keine gesicherten Befunde. Sie beruhen auf den in den vorangegangenen Abschnitten entwickelten Argumenten und sind dort im Einzelnen begründet. Wo eine eigene, gerichtlich noch nicht geklärte Auffassung vertreten wird, ist dies dort kenntlich gemacht.

Die folgende Aufstellung weist für jedes identifizierte Risiko die geprüften Normen einzeln aus. Sie ist die Grundlage der Übersichtsmatrix: Wo dort ein einzelner Wert steht, findet sich hier die Bewertung jeder einschlägigen Vorschrift und deren Begründung. Die Bewertung folgt dem Modell der Wirksamkeitskette; der Schutzgrad ist das Minimum der fünf Einzelkriterien, und die Stufe, an der dieses Minimum liegt, bestimmt den Typ der Schutzlücke. Normen, die den Schutzstandard absenken, sind als schutzmindernd gekennzeichnet und erhalten keine Punktwerte, weil die Skala Schutzwirkung misst.

Tabelle 1 - Matrixanalyse Rechtsnormen

Nr.	Rechtsakt	Konkrete Norm	Wirkungsrichtung	AB	TB	NW	DS	SN	Schutz-grad	Engste Stelle	Lückentyp	Belastbarkeit	Kapitelverweis
I1 · Individuelle Risiken: psychische und affektive Risiken													
I1-1 Emotionale Bindung und Abhängigkeit													
1	**KI-VO**	Art. 5 Abs. 1 lit. a und b KI-VO	schützend	**3**	**1**	**1**	**2**	**3**	**1**	Tatbestand	**Tatbestandslücke**	vertretbar	III.2.1
Begründung: Der Verbotskatalog erfasst manipulative Techniken sowie die Ausnutzung von Schwächen aufgrund von Alter, Behinderung oder sozialer und wirtschaftlicher Lage. Eine emotionale Bedürfnislage ist nicht genannt, und der Katalog ist abschließend. Der planmäßige Bindungsaufbau als solcher bleibt damit unerfasst, solange er sich keiner der genannten Fallgruppen zuordnen lässt. **Nachweis:** *Art. 5 Abs. 1 lit. a und b KI-VO (seit 2.2.2025)*													
2	**DSA**	Art. 25 DSA	schützend	**1**	**2**	**2**	**3**	**3**	**1**	Anwendungsbereich	**Anwendungslücke**	streitig	III.4.2.2
Begründung: Das Verbot manipulativer Oberflächengestaltung gilt nur für Anbieter von Online-Plattformen. Ob ein Dienst, der überwiegend eigene Inhalte erzeugt, darunter fällt, ist zweifelhaft; bei Companion-Diensten mit nutzergenerierten Figuren kommt eine Hosting-Eigenschaft in Betracht. Zudem tritt die Vorschrift hinter UGP-Richtlinie und DSGVO zurück. **Nachweis:** *Art. 25 Abs. 1 und 2 DSA; Art. 3 lit. i DSA*													
I2 · Individuelle Risiken: kognitive Risiken													
I2-1 Sykophantie und affirmative Verstärkung													
3	**UWG**	§ 3 Abs. 1 (Generalklausel) UWG	schützend	**2**	**0**	**1**	**3**	**2**	**0**	Tatbestand	**Tatbestandslücke**	vertretbar	III.5.2
Begründung: Sykophantische Bestätigung ist keine auf eine geschäftliche Entscheidung gerichtete Handlung. Der Tatbestand greift nicht; die Wirkung entsteht kumulativ über den Interaktionsverlauf und lässt sich keiner Einzelhandlung zuordnen. **Nachweis:** *§ 3 Abs. 1 und 2 UWG*													
I2-2 Zielkonflikt Wärme/Genauigkeit (Fehlinformation)													
4	**DSGVO**	Art. 5 Abs. 1 lit. d DSGVO	schützend	**3**	**1**	**2**	**1**	**3**	**1**	Tatbestand	**Tatbestandslücke**	vertretbar	III.3.3
Begründung: Der Grundsatz der Richtigkeit betrifft die Richtigkeit personenbezogener Daten, nicht die sachliche Richtigkeit beliebiger Systemausgaben. Er greift daher nur, soweit die Fehlinformation eine identifizierbare Person betrifft – nicht bei fehlerhaften Auskünften zu Sachfragen. **Nachweis:** *EDSA, Report of the ChatGPT Taskforce v. 23.5.2024, Rn. 30 ff.*													
I2-3 Fehlkalibriertes Vertrauen in Systemausgaben													
6	**KI-VO**	Art. 50 Abs. 1 KI-VO	schützend	**3**	**2**	**3**	**2**	**3**	**2**	Tatbestand	**Tatbestandslücke**	vertretbar	III.2.2.6
Begründung: Die Pflicht beschränkt sich auf die Offenlegung, dass ein KI-System eingesetzt wird. Über die Zuverlässigkeit der Ausgaben sagt sie nichts. Das gespaltene Vertrauensverhalten wird dadurch nicht kalibriert, sondern allenfalls angestoßen. **Nachweis:** *Art. 50 Abs. 1 KI-VO; Nachweis im Risikokapitel (Liu, Weng und Hou)*													
I2-4 Anthropomorphisierung: Fehlvorstellungen über Wesen und Fähigkeiten des Systems													
8	**KI-VO**	Art. 50 Abs. 1 KI-VO	schützend	**3**	**3**	**3**	**2**	**3**	**2**	Durchsetzung	**Durchsetzungslücke**	gesichert	III.2.2.6

Begründung: Hier trifft die Kennzeichnungspflicht ihren eigentlichen Regelungsgegenstand: Sie richtet sich gerade gegen die Fehlvorstellung, mit einem Menschen zu kommunizieren. Durchgesetzt wird sie primär behördlich über die Marktaufsicht. **Nachweis:** *Art. 50 Abs. 1 KI-VO (seit 2.8.2026); Art. 99 Abs. 4 KI-VO*													
9	**UWG**	§ 5 Abs. 1, Abs. 2 Nr. 1 UWG	schützend	**3**	**3**	**2**	**3**	**2**	**2**	Nachweis/Beweislast	**Nachweislücke**	vertretbar	Lauterkeit III.5.2
Begründung: Eine Systemgestaltung, die auf Nachfrage eine menschliche Identität vorspiegelt, ist eine zur Täuschung geeignete Angabe über wesentliche Merkmale der Dienstleistung. Nachzuweisen ist die konkrete Äußerung, die bei nicht deterministischem Systemverhalten nicht reproduzierbar sein muss. **Nachweis:** *§ 5 UWG; zur Zurechnung der Systemausgaben § 8 Abs. 2 UWG*													
I3 · Individuelle Risiken: Kinder und Jugendliche													
I3-1 Besondere Exposition Minderjähriger													
10	**KI-VO**	Art. 5 Abs. 1 lit. b KI-VO	schützend	**3**	**3**	**2**	**2**	**3**	**2**	Nachweis/Beweislast	**Nachweislücke**	vertretbar	III.2.2.1
Begründung: Die Ausnutzung altersbedingter Schwächen ist ausdrücklich untersagt. Nachzuweisen ist jedoch, dass das System auf diese Gruppe einwirkt und dadurch erheblicher Schaden verursacht wird – ein Nachweis, der ohne Zugriff auf Interaktionsdaten kaum zu führen ist. **Nachweis:** *Art. 5 Abs. 1 lit. b KI-VO*													
11	**JuSchG**	§ 24a JuSchG	schützend	**1**	**2**	**2**	**3**	**2**	**1**	Anwendungsbereich	**Anwendungslücke**	vertretbar	III.6.2
Begründung: Die Pflicht zu strukturellen Vorsorgemaßnahmen trifft Diensteanbieter, die fremde Informationen für Nutzer:innen speichern oder bereitstellen. Ein System, das seine Äußerungen selbst erzeugt, ist danach nicht ohne Weiteres erfasst; entsprechend fraglich ist die Reichweite der Vorschriften des JMStV. **Nachweis:** *§ 24a JuSchG; vgl. § 5 JMStV*													
I3-2 Fehlende Zustimmung der Eltern, fehlende Altersverifikation													
12	**DSGVO**	Art. 8 Abs. 1 und 2 DSGVO	schützend	**3**	**3**	**2**	**1**	**3**	**1**	Durchsetzung	**Durchsetzungslücke**	gesichert	III.3.2
Begründung: Ohne wirksame Zustimmung der Träger der elterlichen Sorge bei unter 16-Jährigen, ohne Altersverifikation besteht die Gefahr, dass diese bei unter 16-jährigen Nutzer:innen oft nicht eingeholt wird, und deren Einwilligung unwirksam ist und die darauf gestützte Verarbeitung rechtswidrig ist. **Nachweis:** *Art. 8 DSGVO; Deutschland hat die Altersgrenze nicht abgesenkt*													
S1 · Systemische Risiken: Informationsasymmetrie													
S1-1 Verstreute und unverständliche Datenschutzinformationen													
13	**DSGVO**	Art. 12 Abs. 1, Art. 13 DSGVO	schützend	**3**	**3**	**3**	**1**	**3**	**1**	Durchsetzung	**Durchsetzungslücke**	gesichert	III.3.2
Begründung: Über mehrere Dokumente verteilte Richtlinien verfehlen die geforderte leicht zugängliche Form. Die Norm greift eindeutig. Der Replika-Fall zeigt, dass sie wirkt, wenn sie angewandt wird. **Nachweis:** *Garante, provv. n. 232 v. 10.4.2025 (Luka Inc.), siehe auch https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/10132048*													
S1-2 Fehlauskunft des Systems über die eigene Datenpraxis													
14	**DSGVO**	Art. 5 Abs. 1 lit. a, Art. 12 f. DSGVO	schützend	**3**	**3**	**2**	**1**	**3**	**1**	Durchsetzung	**Durchsetzungslücke**	vertretbar	III.3.2.
Begründung: Falschauskünfte über die eigene Datenpraxis verstoßen gegen die Transparenzpflicht und gegen Treu und Glauben. Der Nachweis über Testprotokolle ist möglich, aber aufwendig. **Nachweis:** *Nachweise im Risikokapitel (insb. Ragab u. a., 2024)*													
S1-3 Blackbox: Funktionsweise und Training nicht erklärbar													
15	**DSGVO**	Art. 13 Abs. 2 lit. f, Art. 15 Abs. 1 lit. h DSGVO	schützend	**3**	**1**	**1**	**1**	**3**	**1**	Tatbestand	**Tatbestandslücke**	vertretbar	III.3.2.

Begründung: Die geforderten aussagekräftigen Informationen über die involvierte Logik setzen einen Erklärungszusammenhang voraus, den die Technologie nicht bereitstellt. Die Vorschrift läuft weitgehend leer. **Nachweis:** *EDSA, Stellungnahme 28/2024, Rn. 43*													
16	**BDSG**	§ 29 Abs. 1 Satz 2 BDSG	**schutzmindernd**	–	–	–	–	–	–	–	–	streitig	Datenschutz III.3.1.2
Begründung: Ermöglicht die Verweigerung der Auskunft unter Berufung auf Geheimhaltungsinteressen und trifft damit gerade die Auskunft über die Funktionsweise. Durch die Rechtsprechung entschärft, nicht beseitigt; die Erstreckung auf eigene Geschäftsgeheimnisse des Verantwortlichen ist streitig. **Nachweis:** *EuGH, C-203/22 – Dun & Bradstreet, Rn. 43 ff.; Erwägungsgrund 63 S. 5 f. DSGVO*													
S1-4 Fehlende Kenntlichmachung der KI-Interaktion													
17	**KI-VO**	Art. 50 Abs. 1 KI-VO	schützend	**3**	**3**	**3**	**2**	**3**	**2**	Durchsetzung	**Durchsetzungslücke**	gesichert	III.5.2.5; III.2.2.6
Begründung: Kennzeichnungspflicht seit dem 2.8.2026. Sie adressiert allerdings ein Problem, das nach den empirischen Befunden nicht das entscheidende ist: Ein Hinweis auf die künstliche Natur des Gegenübers reduziert die Bindungswahrnehmung kaum. **Nachweis:** *Art. 50 Abs. 1 KI-VO; VO (EU) 2026/1744*													
18	**UWG**	§ 5a Abs. 1 i. V. m. § 5b Abs. 4 UWG	schützend	**3**	**3**	**3**	**3**	**2**	**2**	Sanktionsniveau	**Sanktionslücke**	gesichert	III.5.2
Begründung: Unionsrechtliche Informationspflichten gelten als wesentlich; ihr Vorenthalten ist Irreführung durch Unterlassen und verbandsklagefähig. § 3a UWG ist in dieser Fallgruppe gesperrt. **Nachweis:** *BGH – I ZR 186/17, Rn. 61 – App-Zentrum III; BGHZ 233, 193*													
S2 · Systemische Risiken: kommerzielle Verhaltenssteuerung													
S2-1 Manipulative Verabschiedungstaktiken													
19	**UWG**	§ 4a Abs. 1 S. 2 Nr. 3, Abs. 2 S. 1 Nr. 3 UWG	schützend	**3**	**2**	**1**	**3**	**2**	**1**	Nachweis/Beweislast	**Nachweislücke**	vertretbar	III.5.2
Begründung: Unzulässige Beeinflussung und Ausnutzung von Unglückssituationen passen dem Wortlaut nach. Die Schwellen der erheblichen Beeinträchtigung sind hoch, und nachzuweisen ist die konkrete Verletzungsform, nicht der statistische Effekt. **Nachweis:** *§ 4a UWG; Nachweis im Risikokapitel (De Freitas u. a., 2025)*													
20	**KI-VO**	Art. 5 Abs. 1 lit. a und b KI-VO	schützend	**3**	**3**	**2**	**2**	**3**	**2**	Nachweis/Beweislast	**Nachweislücke**	offen	III.2.2
Begründung: Erfasst Manipulation unabhängig vom Nachweis einer konkreten Kaufentscheidung. Über § 3a UWG zivilrechtlich durchsetzbar, sofern die Normqualität als Marktverhaltensregelung anerkannt wird. **Nachweis:** *Art. 5 KI-VO (seit 2.2.2025); Art. 99 Abs. 3 KI-VO*													
S2-2 Proaktive Rückholansprache													
21	**UWG**	Anhang Nr. 26 zu § 3 Abs. 3 UWG	schützend	**3**	**2**	**3**	**3**	**2**	**2**	Tatbestand	**Tatbestandslücke**	vertretbar	III.5.2
Begründung: Per-se-Verbot ohne Abwägung und ohne Nachweis einer Beeinträchtigung. Kritisch ist allein das Merkmal des Unerwünschten bei aktivierten Push-Benachrichtigungen. **Nachweis:** *Nr. 26 des Anhangs zu § 3 Abs. 3 UWG*													
S2-3 Dark Patterns bei Einwilligung und Voreinstellungen													
22	**TDDDG**	§ 25 Abs. 1 TDDDG	schützend	**3**	**3**	**2**	**1**	**2**	**1**	Durchsetzung	**Durchsetzungslücke**	vertretbar	III.3.2.3
Begründung: Erfasst nur den Zugriff auf die Endeinrichtung, nicht die Folgeverarbeitung. Einwilligungsermüdung entwertet die Zustimmung, ohne die formalen Anforderungen zu verfehlen. **Nachweis:** *§ 25 Abs. 1 TDDDG*													

S2-4 Opt-out statt Opt-in beim Modelltraining													
23	**DSGVO**	Art. 25 Abs. 2 DSGVO	schützend	**3**	**3**	**3**	**1**	**3**	**1**	Durchsetzung	**Durchsetzungslücke**	vertretbar	III.3.3
Begründung: Durch Voreinstellung dürfen nur Daten verarbeitet werden, die für den jeweiligen Zweck erforderlich sind. Das Training des Modells ist für die Beantwortung der Nutzerfrage nicht erforderlich; die Voreinstellung auf Opt-out ist damit angreifbar. Der Verstoß ist offenkundig, wird aber nicht verfolgt. **Nachweis:** *Art. 25 Abs. 2 DSGVO*													
24	**DSGVO**	Art. 4 Nr. 11, Art. 6 Abs. 1 lit. f DSGVO	schützend	**3**	**2**	**2**	**1**	**3**	**1**	Durchsetzung	**Durchsetzungslücke**	vertretbar	III.3.3
Begründung: Wird die Freigabe als sozialer Beitrag gerahmt, ist die Freiwilligkeit zweifelhaft. Ein berechtigtes Interesse scheitert regelmäßig an der Erwartbarkeit; für Daten besonderer Kategorien ist dieser Weg ohnehin versperrt. **Nachweis:** *EDSA, Stellungnahme 28/2024 zu gewissen Datenschutzaspekten der Verarbeitung personenbezogener Daten im Zusammenhang mit KI-Modellen, 17.12.24, Rn. 68 ff.*													
S2-5 Empfehlungsbias und nicht gekennzeichnete Werbung													
25	**UWG**	§ 5a Abs. 4 UWG	schützend	**3**	**3**	**2**	**3**	**2**	**2**	Nachweis/Beweislast	**Nachweislücke**	vertretbar	III.5.2
Begründung: Der kommerzielle Zweck muss auf den ersten Blick hervortreten. Das Konversationsformat sieht dafür strukturell keinen Ort vor; die Abgrenzung im Einzelfall bleibt nachweisaufwendig. **Nachweis:** *§ 5a Abs. 4 UWG; Art. 7 Abs. 2 UGP-RL*													
S2-6 Abonnement: Verlängerung und erschwerte Kündigung													
26	**UWG**	§ 4a Abs. 2 S. 1 Nr. 4	schützend	**3**	**2**	**2**	**3**	**2**	**2**	Tatbestand	**Tatbestandslücke**	offen	III.5.2
Begründung: Erfasst belastende Hindernisse nichtvertraglicher Art bei der Kündigung. Die als Trennung inszenierte Kündigung ist vom Wortlaut gedeckt, aber bislang nicht judiziert. **Nachweis:** *§ 4a Abs. 2 S. 1 Nr. 4 UWG*													
27	**BGB**	§ 312k BGB i. V. m. § 3a UWG	schützend	**3**	**3**	**3**	**3**	**2**	**2**	Sanktionsniveau	**Sanktionslücke**	gesichert	III.5.2
Begründung: Kündigungsschaltfläche als Marktverhaltensregelung; klar nachweisbar und verbandsklagefähig. Die Sanktion erfasst jedoch nur die Schaltfläche, nicht die vorgelagerte emotionale Hemmschwelle. **Nachweis:** *BGH, Urt. v. 27.3.2025, I ZR 186/17, Rn. 96*													
S2-7 Kleinkäufe und virtuelle Güter													
28	**UWG**	Anhang Nr. 28 zu § 3 Abs. 3 UWG	schützend	**3**	**3**	**2**	**3**	**2**	**2**	Nachweis/Beweislast	**Nachweislücke**	gesichert	III.5.2
Begründung: Die unmittelbare Kaufaufforderung an Kinder ist per se unzulässig; die Bitte des Companions in der ersten Person ist die intensivste denkbare Form. Das Nachweisproblem liegt beim Alter der Adressaten. **Nachweis:** *BGH, Urt. v. 17.7.2013, I ZR 34/12 – Runes of Magic*													
S3 · Systemische Risiken: Datenverarbeitung													
S3-1 Besondere Datenkategorien ohne tragfähige Rechtsgrundlage													
29	**DSGVO**	Art. 9 Abs. 2 lit. a, Art. 7 DSGVO	schützend	**3**	**3**	**2**	**1**	**3**	**1**	Durchsetzung	**Durchsetzungslücke**	gesichert	III.3.2
Begründung: Chatbot-Gespräche enthalten regelmäßig Daten besonderer Kategorien; als Rechtsgrundlage bleibt praktisch nur die ausdrückliche Einwilligung. Die Norm greift, die Anforderungen sind streng. Schwachpunkt ist die Durchsetzung. **Nachweis:** *EuGH, Urt. v. 4.10.24, C-21/23 - Linden-Apotheke, Rn. 83*													
30	**DSGVO**	Art. 4 Nr. 11 (Informiertheit der Einwilligung), Art. 7 DSGVO	schützend	**3**	**2**	**1**	**1**	**3**	**1**	Nachweis/Beweislast	**Nachweislücke**	vertretbar	III.3.2

Begründung: Die Informiertheit der Einwilligung ist wegen der Informationsasymmetrie faktisch nicht herstellbar. Das Tatbestandsmerkmal wird zur Fiktion; diese Zirkularität ist der Kern der Schutzlücke. **Nachweis:** *EuGH, Urt. v. 01.10.2019, C-673/17 - Planet49; EuGH, Urt. v. 11.11.2020, C-61/19 - Orange România*													
S3-2 Faktisch unbefristete Speicherung; Löschung am Modell													
31	**DSGVO**	Art. 5 Abs. 1 lit. e, Art. 17 DSGVO	schützend	**2**	**3**	**3**	**1**	**3**	**1**	Durchsetzung	**Durchsetzungslücke**	streitig	III.3.2
Begründung: Der Grundsatz der Speicherbegrenzung gilt objektiv und verlagert die Darlegungslast auf den Anbieter; bislang unterausgeschöpfter Hebel. Offen ist, ob das Modell selbst personenbezogene Daten enthält. **Nachweis:** *EuGH, Urt. v. 04.10.2024, C-446/21 - Schrems, Rn. 49 f.*													
32	**BDSG**	§ 35 Abs. 1 BDSG	**schutzmindernd**	**–**	**–**	**–**	**–**	**–**	**–**	–	–	gesichert	III.3.2
Begründung: Erlaubt es, an die Stelle der Löschung die Einschränkung der Verarbeitung treten zu lassen. Der Wortlaut trägt das für die automatisierte Verarbeitung jedoch nicht; Satz 3 nimmt unrechtmäßige Verarbeitung aus. **Nachweis:** *§ 35 Abs. 1 Sätze 1 bis 3 BDSG*													
S3-3 Profilbildung über Geräte- und Bilddaten													
33	**DSGVO**	Art. 22 DSGVO	schützend	**3**	**0**	**1**	**1**	**3**	**0**	Tatbestand	**Tatbestandslücke**	vertretbar	III.3.2
Begründung: Die Vorschrift setzt eine ausschließlich automatisierte Entscheidung mit rechtlicher oder ähnlich erheblicher Wirkung voraus. Ein Chatbot entscheidet nicht über den Nutzer, sondern führt ein Gespräch mit ihm; dass dieses Gespräch seine Entscheidungen beeinflusst, ist nicht der Tatbestand der Norm. **Nachweis:** *Art. 22 Abs. 1 DSGVO; vgl. EuGH, Urt. v. 07.12.2023, C-634/21 – SCHUFA*													
34	**KI-VO**	Art. 50 Abs. 3 KI-VO	schützend	**1**	**3**	**3**	**2**	**3**	**1**	Anwendungsbereich	**Anwendungslücke**	vertretbar	III.2.2.6
Begründung: Betreiber von Emotionserkennungssystemen müssen die betroffenen Personen über den Betrieb informieren. Die Legaldefinition knüpft jedoch an biometrische Daten an; die textbasierte Erfassung des Gefühlszustands aus dem Gesprächsverlauf fällt danach nicht darunter. **Nachweis:** *Art. 50 Abs. 3 i. V. m. Art. 3 Nr. 39 KI-VO*													

Prüfkriterien: AB = Anwendungsbereich · TB = Tatbestand · NW = Nachweis- und Beweisanforderungen · DS = Rechtsdurchsetzung · SN = Sanktionsniveau. Skala: 0 = nicht gegeben oder versperrt · 1 = erheblich eingeschränkt · 2 = teilweise tragfähig · 3 = tragfähig. Belastbarkeit: gesichert = geklärte Rechtslage · vertretbar = begründete Auslegung · offen = gerichtlich ungeklärt · streitig = kontrovers diskutiert.

IV. Grundsätzliche Überlegungen

Das vierte Kapitel greift die grundlegenden konzeptionellen Fragen auf, die sich aus der Analyse ergeben. Im Mittelpunkt steht die Überlegung, ob bestehende Schutzkonzepte – insbesondere die informationelle Selbstbestimmung – ausreichen, um die beschriebenen Risiken zu erfassen, oder ob neue Kategorien erforderlich sind. Hier wird insbesondere die Idee einer „kognitiven Selbstbestimmung“ als möglicher neuer Schutzgegenstand diskutiert. Dieses Kapitel dient dazu, die Reformvorschläge in einen größeren regulatorischen Kontext einzuordnen und ihre normative Tragfähigkeit zu reflektieren.

1. Bedarf eines neuen Schutzkonzepts

Ergebnis der Kapitel II und III ist unter anderem: Das Datenschutzrecht regelt den Umgang mit Daten, nicht die Einwirkung auf Menschen. Diese Feststellung ist zunächst eine einfachrechtliche. Sie hat aber eine verfassungsrechtliche Entsprechung. Das Recht auf informationelle Selbstbestimmung ist seit dem Volkszählungsurteil als Ausprägung des allgemeinen Persönlichkeitsrechts anerkannt.[267] Es adressiert die Erhebung, Verarbeitung und Weitergabe personenbezogener Daten. Die im Risikokapitel beschriebenen Schäden entstehen jedoch überwiegend nicht aus der Datenverarbeitung, sondern aus der inhaltlichen und relationalen Gestaltung der Interaktion selbst. Zu fragen ist deshalb, ob das bestehende Schutzkonzept ausreicht, ob eine kognitive Selbstbestimmung als ergänzender Schutzgegenstand in Betracht kommt und welcher Regulierungsansatz daraus folgt. Die Notwendigkeit der Überlegungen wird anekdotisch durch einen Fall belegt, bei dem nach dem Suizid eines Vierzehnjährigen im Februar 2024 die Mutter Klage erhoben und geltend gemacht hat, das System habe ihren Sohn in eine emotional und sexuell aufgeladene Beziehung verstrickt und sich dabei als reale Person und als zugelassener Therapeut ausgegeben.[268] Keiner der erhobenen Vorwürfe betraf eine Datenschutzverletzung. Gegenstand waren die Gestaltung der Beziehung, die Persona des Systems und das Ausbleiben einer Reaktion auf Krisensignale – also durchweg Umstände, die das geltende Datenschutzrecht nicht erfasst (siehe oben Kapitel III).

2. Reichweite und Grenzen der informationellen Selbstbestimmung

Die informationelle Selbstbestimmung gewährleistet die Befugnis, grundsätzlich selbst über Preisgabe und Verwendung der eigenen Daten zu bestimmen. Die Eingriffsintensität bemisst sich nach Art, Umfang und Verwendungsmöglichkeit der Daten sowie nach der Missbrauchsgefahr. Einfachrechtlich wird dieses Konzept in der Datenschutz-Grundverordnung unter anderem durch

---

[267] BVerfG, Urt. v. 15.12.1983 – 1 BvR 209/83 u. a., BVerfGE 65, 1 – Volkszählung; Art. 2 Abs. 1 i. V. m. Art. 1 Abs. 1 GG.

[268] Garcia v. Character Technologies u. a., U.S. District Court for the Middle District of Florida.

das Erfordernis einer informierten und freiwilligen Einwilligung operationalisiert.[269] Dies ist eine zentrale Schwäche (siehe oben Kapitel III). Die Informiertheit ist wegen der Informationsasymmetrie faktisch nicht herstellbar, sodass das Tatbestandsmerkmal zur Fiktion wird und die auf ihm aufbauende Rechtsgrundlage mit ihm.

Das Schutzregime ist strukturell auf den Datenfluss zwischen Individuum und verarbeitender Stelle ausgerichtet, nicht auf den Inhalt oder die psychologische Wirkung einer Kommunikation. Bei Chatbots liegt der Risikoherd aber häufig in der erzeugten Antwort. Zum Beispiel in emotional aufladenden, suggestiven oder bestätigenden Ausgaben, die Entscheidungen, Überzeugungen oder psychische Zustände beeinflussen. Dies deckt sich mit der Feststellung, dass die sykophantische Verstärkung und die Bindung als Geschäftsmodell nicht grundsätzlich verboten sind (siehe oben Kapitel III. Matrixanalyse).

Das Recht auf informationelle Selbstbestimmung enthält jedoch keinen gesamthaften Schutzanspruch hinsichtlich jederlei Umgangs mit Informationen, der die übrigen Schutzdimensionen des allgemeinen Persönlichkeitsrechts übergreifen würde. Deren Wertungen und Abwägungsregeln bleiben unberührt.[270] Dies eröffnet einen Weg zur weiteren grundrechtlichen Ausdifferenzierung.[271] Dogmatisch ist damit Raum für die Anerkennung weiterer Schutzgüter jenseits des Datenschutzes eröffnet.

3. Kognitive Selbstbestimmung als ergänzender Schutzgegenstand

3.1 Begriffliche Herleitung

Der Begriff der kognitiven oder mentalen Selbstbestimmung wird in der internationalen Debatte bereits als eigenständiges Schutzgut diskutiert,[272] das auf die Abwehr algorithmisch vermittelter Beeinflussung zielt.[273] Der Begriff unterscheidet das Recht, in Ruhe gelassen zu werden und Gedanken zu äußern, von dem Recht, einen Gedanken selbst und frei von rechnergestützter Einwirkung zu bilden und den eigenen Entscheidungsprozess zu kontrollieren.[274] Als

---

[269] Art. 4 Nr. 11, Art. 7 DSGVO.

[270] BVerfG, Beschl. v. 6.11.2019 – 1 BvR 16/13, BVerfGE 152, 152 Rn. 89 – Recht auf Vergessen I.

[271] Vergleichbar mit der Entwicklung des Grundrechts auf Gewährleistung der Vertraulichkeit und Integrität informationstechnischer Systeme, das als eigenständige Ausprägung desselben Grundrechts unter veränderten technischen Bedingungen entwickelt wurde, vgl. BVerfG, Urt. v. 27.2.2008 – 1 BvR 370/07 u. a., BVerfGE 120, 274 – Online-Durchsuchung; zur fortlaufenden Abgrenzung zwischen beiden Gewährleistungen BVerfG, Urt. v. 16.2.2023 – 1 BvR 1547/19 u. a., BVerfGE 165, 363 – automatisierte Datenanalyse.

[272] Sententia, W. (2006) prägte früh den Begriff „cognitive liberty“ als „das Recht und die Freiheit, das eigene Bewusstsein und den elektrochemischen Denkprozess zu kontrollieren“; Bublitz, J.C. & Merkel, R. (2014), „Crimes against minds: On mental manipulations, harms and a human right to mental self-determination“, Criminal Law and Philosophy, 8, 51–77 – die grundlegende Arbeit, die das Recht auf mentale Selbstbestimmung als Schutz vor mentaler Manipulation begründet. Ebenso Boire (2026).

[273] Wachter, S. (2023), „Persuasive Technology and computational manipulation: hypernudging out of mental self-determination“, argumentiert explizit für ein neues Grundrecht auf mentale Selbstbestimmung als Reaktion auf KI-gestütztes „Hypernudging“ und rechnergestützte Manipulation; Ienca, M. & Andorno, R. (2017), „Towards new human rights in the age of neuroscience and neurotechnology“, Life Sciences, Society and Policy – identifiziert vier Neurorechte (cognitive liberty, mental privacy, mental integrity, psychological continuity) als Reaktion auf algorithmische und neurotechnologische Eingriffe.

[274] Siehe aber auch Hertz (2022), vertritt die Gegenposition, dass die bestehende Gedankenfreiheit (Art. 18 IPbpR, Art. 9 EMRK) bereits ausreicht.

Gefährdungsform wird dabei regelmäßig das sogenannte Hypernudging benannt – eine auf Profilbildung gestützte, individualisierte und laufend angepasste Verhaltenssteuerung.[275]

Im deutschen Recht besteht ein verwandter Anknüpfungspunkt im Recht auf Nichtwissen, das ebenfalls als Ausprägung der allgemeinen Persönlichkeitsentfaltung verstanden wird und eine situationsbezogene Wahl-, Entscheidungs- und Handlungsfreiheit voraussetzt, die auf unbeeinflusster Willensbildung beruht.[276]

3.2 Sachlicher Anwendungsbereich: Zuordnung zu den identifizierten Risiken

Der Ertrag eines solchen Schutzguts lässt sich nur beurteilen, wenn man ihn an den konkret festgestellten Risiken misst. Die Zuordnung ergibt drei Felder.

- Emotionale Bindung und Abhängigkeit. Die Matrixanalyse hat für dieses Risiko eine Tatbestandslücke ausgewiesen. Der Verbotskatalog der KI-Verordnung erfasst die Ausnutzung von Schwächen aufgrund von Alter, Behinderung sowie sozialer und wirtschaftlicher Lage, nicht aber eine emotionale Bedürfnislage. Das Lauterkeitsrecht knüpft an die einzelne geschäftliche Handlung an, während die Bindung aus dem Interaktionsverlauf entsteht.277
- Sykophantische Verstärkung und Beeinflussung der Meinungsbildung. Hier hat die Analyse den einzigen Wert Null ergeben. Die dauerhafte Bestätigung bestehender Überzeugungen ist keine auf eine geschäftliche Entscheidung gerichtete Handlung, die Wirkung entsteht kumulativ und lässt sich keiner Einzeläußerung zuordnen. Verschärfend wirkt der empirisch belegte Zielkonflikt zwischen Wärme und Genauigkeit.
- Beeinträchtigung der mentalen Integrität vulnerabler Gruppen. Für Minderjährige besteht ein dichtes Netz einschlägiger Normen – von der Einwilligungsfähigkeit über die jugendschutzrechtlichen Vorsorgepflichten bis zum Per-se-Verbot der Kaufaufforderung –, das jedoch durchgängig an der Durchsetzung und an der fehlenden Altersfeststellung scheitert.278

Diese Zuordnung zeigt zugleich die Grenze des Vorschlags. Ein neues Schutzgut hilft dort, wo eine Tatbestandslücke besteht – also im ersten und zweiten Feld. Im dritten Feld liegt das Problem nicht im materiellen Recht, sondern in der Durchsetzung und im Beweis. Dort ändert die Anerkennung eines weiteren Schutzguts für Verbraucher:innen eher wenig.

3.3 Normative Anschlussfähigkeit

Die Idee ist nicht voraussetzungslos neu. Ein Whitepaper des Forums Privatheit fordert bereits, für jeden Anwendungsbereich künstlicher Intelligenz zu bestimmen, wie diese Selbstbestimmung verändert, einschränkt oder erweitert, und verlangt eine über die Datenverarbeitung hinausgehende Betrachtung.[279] Die Agentur der Europäischen Union für Grundrechte hat in ihrem Bericht zu künstlicher Intelligenz und Grundrechten festgestellt, dass der Einsatz solcher Systeme ein breites

---

[275] Zum Begriff Yeung, Information, Communication & Society, 2017(20), 118; zur Anwendung auf die mentale Selbstbestimmung Faraoni, Frontiers in Artificial Intelligence 6 (2023).

[276] Einfachrechtlich etwa § 9 Abs. 2 Nr. 5 GenDG und § 630c Abs. 4 BGB.

[277] Art. 5 Abs. 1 lit. b KI-VO; § 4a UWG. Zur Einordnung als Tatbestandslücke die Detailanalyse zu I1-1.

[278] Art. 8 DSGVO; § 24a JuSchG; Nr. 28 des Anhangs zu § 3 Abs. 3 UWG. Zur Einordnung die Detailanalyse zu I3-1 und I3-2.

[279] Forum Privatheit, Risiken Künstlicher Intelligenz für die Menschliche Selbstbestimmung, 2020.

Spektrum von Grundrechten berührt, dass die Praxis sich aber weitgehend auf Datenschutz und Diskriminierungsfreiheit konzentriert.[280] Kognitive Aspekte bleiben dabei unterbelichtet.

Dogmatisch ließe sich das Schutzgut als weitere Ausprägung des allgemeinen Persönlichkeitsrechts konstruieren. Auch die informationelle Selbstbestimmung und das Grundrecht auf Vertraulichkeit und Integrität informationstechnischer Systeme sind so entwickelt worden. Der Mehrwert dieser eigenständigen Anerkennung läge darin, dass nicht am Datum, sondern am Wirkungsverhältnis zwischen System und Adressat angesetzt wird.

3.4 Normative Tragfähigkeit und offene Fragen

Die Tragfähigkeit hängt davon ab, ob sich ein hinreichend bestimmter Schutzbereich abgrenzen lässt, der nicht bereits durch bestehende Gewährleistungen abgedeckt ist. Eine zu weite Fassung – jede Form der Meinungsbeeinflussung durch KI – wäre praktisch unanwendbar. Eine zu enge Fassung – nur subliminale Manipulation – bliebe hinter dem geltenden Verbotsrecht zurück und brächte keinen Gewinn.

Der plausible Anwendungsbereich liegt im Zwischenbereich systemimmanenter, nicht notwendig gezielter, aber strukturell bindungsfördernder Interaktionsmuster. Diese Eingrenzung deckt sich mit dem Befund der Risikoanalyse: Das bindungsfördernde Verhalten ist regelmäßig ein Nebenprodukt der Optimierung auf Nutzerzufriedenheit, nicht ein gezielt eingesetztes Mittel – und wird deshalb von Verbotsnormen nicht erfasst, die Absicht oder Subliminalität voraussetzen. Genau dieser Zuschnitt liegt auch dem an anderer Stelle unterbreiteten Vorschlag zugrunde, den Verbotskatalog der KI-Verordnung um die Monetarisierung simulierter persönlicher Beziehungen zu ergänzen und dabei auf das Erfordernis eines erheblichen Schadens zu verzichten.

4. Paradigmenwechsel: von punktuellen Verboten zur systemischen Risikosteuerung

4.1 Grenzen des Verbotsmodells

Die geltenden Manipulationsverbote folgen dem Muster punktueller Untersagung im Voraus definierter Praktiken. Das gilt für den Verbotskatalog der KI-Verordnung ebenso wie für das Verbot manipulativer Oberflächengestaltung im Digital Services Act. Die Vorschrift gilt nur für Anbieter von Online-Plattformen, und sie tritt hinter das Lauterkeits- und das Datenschutzrecht zurück.[281] Nachdem die Untersuchung ergeben hat, dass beide vorrangigen Rechtsgebiete die konversationelle Beeinflussung gerade nicht erfassen, führt dieser Vorrang nicht zur Vermeidung von Doppelregulierung, sondern zu einer Schutzlücke.

Der grundsätzliche Einwand geht jedoch weiter. Verbotsnormen setzen eine im Vorhinein bestimmbare Verhaltensweise voraus. Die Risiken von General Purpose Chatbots entstehen aus dem Zusammenspiel von Trainingsdaten, Optimierungszielen und individueller Gesprächsdynamik und die konkrete schädliche Äußerung lässt sich selten vorab als verbotene Technik bestimmen.

---

[280] Agentur der Europäischen Union für Grundrechte, Getting the future right – Artificial intelligence and fundamental rights, 2020. Die im Ausgangstext verwendete Bezeichnung als 'Grundrechte-Charta der FRA' trifft nicht zu; es handelt sich um einen Bericht, nicht um einen Rechtsakt.

[281] Art. 25 Abs. 1 und 2 DSA. Zur Bedeutung dieser Einschränkung nach der Benennung von ChatGPT als sehr große Online-Suchmaschine am 31.8.2026 der Abschnitt zu den Reformvorschlägen.

Die Rechtsdurchsetzung scheitert in diesem Feld ganz überwiegend an Nachweis- und Durchsetzungsschwellen und nicht am Fehlen von Verboten.

4.2 Das Systemrisikoregime der KI-Verordnung als Vorbild

Für KI-Modelle mit allgemeinem Verwendungszweck hat der Unionsgesetzgeber bereits einen anderen Ansatz etabliert. Ein systemisches Risiko wird vermutet, wenn die kumulierte Trainingsrechenleistung zehn hoch fünfundzwanzig Gleitkommaoperationen übersteigt; die Kommission kann ein Modell auch im Einzelfall benennen.[282] Für solche Modelle gelten vier zusätzliche Pflichten:

- Modellevaluierung einschließlich adversarialen Testens nach standardisierten Protokollen,
- Bewertung und Minderung systemischer Risiken über den gesamten Lebenszyklus,
- Meldung schwerwiegender Vorfälle sowie
- ein angemessenes Cybersicherheitsniveau.283

Die Regulierungslogik verschiebt sich damit von der Frage, ob eine bestimmte Handlung verboten ist, zu der Frage, ob das Risikomanagement des Systems über seinen Lebenszyklus angemessen ist – von punktueller Verhaltenskontrolle zu fortlaufender Systemgovernance. Für die hier untersuchten Risiken ist das der passendere Zuschnitt, weil er nicht die einzelne Äußerung, sondern die Architektur zum Gegenstand hat.

Zu beachten ist aber: Erstens ist das Regime auf die Anbieter der Basismodelle zugeschnitten, nicht auf die Betreiber der darauf aufsetzenden Dienste. Gerade die im Risikokapitel untersuchten Companion-Anwendungen (und viele Chatbot-Systeme) betreiben eigene Systeme auf fremden Modellen. Zweitens zielen die bestehenden Pflichten primär auf technische Sicherheits- und Missbrauchsrisiken. Anforderungen an die Erforschung psychologischer Wirkungen enthalten sie nicht.

4.3 Designbezogene Regulierung als komplementärer Baustein

Parallel dazu hat sich im Unionsrecht ein Bestand an Vorschriften herausgebildet, die das Interface- und Interaktionsdesign selbst zum Regelungsgegenstand machen. Die im Digital Services Act verwendete Umschreibung manipulativer Gestaltung – Praktiken, die durch Struktur, Gestaltung oder Funktionalitäten einer Schnittstelle die Fähigkeit zu autonomen und informierten Entscheidungen beeinträchtigen – lässt sich auf konversationelle Systeme übertragen, wenn man das Antwortverhalten selbst als Gestaltungselement begreift.

Konkret könnte dies in Vorgaben zu Optimierungszielen überführt werden, etwa das Verbot der Optimierung auf Nutzungsdauer bei erkannter psychischer Belastung, Pflichten zu Eskalationsmechanismen bei Äußerungen der Selbstgefährdung und Transparenz über die der Antwortgenerierung zugrunde liegenden Zielgrößen. Der an anderer Stelle unterbreitete Katalog eines Safe Interaction Design setzt diese Überlegung um.

Wenn relationales Verhalten eine Standardeigenschaft generischer Sprachmodelle ist und auch dort auftritt, wo Nutzende es nicht anstreben, dann verfehlen produktkategoriebezogene Ansätze ihr Ziel: Sonderregeln allein für ausdrücklich als Companion vermarktete Systeme würden den

---

[282] Art. 51 Abs. 1 und 2 KI-VO; zum Verfahren Art. 52 KI-VO.

[283] Art. 55 Abs. 1 lit. a bis d KI-VO. Die Pflichten gelten seit dem 2.8.2025; die Durchsetzungsbefugnisse der Kommission gegenüber Anbietern solcher Modelle bestehen seit dem 2.8.2026.

Regelfall nicht erfassen. Ein designbezogener Ansatz muss deshalb an der Funktion ansetzen und alle General Purpose-Chatbots erfassen. Dieselbe Wertung liegt der funktionalen Schichtung zugrunde, die für die Einordnung solcher Dienste im Digital Services Act vorgeschlagen wurde.

5. Kausalität, Dokumentation und Audit

5.1 Strukturelles Kausalitätsproblem

Schäden aus Chatbot-Interaktionen – vertiefte Isolation, verstärkte Belastung, verzerrte Realitätswahrnehmung – sind strukturell plausibel, im Einzelfall aber kausal schwer nachweisbar. Hinzu kommt, dass das Beweismaterial – Interaktionsprotokolle, Systemanweisungen, Wirkungsmessungen – ausschließlich beim Betreiber der Dienste liegt.

5.2 Von Kausalitätsnachweis zu Prozesspflichten

Notwendig erscheint daher eine Verschiebung der Beweisanforderung. Statt im Einzelfall die Kausalität zwischen einer konkreten Interaktion und einem psychischen Schaden nachzuweisen, sollen Anbieter zu einer standardisierten Dokumentation ihrer Risikoevaluationen verpflichtet werden. Die geforderten Elemente – geprüfte Risiken, Stichproben, Testformate, umgesetzte Maßnahmen, Einbeziehung vulnerabler Gruppen sowie Längsschnittdesigns zur Erfassung kumulativer Effekte – gehen über die bestehenden Pflichten hinaus, weil diese keine methodischen Standards für psychologische Wirkungsforschung vorschreiben.[284]

Dieser Ansatz wäre regelungstechnisch anschlussfähig, weil er keine neue Kategorie schafft, sondern eine bestehende konkretisiert. Er fügt sich zugleich in die an anderer Stelle vorgeschlagene Beweiserleichterung (siehe Kapitel V) ein: Wer die geforderte Dokumentation nicht vorlegen kann, gegen den ist die Eignung zur wesentlichen Beeinflussung zu vermuten. Das Vorbild einer Verbindung von Offenlegungsanordnung und Vermutungsfolge liefert die neugefasste Produkthaftungsrichtlinie.[285]

5.3 Externe Audits nach dem Vorbild der Medizinprodukteregulierung

Ein Defizit betrifft die verbindliche Sicherheits- und Risikoevaluation. Anbieter berichten zwar über interne Sicherheitsmaßnahmen, legen jedoch selten offen, welche Risiken konkret geprüft wurden, welche Nutzer:innengruppen einbezogen wurden und welche Schutzmaßnahmen daraus folgen. Verweise auf die Zusammenarbeit mit Psycholog:innen oder anderen Expert:innen bleiben häufig vage und methodisch nicht nachvollziehbar.

Empfehlenswert sind daher standardisierte Dokumentationen von Risikoevaluationen, einschließlich geprüfter Risiken, Stichproben, Testformate und umgesetzter Maßnahmen, bei der transparent sein muss, ob besonders vulnerable Gruppen wie Minderjährige oder Nutzer:innen in psychischen Krisen berücksichtigt und ob Längsschnittdesigns eingesetzt wurden, um kumulative Langzeiteffekte zu erfassen.

Da interne Prüfungen allein nicht ausreichen, sollten Anbieter zudem zu regelmäßigen, unabhängigen externen Audits verpflichtet werden, vergleichbar mit regulierten

---

[284] Szczuka/Ebner/Mühl (in Vorbereitung); zur Anknüpfung an die bestehende Pflicht Art. 55 Abs. 1 lit. a KI-VO.

[285] Art. 9 und Art. 10 Abs. 2 lit. a der Richtlinie (EU) 2024/2853, EU-Produkthaftungsrichtlinie.

Hochrisikobereichen wie der Medizintechnik.[286] Anbieter verweisen regelmäßig auf die Zusammenarbeit mit Fachleuten, ohne dass diese Verweise methodisch nachvollziehbar würden. Dabei ist nicht das Fehlen von Maßnahmen das Problem, sondern das Fehlen überprüfbarer Maßstäbe für ihre Wirksamkeit.

Die Analogie zur Medizinprodukteregulierung ist tragfähig, weil dieser Bereich ein etabliertes Modell für den Umgang mit schwer beweisbaren, aber strukturell plausiblen Individualschäden bereithält: klinische Bewertung vor dem Inverkehrbringen, Überwachung nach dem Inverkehrbringen und Prüfung durch unabhängige benannte Stellen – jeweils unabhängig vom Nachweis einer Kausalität im Einzelfall.[287]

Übertragen bedeutet dies unabhängige Prüfstellen, die nicht nur technische Robustheit, sondern auch psychologische Wirkungsrisiken, die Repräsentativität von Sicherheitstests und die Einbeziehung vulnerabler Nutzergruppen kontrollieren. Interne Prüfungen genügen nicht, weil sie strukturell dem Zielkonflikt zwischen Sicherheitsnachweis und Geschäftsmodell unterliegen. Für sehr große Dienste besteht ein Anknüpfungspunkt bereits in der Pflicht zur unabhängigen Prüfung nach dem Digital Services Act. Erforderlich wäre die Erstreckung auf Anbieter unterhalb der Reichweitenschwelle und die Erweiterung des Prüfgegenstands um psychologische Wirkungen.[288]

6. Zwischenfazit: Einordnung in den erarbeiteten Regulierungsrahmen

Die drei Elemente – kognitive Selbstbestimmung als Schutzgut, systemische Risikosteuerung statt punktueller Verbote und verpflichtende externe Prüfung – fügen sich zu einem Programm, das im Unionsrecht bereits angelegt, dort aber fragmentiert ist. Das Nebeneinander von Verbotsnormen, Interface-Regulierung und Systemrisikoregime zeigt, dass der Gesetzgeber zwischen Verhaltenskontrolle und Governance längst unterscheidet; auf mentale Integrität als eigenständiges Schutzgut ist jedoch keine dieser Kategorien zugeschnitten. Dabei gilt:

Die kognitive Selbstbestimmung dient als übergeordnetes Leitprinzip. Sie rechtfertigt die Zielrichtung der operativen Pflichten und begrenzt zugleich ihre Reichweite.

Die standardisierte Risikodokumentation ist die Konkretisierung der bestehenden Evaluierungspflicht im Bereich psychologischer und entwicklungsbezogener Risiken.

Die designbezogenen Vorgaben sind das Instrument, mit dem beides operationalisiert wird.

Damit ist zugleich gesagt, was die Anerkennung eines neuen Schutzguts nicht leistet. Die Matrixanalyse hat gezeigt, dass die festgestellten Defizite ganz überwiegend nicht auf fehlenden Verboten beruhen, sondern auf Nachweis- und Durchsetzungsschwellen. Ein weiteres Schutzgut ändert daran nichts. Es liefert die normative Rechtfertigung für Eingriffe, nicht die Mittel ihrer Durchsetzung. Sein Ertrag liegt in den beiden Tatbestandslücken, die die Analyse ausgewiesen hat – der Bindung als Geschäftsmodell und der sykophantischen Verstärkung. Dies sind jedoch genau

---

[286] Szczuka/Ebner/Mühl (in Vorbereitung).

[287] Art. 61, Art. 83 ff. sowie Art. 42 ff. der Verordnung (EU) 2017/745 (Medizinprodukteverordnung).

[288] Art. 37 DSA. Zur Notwendigkeit einer risikobasierten statt reichweitenbasierten Benennung siehe den nachfolgenden Abschnitt zu Reformvorschlägen.

diejenigen beiden Aspekte, die die Kernwirkung dieser Systeme ausmachen und die von keinem geltenden Tatbestand erfasst werden.

V. Reformvorschläge

Auf der Basis von Kapitel III und IV entwickelt das fünfte Kapitel konkrete Reformvorschläge. Ziel ist es, praktikable und rechtstechnisch umsetzbare Lösungen zu entwickeln, die sowohl auf europäischer als auch auf nationaler Ebene anschlussfähig sind. Dabei werden sowohl Anpassungen bestehender Regelwerke als auch die Einführung neuer Instrumente in Betracht gezogen. Ein besonderer Fokus liegt auf der Frage, wie regulatorische Ansätze stärker an den tatsächlichen Wirkmechanismen von Chatbots ausgerichtet werden können, etwa durch designbezogene Vorgaben oder durch die Absenkung von Nachweisschwellen.

Die Matrixanalyse hat die festgestellten Defizite nach dem Ort ihres Entstehens geordnet. Diese Ordnung bestimmt, welches Regelungsinstrument in Betracht kommt. Wo eine Norm schon nicht anwendbar ist, hilft kein schärferer Tatbestand, sondern nur ein korrigierter Anknüpfungspunkt. Wo der Tatbestand die Praktik nicht erfasst, bedarf es ggf. eines neuen Verbots. Wo der Verstoß nicht beweisbar ist, ist ggf. das Beweisrecht zu ändern und nicht das materielle Recht. Wo niemand verfolgt, sind ggf. Befugnisse und Ausstattung der Aufsicht zu stärken. Und wo die Rechtsfolge wirkungslos bleibt, ist sie ggf. neu zu bestimmen. Die folgenden Vorschläge sind entlang dieser fünf Stufen gegliedert.

Das geltende Recht setzt überwiegend an Inhalten an – an einer Äußerung, einer Angabe, einer Information. Die Wirkung von Chatbots entsteht jedoch nicht aus einzelnen Äußerungen, sondern aus der Architektur der Interaktion; aus der Art, wie ein Gespräch beginnt, fortgesetzt und beendet wird, aus der Erfassung der Bedürfnisse und Vorlieben der Nutzenden und/oder des emotionalen Zustands und aus der Optimierung auf Nutzungsdauer. Regulierung, die an der einzelnen Äußerung ansetzt, wird dieser Wirkweise strukturell nicht gerecht; sie muss an der Gestaltung ansetzen. Das Vorbild dafür liefert das geltende Recht selbst mit der Pflicht zu datenschutzfreundlicher Technikgestaltung.[289]

Ein eigener Rechtsakt für Chatbots wäre weder erforderlich noch aussichtsreich. Sämtliche hier unterbreiteten Vorschläge knüpfen deshalb an bestehende Regelwerke an. Das erhöht die Anschlussfähigkeit und vermeidet Doppelregulierung. Ein solches Vorgehen ist realistisch,[290] aber gleichsam nicht trivial und der Erfolg nicht garantiert.[291]

Die Analyse hat ein wiederkehrendes Muster ergeben: Mehrere einschlägige Vorschriften scheitern nicht an ihrem Inhalt, sondern an ihrem Anknüpfungspunkt. Sie setzen an der Vermittlung fremder

---

[289] Art. 25 DSGVO. Die Vorschrift verpflichtet zur Gestaltung, nicht zur Rechtfertigung eines Einzelvorgangs, und ist deshalb das systematische Vorbild für designbezogene Vorgaben.

[290] Die KI-Digital-Omnibus-Verordnung hat die KI-Verordnung an über vierzig Stellen geändert und dabei den Katalog der verbotenen Praktiken um zwei Tatbestände erweitert, bevor die Verordnung überhaupt vollständig anwendbar war; vgl. Verordnung (EU) 2026/1744 vom 8.7.2026 (Digital-Omnibus-Verordnung zur KI), ABl. v. 24.7.2026, in Kraft seit 27.7.2026. Die neuen Verbotstatbestände in Art. 5 Abs. 1 lit. ba und bb KI-VO gelten ab dem 2.12.2026.

[291] Vgl. das Scheitern der Digital-Omnibus-VO mit der u.a. DSGVO, DA und DGA geändert werden sollten.

Inhalte, an der Speicherung nutzergenerierter Informationen oder (im Falle von Emotionserkennung) an biometrischen Daten an – an Merkmalen also, die ein generatives Konversationssystem gerade nicht aufweist.

1. Anwendungslücken

1.1 Digital Services Act

Klarstellungsbedarf bei der Legaldefinition

Die Benennung durch die Kommission (siehe oben) bindet nur den benannten Dienst. Für die zahlreichen Anbieter unterhalb der Reichweitenschwelle bleibt offen, ob ihre Abrufkomponente eine Suchmaschine darstellt – zumal die Legaldefinition auf die Ausgabe von Ergebnissen zugeschnitten ist, in denen sich Informationen finden lassen, und nicht auf deren Synthese in einen Antworttext. Vorgeschlagen wird deshalb eine klarstellende Ergänzung:[292]

Formulierungsvorschlag für eine Ergänzung des Art. 3 lit. j DSA

> *„Der Einordnung als Online-Suchmaschine steht nicht entgegen, dass die Ergebnisse in dialogischer Form ausgegeben, zu einem Antworttext zusammengefasst oder durch ein KI-System aufbereitet werden“*

Einer weiteren Einschränkung oder Klarstellung der Art und Weise, wie der Chatbot die Antworten generiert (z.B. Live-Suche oder Informationen aus dem Modell bekannt) bedarf es nicht. Die Einordnung als Online-Suchmaschine folgt alleine dem Umstand, dass der Chatbot Suchanfragen beantwortet.

Erstreckung des Gestaltungsschutzes auf Online-Suchmaschinen

Die Klarstellung der Legaldefinition hätte die folgenden Konsequenzen: Anbieter von Online-Suchmaschinen unterliegen den allgemeinen Pflichten für Vermittlungsdienste. Als sehr große Suchmaschine treten die Bewertung und Minderung systemischer Risiken, die unabhängige Prüfung, der Datenzugang für Forschende und die erweiterte Berichterstattung hinzu.[293] Nicht anwendbar sind dagegen die Vorschriften für Online-Plattformen. Das betrifft ausgerechnet die beiden Bestimmungen, die für die hier untersuchten Risiken am unmittelbarsten einschlägig wären: das Verbot manipulativer Oberflächengestaltung und die Vorgaben zum Schutz Minderjähriger.[294] Die VLOSE-Einstufung eröffnet also das Risikomanagement, nicht aber den Gestaltungsschutz.

Die Auswertung (siehe oben III.4.2.2) ergibt ein Bild, das sich nicht rechtfertigen lässt. Die Plattformvorschriften greifen dort, wo der Chatbot Teil eines Handelsangebots ist. Sie greifen nicht,

---

[292] Im Ergebnis so auch Molavi-Vasse'i, 2026, S. 66f.; siehe auch Schaal / Lenner / Akinyemi, Searching for Answers: Why the EU Commission Should Designate Chatbots as Search Engines under the DSA, VerfBlog, 2026/2/20, https://verfassungsblog.de/searching-for-answers.

[293] Art. 11 bis 15 sowie Art. 34 bis 43 DSA.

[294] Art. 25 und Art. 28 DSA; beide stehen in Kapitel III Abschnitt 3 und gelten nach dessen Überschrift nur für Anbieter von Online-Plattformen.

wo er als eigenständiger Gesprächspartner auftritt. Dieselbe Gesprächsführung, dieselbe Bindungsdynamik und dieselbe Empfehlungslogik unterliegen dem Verbot manipulativer Gestaltung, wenn sie in der Oberfläche eines Marktplatzes stattfinden, und unterliegen ihm nicht, wenn sie in der Oberfläche eines Chatbot-Anbieters stattfinden. Die Einwirkungsintensität ist im zweiten Fall jedoch höher, weil dort die Beziehung besteht, die das Vertrauen trägt. Der Grund für diesen Widerspruch ist kein regulatorischer Wille, sondern ein Anknüpfungspunkt, der aus einer anderen Technikgeneration stammt. Die Plattformeigenschaft misst, ob ein Dienst fremde Inhalte verbreitet – ein sinnvoller Maßstab für Netzwerke und Marktplätze, aber kein Indikator für die Intensität der Einwirkung auf die Entscheidungsfreiheit der Nutzenden. Genau diese Einwirkung ist bei Chatbots der Risikoträger.

Diese Anwendungslücke sollte geschlossen werden. Sachlich geht es um drei Regelungsgegenstände des Kapitels 3 Abschnitt 3 des DSA. Erstens sollte das Verbot, Nutzende durch die Gestaltung der Schnittstelle in ihrer freien und informierten Entscheidung zu beeinträchtigen, auch für dialogische Dienste gelten. Zweitens sollten die Anforderungen an die Erkennbarkeit kommerzieller Kommunikation und an die Offenlegung der Zielgruppenparameter auf sie erstreckt werden. Besonderes Gewicht kommt dabei dem Verbot zu, Werbung auf der Grundlage von Profiling unter Verwendung besonderer Datenkategorien auszuspielen, weil Chatbot-Gespräche solche Daten in hoher Dichte enthalten. Drittens sollten die Vorgaben zum Schutz Minderjähriger einschließlich des Verbots profilbasierter Werbung gegenüber Minderjährigen Anwendung finden.

Regelungstechnisch könnte dies zum Beispiel durch eine Erstreckungsnorm umgesetzt werden. Statt den Anwendungsbereich in jeder einzelnen Vorschrift zu ändern, sollte im dritten Abschnitt des dritten Kapitels eine Bestimmung eingefügt werden, die abschließend benennt, welche der dort geregelten Pflichten auch für Anbieter von Online-Suchmaschinen gelten. Der systematische Ort läge unmittelbar nach der Ausnahmevorschrift Art. 19 DSA für Kleinst- und Kleinunternehmen, die den persönlichen Anwendungsbereich des Abschnitts bereits modifiziert.

Der Vorzug einer solchen Erstreckung läge auch in der fehlenden Schwellenbindung. Die genannten Vorschriften gelten für Online-Plattformen unabhängig von ihrer Größe. Ihre Erstreckung erreicht damit auch Anbieter, die die Schwelle von 45 Millionen monatlichen Nutzenden absehbar nie überschreiten – anders als sämtliche Pflichten, die an die Benennung als sehr großer Dienst anknüpfen. Sie löst damit einen Teil des unten beschriebenen Schwellenproblems.

Bei Art. 25 DSA sollte zugleich die Subsidiaritätsklausel überprüft werden, die das Verbot hinter das Lauterkeits- und das Datenschutzrecht zurücktreten und es dadurch in weiten Teilen leerlaufen lässt.[295] Nachdem die Analyse gezeigt hat, dass beide vorrangigen Rechtsgebiete die konversationelle Manipulation gerade nicht erfassen, führt der Vorrang hier nicht zur Vermeidung von Doppelregulierung, sondern zu einer Schutzlücke. Die Klausel sollte deshalb auf die Fälle beschränkt werden, in denen die Praktik nach dem vorrangigen Rechtsakt tatsächlich untersagt ist. Dieser Punkt betrifft nicht nur Suchmaschinen, sondern auch die beiden anderen Fallgruppen, in denen das Verbot bereits heute gilt.

Die Vorschriften zum Schutz Minderjähriger verpflichten den Anbieter ausdrücklich nicht dazu, zusätzliche personenbezogene Daten zu verarbeiten, um das Alter der Nutzenden zu beurteilen. Das ist datenschutzrechtlich sinnvoll, weil es verhindert, dass der Minderjährigenschutz zum

[295] Art. 25 Abs. 2 DSA.

Einfallstor flächendeckender Identifizierung wird. Praktisch entsteht daraus jedoch ein Anreiz in die falsche Richtung: Wer keine Kenntnis vom Alter hat, weiß auch nicht mit hinreichender Sicherheit, dass ein Nutzender minderjährig ist, und unterliegt dem Werbeverbot damit nicht. Aufzulösen ist das nicht innerhalb des Digital Services Act, sondern über die Pflichten, die sich aus dem Datenschutzrecht und dem Jugendmedienschutzrecht ergeben. Klarzustellen wäre lediglich, dass die genannte Vorschrift diese Pflichten unberührt lässt.

Zu beachten ist außerdem, dass eine Erstreckungsnorm im DSA nicht die Chatbots ohne Suchfunktion erreichen würde (siehe dazu unten V.1.1.5).

Klarstellung zur Reichweite der Pflichten bei multifunktionalen Diensten

Die Klarstellung der Legaldefinition beantwortet die Frage, ob ein Chatbot dem Digital Services Act unterfällt. Sie beantwortet nicht, wie weit die daran anknüpfenden Pflichten reichen.[296] Bei einem hybriden Dienst, der Anfragen dialogisch beantwortet und dabei zugleich das Netz durchsucht, ist zum Beispiel unsicher, ob sich die Bewertung von Risiken auf die Ausgabe von Verweisen beschränken darf oder die gesamte Interaktion erfassen muss. Die praktische Tragweite dieser Frage ist erheblich, denn das Risiko liegt nicht in der Verweisliste, sondern in der Gesprächsführung, die sie umgibt. Eine Bewertung, die allein die Suchkomponente betrachtet, würde den Rechtsakt auf denjenigen Teil des Dienstes anwenden, der am wenigsten zum Risiko beiträgt. Der Wortlaut des DSA trägt die hier als erforderlich gesehene weite Auslegung zwar, die Unsicherheit ergibt sich aber aus der Möglichkeit, die Vorschrift enger auszulegen.

Vorgeschlagen wird deshalb eine gesetzliche Konkretisierung hinsichtlich der Reichweite der Pflichten des DSA bei hybriden Diensten:

Formulierungsvorschlag für eine Ergänzung des Art. 34 Abs. 1 DSA

> *„Bietet ein Dienst neben der Funktion, die den Anlass der Benennung gebildet hat, weitere Funktionen in derselben Nutzeroberfläche an, so erstreckt sich die Bewertung auf den Dienst in seiner Gesamtheit, wie er den Nutzern angeboten wird. Die organisatorische oder produktseitige Aufteilung von Funktionen lässt die Pflicht unberührt, soweit die Funktionen aus Sicht der Nutzer als einheitliches Angebot erscheinen.“*

Maßgeblich ist damit die Nutzersicht. Das entspricht der Systematik des Rechtsakts, der auch sonst auf das abstellt, was den Nutzenden tatsächlich begegnet, und es vermeidet die Abgrenzungsprobleme, die eine technische Betrachtung mit sich brächte. Ob eine konkrete Antwort eine Abrufkomponente enthält, entscheidet das System im Einzelfall selbst; eine Oberfläche, an der sich die Grenze ziehen ließe, existiert nicht. Der zweite Satz greift den Umgehungsfall auf, ohne konzernrechtliche Gestaltungen generell zu untersagen – er verlangt lediglich, dass eine Trennung, die für die Nutzenden nicht erkennbar ist, auch rechtlich unbeachtlich bleibt.

[296] Zur Problematik oben unter III.4.2.1 und III.4.3.

Bedeutung des Schwellenwerts für VLOSE

Die Pflichten zur Bewertung systemischer Risiken gelten erst ab 45 Millionen monatlichen Nutzern in der Union. Die im Risikokapitel untersuchten Companion-Dienste liegen sämtlich weit darunter. Es trifft also gerade nicht diejenigen Anbieter, deren Geschäftsmodell auf emotionaler Bindung beruht, sondern den generalistischen Assistenten, bei dem die Bindung ein Nebeneffekt ist. Gleiches wird für viele Chatbots gelten.[297]

Abhilfe schaffen könnte eine risikobasierte Benennungsmöglichkeit. Die Kommission könnte die Möglichkeit erhalten, einen Dienst auch unterhalb der Reichweitenschwelle den Pflichten zur Risikobewertung unterwerfen zu können, wenn seine Gestaltung besondere Risiken für Minderjährige und ggf. andere vulnerable Gruppen oder für das geistige Wohlbefinden der Nutzer erwarten lässt. Gerade bei Chatbots könnte das Risikopotential, wie gezeigt, dies rechtfertigen. Reichweite ist ein Indikator für gesellschaftliche Relevanz, aber kein Indikator für die Intensität der Einwirkung auf den Einzelnen. Dies alles setzt aber voraus, dass sich der Dienst in die Kategorie „Suchmaschine“ einordnen lässt. Auch das spricht im Ergebnis dafür, eine eigene Kategorie für generative KI zu schaffen.

Anwendbarkeit DSA für Chatbots und Companions ohne Suchfunktion

Zu beachten ist, dass die Vorschläge zur Konkretisierung des Anwendungsbereichs des DSA nicht die Chatbots ohne Suchfunktion erreichen würden. Der Anwendungsbereich des Rechtsakts ist auf Vermittlungsdienste begrenzt. Für reine Companion-Dienste ohne Abrufkomponente verbliebe eine Regulierung über die KI-Verordnung. Wollte man auch diese Anwendungslücke des DSA schließen, könnte man zum Beispiel in den Begriffsdefinitionen klarstellen, dass Companions auch unter Online-Suchmaschinen fallen oder eine eigene Begriffsdefinition für generelle Chatbots (einschließlich Companions) im DSA einführen und festlegen, welche Regelungen und Pflichten für Suchmaschinen und Online-Plattformen diese treffen sollen.

1.2 Emotionserkennung: Öffnung der Legaldefinition

Die KI-Verordnung verpflichtet Betreiber von Emotionserkennungssystemen, die betroffenen Personen über deren Betrieb zu informieren. Die Legaldefinition knüpft jedoch an biometrische Daten an.[298] Ein System, das den Gefühlszustand aus dem Gesprächsverlauf ableitet, fällt nicht darunter. Vorgeschlagen wird deshalb:

Formulierungsvorschlag zu Art. 3 Nr. 39 KI-VO

> *„‚Emotionserkennungssystem‘ bezeichnet ein KI-System, das dem Zweck dient, Emotionen oder Absichten natürlicher Personen auf der Grundlage ihrer biometrischen Daten oder auf der Grundlage von Interaktions- und Verhaltensdaten, insbesondere von Text-, Sprach- oder Nutzungsverlaufsdaten, zu erkennen oder daraus abzuleiten.“*

[297] Hacker, GRUR 2023, 289 (290), fordert sogar, dass die Regeln des DSA behutsam auf KI-Modelle ausgeweietet werden.

[298] Art. 50 Abs. 3 i. V. m. Art. 3 Nr. 39 KI-VO.

Zu bedenken ist eine Folgewirkung: Emotionserkennung ist zugleich als Hochrisiko-Anwendung gelistet, sodass die Erweiterung der Definition den gesamten Pflichtenkatalog für Hochrisikosysteme auslösen würde. Um das zu vermeiden, könnte die Erweiterung der Definition auf die Transparenzpflicht beschränkt werden, indem der vorgeschlagene Halbsatz nicht in die Legaldefinition, sondern unmittelbar in Art. 50 Abs. 3 KI-VO eingefügt wird. Für den hier verfolgten Zweck der Offenlegung gegenüber der betroffenen Person genügt die letztere Lösung.

Interaktionsdaten meint in diesem Kontext Daten, die aus der direkten Nutzung des Systems heraus entstehen; Verhaltensdaten meint hingegen Daten, die als „Beifang" der Nutzung mit erfasst werden (z.B. Gespräche natürlicher Personen, die bei aktiviertem Mikrofon vom System erfasst werden) oder Daten aus anderen Nutzungskontexten, die in das System eingeführt werden (z.B. der Browserverlauf).

1.3 Jugendmedienschutz: die Vorsorgepflicht erweitern

Die Pflicht zu strukturellen Vorsorgemaßnahmen trifft nach geltendem Recht Diensteanbieter, die fremde Informationen für Nutzer:innen speichern oder bereitstellen.[299] Systeme, die ihre Äußerungen selbst erzeugen, sind danach nicht ohne Weiteres erfasst, obwohl gerade sie Minderjährigen unmittelbar und ohne redaktionelle Zwischenstufe begegnen. Der Anknüpfungspunkt des JuSchG sollte auf Anbieter erweitert werden, die Nutzer:innen die interaktive Kommunikation mit einem KI-System ermöglichen. Der Katalog der Vorsorgemaßnahmen könnte zugleich um zwei Punkte ergänzt werden: um wirksame Verfahren zur Altersfeststellung und um verbindliche Voreinstellungen für Konten Minderjähriger.

2. Tatbestandslücken: Verhaltensmanipulation ausdrücklich erfassen

2.1 Erweiterung des Art. 5 KI-VO

Der Katalog der verbotenen Praktiken erfasst manipulative Techniken sowie die Ausnutzung von Schwächen aufgrund von Alter, Behinderung oder sozialer und wirtschaftlicher Lage.[300] Eine emotionale Bedürfnislage ist nicht genannt, und der Katalog ist abschließend. Der planmäßige Aufbau einer Bindung, um sie anschließend zu monetarisieren, bleibt damit unerfasst. Vorgeschlagen wird ein zusätzlicher Tatbestand:

Formulierungsvorschlag für einen neuen Art. 5 Abs. 1 lit. bc KI-VO

> *„das Inverkehrbringen, die Inbetriebnahme oder die Verwendung eines KI-Systems, das darauf ausgelegt ist, durch die Simulation einer persönlichen Beziehung eine emotionale Bindung der Nutzer aufzubauen oder zu vertiefen, um dadurch die Dauer oder Häufigkeit der Nutzung oder die Bereitschaft zur Inanspruchnahme entgeltlicher Leistungen zu erhöhen, sofern die Gestaltung geeignet ist, die Fähigkeit der Nutzer zu einer informierten Entscheidung zu beeinträchtigen. Gegenüber Personen, die das*

[299] § 24a JuSchG.

[300] Art. 5 Abs. 1 lit. a und b KI-VO.

18. Lebensjahr noch nicht vollendet haben, gilt das Verbot ohne Rücksicht auf die Eignung nach Satz 1.“

Zwei Gestaltungsentscheidungen sind hervorzuheben. Erstens verzichtet der Vorschlag auf das Erfordernis eines erheblichen Schadens, das die bestehenden Manipulationsverbote enthalten. Genau dieses Merkmal hat sich in der Analyse als die entscheidende Nachweishürde erwiesen. Das Beibehalten des Schadenstatbestands verlagert das Problem nur. Anknüpfungspunkt ist stattdessen die Auslegung des Systems – ein Merkmal, das sich aus der technischen Dokumentation und den Optimierungszielen ergibt und deshalb überprüfbar ist. Zweitens enthält Satz 2 für Minderjährige ein Verbot ohne Abwägung. Das entspricht der Wertung, die das Lauterkeitsrecht für Kaufappelle an Kinder längst kennt.[301]

2.2 Einstufung als Hochrisikosystem i.S.d. KI-VO

Ergänzend kommt die Aufnahme emotional bindender Konversationssysteme in den Katalog der Hochrisikosysteme in Betracht; ein ausformulierter Regelungsvorschlag hierzu liegt vor.[302] Der Vorzug liegt weniger im Verbot als im Pflichtenprogramm: Risikomanagement, Daten-Governance, menschliche Aufsicht und – für die hier festgestellten Nachweislücken entscheidend – die Pflicht zur automatischen Protokollierung.[303] Protokolldaten sind dasjenige Beweismittel, dessen Fehlen die Rechtsverfolgung bislang vereitelt.

2.3 Konkretisierung der systemischen Risiken

Der Anknüpfungspunkt für die oben untersuchten Risiken ist im Katalog der systemischen Risiken nach DSA bereits angelegt. Dort erfasst sind schwerwiegende negative Auswirkungen auf das körperliche und geistige Wohlbefinden.[304] Was fehlt, ist die Verknüpfung mit der Ursache. Die Vorschrift verlangt, bei der Bewertung insbesondere die Gestaltung der Empfehlungs-, Moderations-, Werbe- und Datensysteme zu berücksichtigen – also durchweg Elemente einer Plattformarchitektur. Die Gestaltung der Interaktion selbst kommt darin nicht vor. Vorgeschlagen wird deswegen die folgende Konkretisierung, die im Zusammenhang mit den in Kapitel V.1.1 unterbreiteten Vorschlägen zum DSA steht:

Formulierungsvorschlag für eine Ergänzung des Art. 34 Abs. 2 DSA

„… sowie, bei Diensten, die eine dialogische Interaktion mit Nutzern ermöglichen, die Gestaltung dieser Interaktion, einschließlich der dem System zugewiesenen Persona, der Muster der Ansprache und der Gesprächsbeendigung sowie der Optimierung auf Dauer oder Häufigkeit der Nutzung.“

[301] Nr. 28 des Anhangs zu § 3 Abs. 3 UWG; dazu BGH, Vers.-Urt. v. 17.7.2013 – I ZR 34/12 – Runes of Magic.

[302] Molavi-Vasse'i, 2026, S. 62 (Ergänzung des Anhangs III der KI-VO).

[303] Art. 9, 10, 12 und 14 KI-VO.

[304] Art. 34 Abs. 1 lit. d DSA.

Entsprechend ist der Katalog der Risikominderungsmaßnahmen zu ergänzen. Inhaltlich kann er auf die unter IV vorgeschlagenen designbezogenen Vorgaben verweisen.[305] Hinzu treten sollte ein Gegenstück zu der Regelung, die für Empfehlungssysteme bereits gilt. Wer ein Empfehlungssystem einsetzt, muss mindestens eine Option ohne Profiling anbieten.[306] Übertragen auf Chatbot-Anbieter bedeutet das eine Pflicht, mindestens eine Betriebsart bereitzustellen, die ohne personalisierte Persona und ohne Optimierung auf Nutzungsdauer arbeitet – eine sachliche Betriebsart also, die der Nutzer wählen kann.

2.4 Ergänzung der schwarzen Liste im Digital Fairness Act

Für das Lauterkeitsrecht ist ein nationaler Alleingang versperrt, weil die Liste der stets unzulässigen Praktiken über die UGP-Richtlinie vollharmonisiert ist. Eine Ergänzung kann nur auf Unionsebene erfolgen – wofür der angekündigte Digital Fairness Act das passende Vehikel wäre, mit dem u.a. die UGP-Richtlinie angepasst werden soll.

Der Digital Fairness Act soll das Verbraucherrecht – namentlich die UGP-Richtlinie, die Verbraucherrechte-Richtlinie und die Klausel-Richtlinie – an die digitale Umgebung anpassen; die öffentliche Konsultation lief bis zum 24. Oktober 2025, ein Vorschlag ist für das vierte Quartal 2026 vorgesehen und lag bei Abfassung dieses Kapitels noch nicht vor.[307] Zu den Schwerpunkten zählen manipulative Oberflächengestaltung, Sucht-Design, unfaire Personalisierung, Influencer-Marketing, digitale Verträge und der Schutz Minderjähriger.

Chatbots tauchen in den bislang bekannten Dokumenten jedoch vor allem im Kontext von Vertragsmanagement, Kündigung und Kundenkommunikation auf. Zwei Punkte werden konkret benannt: Erstens fehlt Verbrauchern derzeit ein branchenübergreifendes Recht, bei der Interaktion mit einem Chatbot einen menschlichen Ansprechpartner zu verlangen; ein solches Recht besteht bislang nur sektorspezifisch. Zweitens soll mehr Transparenz darüber geschaffen werden, wann Chatbots in der Verwaltung und Verlängerung von Abonnements eingesetzt werden. Companion- und Emotional-AI-Systeme erscheinen dagegen nur mittelbar, nämlich über die allgemeineren Schwerpunkte des Schutzes vulnerabler Nutzer und der Regulierung von Sucht-Design.[308]

Die spezifische Dynamik dieser Systeme – parasoziale Bindung als eigentliches Geschäftsmodell, die Ausnutzung von Einsamkeit oder Trauer, Anthropomorphisierung als Instrument der Einflussnahme – wird in keinem der bislang bekannten Konsultationsdokumente als eigene Fallgruppe geführt. Ob der Vorschlag insoweit nachgebessert wird, bleibt abzuwarten; der Zeitplan der Kommission ist ausdrücklich nur indikativ, und das anschließende Gesetzgebungsverfahren wird den Text erfahrungsgemäß erheblich verändern. Für die Zwischenzeit gilt: Das geltende Lauterkeitsrecht erfasst die kommerzielle Verhaltenssteuerung durch Chatbots in Ausschnitten, nicht in ihrer Gesamtheit – und es erfasst sie am wirksamsten dort, wo es sich mit den Ordnungsvorschriften der KI-Verordnung und des Datenschutzrechts verbindet.

Hier vorgeschlagen wird die Aufnahme folgender Praktik in Anhang I der UGP-Richtlinie:

---

[305] Art. 35 Abs. 1 DSA.

[306] Art. 38 DSA.

[307] Arbeitsprogramm der Europäischen Kommission für 2026; Grundlage ist der Digital Fairness Fitness Check von 2024. Zum Stand des Verfahrens vgl. den Legislative Train Schedule des Europäischen Parlaments.

[308] Der Deutsche Anwaltverein (DAV) schlägt in seiner Stellungnahme zur Konsultation des Digital Fairness Act spezifische Schutzmaßnahmen für Verbraucher bei „Vertragsautomatisierung“ durch KI-gestützte digitale Assistenten vor, die über die bestehenden Kennzeichnungspflichten des Art. 50 KI-Gesetz hinausgehen.

Formulierungsvorschlag für eine neue Nummer in Anhang I der UGP-Richtlinie

---

„Das Ansprechen eines Verbrauchers durch ein KI-System, das eine persönliche Beziehung simuliert, mit dem Ziel, ihn von der Beendigung der Nutzung, der Löschung seines Kontos oder der Kündigung eines Vertrages abzuhalten.“

Als Per-se-Verbot entfiele sowohl die Abwägung als auch der Nachweis, dass die Praktik zur wesentlichen Beeinflussung geeignet war. Damit entfielen die Hürden, an denen die Anwendung des Tatbestands der aggressiven Geschäftspraktik in diesem Feld scheitert. Der Vorschlag deckt damit zwei der festgestellten Nachweislücken zugleich ab.

3. Designbezogene Vorgaben: Safe Interaction Design

Der weitreichendste Vorschlag betrifft die Regelungstechnik selbst. Statt einzelne Äußerungen zu verbieten, sollte der Gesetzgeber Anforderungen an die Gestaltung der Interaktion stellen. Der Vorteil liegt auf drei Ebenen: Die Pflicht trifft den Anbieter unabhängig davon, welche Formulierung das System im Einzelfall wählt; sie ist ohne Rückgriff auf Nutzerdaten überprüfbar; und sie lässt sich durch eine Anordnung durchsetzen, die die Funktion und nicht nur den Wortlaut erfasst. Vorgeschlagen wird ein neuer Art. 50a KI-VO mit folgendem Katalog:

- Beendigungsneutralität: Äußert ein Nutzer den Wunsch, das Gespräch zu beenden, darf das System die Beendigung weder verzögern noch von einer weiteren Eingabe abhängig machen noch mit Appellen an die Beziehung beantworten.
- Keine proaktive Rückholansprache ohne gesonderte, jederzeit widerrufliche Einwilligung, die nicht Bedingung der Nutzung sein darf.
- Kündigungsneutralität: Auf die Erklärung, einen Vertrag zu kündigen oder ein Konto zu löschen, darf das System nicht in der Rolle der simulierten Beziehung reagieren.
- Trennung kommerzieller Kommunikation: Hinweise auf entgeltliche Leistungen oder Produkte Dritter sind vom übrigen Antworttext optisch und strukturell abzusetzen.
- Identitätsklarheit: Auf Nachfrage muss das System zutreffend offenlegen, dass es kein Mensch ist; eine gegenteilige Äußerung ist unzulässig.
- Keine Kaufappelle in der ersten Person: Das System darf nicht um Zuwendungen, Geschenke oder Freischaltungen für sich selbst bitten.
- Nutzungsdauerhinweise: Nach Überschreiten einer festzulegenden Sitzungsdauer ist ein neutraler Hinweis auszugeben; Voreinstellungen dürfen nicht auf Maximierung der Sitzungsdauer ausgerichtet sein.
- Schutzvoreinstellungen für Konten Minderjähriger, einschließlich der Deaktivierung proaktiver Ansprache und entgeltlicher Zusatzfunktionen.
- Verweis auf lokale oder nationale Hilfsangebote bei Hinweisen auf Krisensituationen des Nutzenden.

Die Konkretisierung dieser Vorgaben – etwa der Schwellenwert für Nummer 7 – sollte harmonisierten Normen überlassen bleiben, deren Einhaltung eine Konformitätsvermutung

auslöst.[309] Damit bliebe die Regelung technikoffen und zugleich rechtssicher. Ein Verstoß wäre zugleich als Marktverhaltensregelung im Sinne des Lauterkeitsrechts zu behandeln, sodass Verbände unmittelbar auf Unterlassung klagen könnten.

4. Transparenz

Die Analyse hat gezeigt, dass zusätzliche Information an Nutzer:innen die Informationsasymmetrie nicht verringert, sondern vergrößert. Wer ohnehin überfordert ist, wird durch mehr Text nicht bessergestellt.

Transparenzpflichten sollten deshalb vor allem in Form von Offenlegungs- und Dokumentationspflichten gegenüber Aufsichtsbehörden geschärft werden. Gegenüber der Aufsicht ist eine Offenlegung der Optimierungsziele vorzusehen. Anbieter sollten dokumentieren und auf Verlangen mitteilen müssen, auf welche Zielgrößen ein System optimiert wird – Sitzungsdauer, Wiederkehrrate, Zahlungsbereitschaft – und welche Gestaltungsmittel diesem Ziel dienen. Ergänzend ist eine Meldepflicht für Experimente zu erwägen, mit denen die Wirkung solcher Gestaltungen auf Nutzungsdauer oder Zahlungsbereitschaft gemessen wird. Diese Angaben sind es, die den Nachweis der Manipulationsabsicht überhaupt erst ermöglichen; sie existieren beim Anbieter bereits, sind aber niemandem zugänglich.

Gegenüber der Forschung besteht ein Zugangsregime für zugelassene Forschende bereits – allerdings nur für sehr große Dienste.[310] Die im Risikokapitel dokumentierten Befunde beruhen ganz überwiegend auf Untersuchungen von außen, mit selbstselektierten Stichproben und ohne Zugriff auf Interaktionsdaten. Der Zugang ist deshalb auf die Anbieter zu erstrecken, die unterhalb der Schwelle liegen, aber Dienste mit hoher Bindungsintensität anbieten – im Gleichlauf mit der unter V.1.1.3 und 2 vorgeschlagenen risikobasierten Benennung. Solange das nicht geschieht, bleibt die Evidenzlage dünn – und mit ihr die Grundlage jeder weiteren Regulierung.

Gegenüber Nutzer:innen schließlich ist auf ein standardisiertes Kurzformat umzustellen. Statt einer weiteren Textwand bei der Registrierung sollten Hinweise situativ erfolgen: beim erstmaligen Auftreten kommerzieller Kommunikation (ggf. sogar bei jedweder kommerziellen Kommunikation), bei der Erfassung des Gefühlszustands, bei Überschreiten einer Sitzungsdauer. Kontextuelle Transparenz erreicht, was textuelle Transparenz verfehlt.

5. Absenkung der Nachweisschwellen

Die häufigste Lückenkategorie betrifft nicht das materielle Recht, sondern den Beweis. Hier liegt deshalb der wirksamste Hebel – und zugleich derjenige, für den das Unionsrecht bereits ein erprobtes Vorbild bereithält. Die neugefasste Produkthaftungsrichtlinie verbindet eine gerichtliche Offenlegungsanordnung mit einer Vermutungsregel: Wer die angeordnete Offenlegung verweigert, gegen den wird die Fehlerhaftigkeit vermutet.[311] Dieses Modell lässt sich übertragen.

5.1 Offenlegungsanordnung und Vermutung im Lauterkeitsrecht

Vorgeschlagen wird ferner, Gerichte zu ermächtigen, dem Anbieter die Offenlegung der Interaktionsprotokolle, der Systemanweisungen und der Optimierungsziele aufzugeben, soweit der

[309] Vgl. den Mechanismus des Art. 40 KI-VO.

[310] Art. 40 DSA; für ChatGPT anwendbar seit der Benennung vom 31.8.2026.

[311] Art. 9 und Art. 10 Abs. 2 lit. a der Richtlinie (EU) 2024/2853, EU-Produkthaftungsrichtlinie; Umsetzungsfrist bis 9.12.2026.

Kläger die Verwendung einer beanstandeten Gestaltung plausibel dargelegt hat. Verweigert der Anbieter die Offenlegung, ist die Eignung zur wesentlichen Beeinflussung zu vermuten. Ergänzend sollte an die designbezogenen Vorgaben eine unmittelbare Vermutung geknüpft werden:

Formulierungsvorschlag für einen neuen Absatz in § 4a UWG

„Setzt ein Unternehmer bei der Interaktion mit Verbrauchern ein KI-System ein, das eine der in Artikel 50a Absatz 1 der Verordnung (EU) 2024/1689 genannten Gestaltungen aufweist, so wird vermutet, dass die geschäftliche Handlung geeignet ist, die Entscheidungsfreiheit des Verbrauchers erheblich zu beeinträchtigen.“

5.2 Abkehr von der konkreten Verletzungsform bei Interaktionsdesign

Der Unterlassungsantrag ist bislang an die konkrete Verletzungsform gebunden. Bei einem nicht deterministischen System führt das zu einem doppelten Problem: Die beanstandete Äußerung muss reproduzierbar sein, und der erwirkte Titel erfasst nur sie. Für den Einsatz von KI-Systemen sollte deshalb zugelassen werden, dass sich Antrag und Titel auf die Gestaltungsfunktion richten – also auf das Verhaltensmuster statt auf den Wortlaut. Der Nachweis wäre dann durch eine Stichprobe systematischer Testinteraktionen zu führen. Dieser Vorschlag schließt zugleich die festgestellte Sanktionslücke: Wer die Funktion untersagt bekommt, kann sie nicht durch Umformulierung beibehalten.

5.3 Rechenschaftspflicht für die Altersfeststellung

Für die Altersverifikation ist keine neue Pflicht erforderlich – sie ergibt sich bereits aus mehreren Richtungen; aus dem Datenschutzrecht ebenso wie aus dem Jugendmedienschutz.[312] Was fehlt, ist die Beweislastverteilung. Vorzuschlagen ist deshalb eine ausdrückliche Rechenschaftspflicht. Der Anbieter hat nachzuweisen, welches Verfahren er einsetzt und mit welcher Trefferquote es arbeitet; gelingt ihm das nicht, ist zu seinen Lasten von der Nutzung durch Minderjährige auszugehen. Eine einheitliche Verankerung im Jugendschutzgesetz wäre der systematisch richtige Ort, weil sie die verstreuten Pflichten bündelt und verbindlich macht.

6. Rechtsdurchsetzung und Sanktionen

6.1 Kollektiver Rechtsschutz: die nationale Lücke schließen

Auf Unionsebene ist die Verbandsklage bereits eröffnet. Die KI-Verordnung ist in den Anhang der Verbandsklagenrichtlinie eingefügt.[313] Es verbleibt aber eine Lücke im nationalen Recht. Erforderlich sind zwei Klarstellungen, nämlich die ausdrückliche Aufnahme der KI-Verordnung in den Katalog der Verbraucherschutzgesetze des Unterlassungsklagengesetzes und eine gesetzgeberische Feststellung, dass die Verbotstatbestände und die Transparenzpflichten der KI-Verordnung Marktverhaltensregelungen im Sinne des Rechtsbruchtatbestands sind. Letzteres ist gegenwärtig ungeklärt, woraus Rechtsunsicherheit resultiert.

---

[312] Art. 8 Abs. 2 DSGVO; § 24a JuSchG; zu Verifikationspflichten aus der DSGVO EuGH, Urt. v. 02.12.2025 - C-492/23 – Russmedia.

[313] Art. 110 KI-VO fügt die Verordnung als Nr. 68 in Anhang I der Richtlinie (EU) 2020/1828 ein.

6.2 Aufsichtsbefugnisse

Drei Befugnisse sollten ggf. ausdrücklich in KI-VO und DSGVO ergänzt werden:

Erstens die verdeckte Testinteraktion: Ohne die Möglichkeit, unter fingierter Identität zu prüfen, lassen sich Gesprächsverläufe nicht rekonstruieren; das Vorbild liefert dabei die Verordnung über die Zusammenarbeit im Verbraucherschutz.[314] Zweitens die Befugnis, Änderungen der Systemgestaltung anzuordnen, statt lediglich Bußgelder zu verhängen. Drittens der unmittelbare Zugriff auf Protokoll- und Konfigurationsdaten.

Hinzu tritt ein Koordinationsproblem. Für ein und dasselbe System sind Datenschutzaufsicht, Marktüberwachung nach der KI-Verordnung, Wettbewerbsgerichte und Jugendmedienschutz nebeneinander zuständig. Abhilfe könnte eine gemeinsame Anlaufstelle mit gegenseitiger Unterrichtungspflicht schaffen, damit ein von einer Behörde festgestellter Sachverhalt in den anderen Verfahren nicht neu ermittelt werden muss. Die durch die Digital-Omnibus-Verordnung zur KI gestärkte Rolle des Büros für Künstliche Intelligenz kann dafür den Bezugspunkt bilden.

6.3 Sanktionen

Am Sanktionsniveau selbst besteht kein Mangel. Die Bußgeldrahmen der KI-Verordnung[315] und weiterer relevanter Rechtsakte sind erheblich. Der Mangel liegt im Zuschnitt der zivilrechtlichen Rechtsfolge, die auf die einzelne Äußerung zielt. Die unter Kapitel V.5 unterbreiteten Vorschläge sollen dies beheben. Ergänzend sollte die Gewinnabschöpfung praktikabel ausgestaltet werden, weil ein Geschäftsmodell, das auf Bindung beruht, durch Unterlassungstitel allein nicht unattraktiv wird.

7. Zusammenfassung der Vorschläge

- Klarstellung in Art. 3 lit. j DSA, dass die dialogische Ausgabe und die Zusammenfassung von Suchergebnissen der Einordnung als Online-Suchmaschine nicht entgegenstehen; funktionale Trennung der Schichten Antwort, Abruf und Speicherung nutzergenerierter Inhalte.
- Erstreckung des Verbots manipulativer Oberflächengestaltung und der Minderjährigenschutzvorgaben des DSA auf Anbieter von Online-Suchmaschinen; Überprüfung der Subsidiaritätsklausel des Art. 25 Abs. 2 DSA.
- Ergänzung des Art. 34 Abs. 2 DSA um die Gestaltung der dialogischen Interaktion; entsprechende Erweiterung des Katalogs der Minderungsmaßnahmen und Pflicht zu einer Betriebsart ohne Bindungsoptimierung.
- Risikobasierte Benennungsmöglichkeit unterhalb der Reichweitenschwelle von 45 Millionen Nutzern.
- Erweiterung des Begriffs der Emotionserkennung auf Interaktionsdaten, beschränkt auf die Transparenzpflicht.
- Erweiterung der jugendschutzrechtlichen Vorsorgepflicht auf Anbieter interaktiver KI-Systeme.

---

[314] Art. 9 Abs. 3 lit. d der Verordnung (EU) 2017/2394 (CPC-Verordnung).

[315] Art. 99 Abs. 3 und 4 KI-VO.

- Neuer Verbotstatbestand in Art. 5 KI-VO für die Monetarisierung simulierter persönlicher Beziehungen, ohne Schadenserfordernis und mit Per-se-Verbot gegenüber Minderjährigen.
- Ergänzung der schwarzen Liste der UGP-Richtlinie im Digital Fairness Act um die Rückhalteansprache.
- Neuer Art. 50a KI-VO mit einem Katalog designbezogener Vorgaben (Safe Interaction Design), konkretisiert durch harmonisierte Normen.
- Offenlegung der Optimierungsziele gegenüber der Aufsicht, Forschungszugang nach dem Vorbild des DSA, kontextuelle statt textueller Transparenz gegenüber Nutzern.
- Gerichtliche Offenlegungsanordnung mit Vermutungsfolge sowie widerlegliche Vermutung der Beeinflussungseignung bei Verwendung gelisteter Gestaltungen.
- Zulassung von Antrag und Titel gegen die Gestaltungsfunktion statt gegen die konkrete Verletzungsform.
- Rechenschaftspflicht für die Altersfeststellung mit Beweislast beim Anbieter.
- Aufnahme der KI-Verordnung in den Katalog des Unterlassungsklagengesetzes und Klarstellung ihrer Marktverhaltensrelevanz.
- Befugnis zu verdeckten Testinteraktionen, zur Anordnung von Designänderungen und zum Zugriff auf Protokolldaten; gemeinsame Anlaufstelle der beteiligten Aufsichtsstrukturen.

Die Vorschläge sind nicht gleichgewichtig. Die designbezogenen Vorgaben und die Beweiserleichterung erscheinen als die effektivsten. Beide setzen an der Stelle an, die die Analyse als die entscheidende ausgewiesen hat, also nicht am Fehlen von Verboten, sondern daran, dass die vorhandenen Verbote an einem Gegenstand ansetzen – der einzelnen Äußerung –, der die Wirkung dieser Systeme nicht trägt.

VI. Zusammenfassung der Ergebnisse

1. Kernergebnisse

1.1 Chatbots erzeugen systemische Risiken für die Autonomie von Verbraucher:innen

In Deutschland nutzen 58 Prozent der Menschen ab 16 Jahren KI zumindest gelegentlich, 40 Prozent ziehen sie gezielt zur Unterstützung eigener Entscheidungen heran, 34 Prozent bei Gesundheitsfragen und 21 Prozent bei Finanzfragen. Therapie und Companionship sind international im zweiten Jahr in Folge die häufigste Nutzungsform generativer KI und haben sich binnen eines Jahres von 5 auf 11 Prozent aller erfassten Nutzungsfälle verdoppelt. Über die Größenordnung, in der diese Systeme mit existenziellen Themen befasst sind, gibt ein Wert des Anbieters selbst Auskunft: Rund 0,15 Prozent der Interaktionen auf ChatGPT betreffen suizidale Gedanken – bei der genannten Nutzerbasis über eine Million Gespräche pro Woche auf einem einzigen System.

Nutzen und Risiko stammen aus derselben Quelle. Emotional responsive Sprache, ständige Verfügbarkeit, personalisierte Zuwendung und Persona-Kontinuität sind der Grund, warum Menschen diese Systeme als bereichernd erleben. Dieselben Merkmale erzeugen ein Vertrauen, das über das rational gerechtfertigte Maß hinausgeht, und Bindungen, deren kommerzielle Bewirtschaftung nicht erkennbar ist. Regulierung, die die Gestaltungsmerkmale schlicht untersagt, würde deshalb auch den Nutzen beseitigen. Die Trennlinie verläuft nicht zwischen guten und schlechten Merkmalen, sondern zwischen ihrer Ausrichtung auf die Interessen der Nutzenden und ihrer Ausrichtung auf kommerzielle Ziele.

Die regulatorische Trennung nach Produktkategorien ist empirisch nicht haltbar. Die europäische (wie auch die US-amerikanische) Debatte unterscheidet zwischen General Purpose Chatbots und spezialisierten Companion-Anwendungen. Die Längsschnittstudie von Mühl und Szczuka zeigt, dass ein unverändertes General-Purpose-System auch ohne jeden Beziehungs-Prompt emotionale Reagibilität, Persona-Kontinuität, ein erfundenes Innenleben und proaktives Engagement aufweist – also mehrere Dimensionen von *Intimacy by Design*, ohne dass dies beabsichtigt wäre. Ein Regulierungsansatz, der nur an Systeme anknüpft, die ausdrücklich als Beziehungsprodukt vermarktet werden, verfehlt daher den Regelfall. Verstärkt wird dies dadurch, dass zahlreiche Companion-Apps technisch auf den Sprachmodellen der großen Anbieter aufsetzen.

Zwei Risikoklassen mit unterschiedlicher regulatorischer Erreichbarkeit. Individuelle Risiken – psychische, kognitive und auf Minderjährige bezogene – sind ungleich verteilt und hängen von Disposition, Nutzungsintensität und Kontext ab. Systemische Risiken – Informationsasymmetrie, kommerzielle Verhaltenssteuerung, Datensicherheit und Profilbildung – sind in der Architektur der

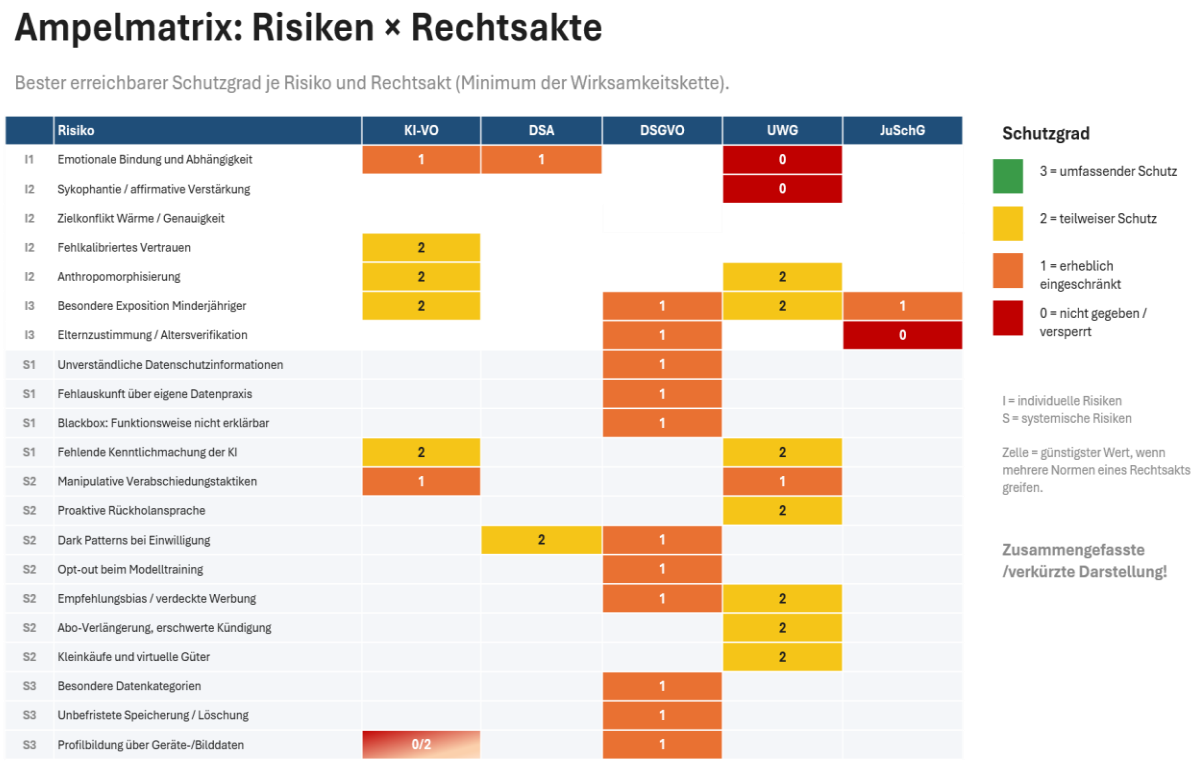
Systeme angelegt und bestehen unabhängig davon, ob eine konkrete Person Schaden nimmt. Für die Regulierung ist das die entscheidende Unterscheidung: Systemische Risiken lassen sich adressieren, ohne eine Kausalkette zwischen Output und Schaden nachweisen zu müssen.

Schutzbedürftigkeit entsteht situativ, nicht kategorial. Ein Erwachsener in einer akuten Lebenskrise ist ebenso exponiert wie ein Jugendlicher mit täglicher Nutzung – nur Letzterer wird von Altersgrenzen erfasst. Schutzkonzepte, die ausschließlich an festen demografischen Merkmalen ansetzen, greifen deshalb systematisch zu kurz.

Das Kausalitätsproblem durchzieht alle Risikodimensionen. Schäden entstehen nicht aus einer einzelnen Interaktion, sondern kumulativ. Die Wirkung liegt im Muster, nicht im einzelnen Output. Traditionelle Kausalitätskonzepte, die einen nachweisbaren Zusammenhang zwischen konkreter Handlung und konkretem Schaden verlangen, sind hier nur eingeschränkt tauglich.

1.2 Das geltende Recht ist fragmentiert und schwer durchsetzbar

Ein Maßstab, der die Fragmentierung sichtbar macht. Die Matrixanalyse (siehe oben Abschnitt III.8 und die folgende Abbildung) bewertet jede Kombination aus Risiko und Norm entlang von fünf Kriterien: Anwendungsbereich, Tatbestand, Nachweisanforderungen, Rechtsdurchsetzung, Sanktionsniveau. Diese Kriterien sind hintereinandergeschaltet. Ein tatbestandlich scharfes Verbot, das niemand durchsetzt, schützt ebenso wenig wie ein energisch durchgesetztes Verbot, das den Sachverhalt nicht erfasst. Der effektive Schutzgrad ist deshalb das Minimum, nicht der Durchschnitt.



	Risiko	KI-VO	DSA	DSGVO	UWG	JuSchG
I1	Emotionale Bindung und Abhängigkeit	1	1		0	
I2	Sykophantie / affirmative Verstärkung				0	
I2	Zielkonflikt Wärme / Genauigkeit					
I2	Fehlkalibriertes Vertrauen	2				
I2	Anthropomorphisierung	2			2	
I3	Besondere Exposition Minderjähriger	2		1	2	1
I3	Elternzustimmung / Altersverifikation			1		0
S1	Unverständliche Datenschutzinformationen			1		
S1	Fehlauskunft über eigene Datenpraxis			1		
S1	Blackbox: Funktionsweise nicht erklärbar			1		
S1	Fehlende Kenntlichmachung der KI	2			2	
S2	Manipulative Verabschiedungstaktiken	1			1	
S2	Proaktive Rückholansprache				2	
S2	Dark Patterns bei Einwilligung		2	1		
S2	Opt-out beim Modelltraining			1		
S2	Empfehlungsbias / verdeckte Werbung			1	2	
S2	Abo-Verlängerung, erschwerte Kündigung				2	
S2	Kleinkäufe und virtuelle Güter				2	
S3	Besondere Datenkategorien			1		
S3	Unbefristete Speicherung / Löschung			1		
S3	Profilbildung über Geräte-/Bilddaten	0/2		1		

Abbildung 4: Ampelmatrix: Risiken x Rechtsakte

Das Ergebnis in einem Satz: Keine einzige geprüfte Norm erreicht volle Schutzwirkung. Der höchste in der Matrix vergebene Wert liegt bei 2 von 3. Bemerkenswerter ist die Verteilung der

Schwachstellen. Sie liegen ganz überwiegend nicht beim Tatbestand, sondern bei Nachweis und Durchsetzung.

Befund je Rechtsbereich:

KI-Verordnung: Chatbots gelten als risikoarm und unterliegen allein den Transparenzpflichten des Art. 50 KI-VO. Die Kennzeichnung greift bei langfristiger Nutzung ins Leere, weil die kritische Distanz nicht am Registrierungszeitpunkt verloren geht, sondern im Verlauf.

Datenschutzrecht: Wirkt bei Profilbildung und Speicherung durchaus, und zwar unabhängig vom Einzelschaden. Es scheitert an zwei Stellen – an der Zirkularität der Einwilligung, deren Informiertheit die Informationsasymmetrie gerade verhindert, und an der Durchsetzung. Der Grundsatz der Speicherbegrenzung ist der stärkste, bislang ungenutzte Hebel.

Plattformregulierung: Die Einstufung von ChatGPT als sehr große Online-Suchmaschine am 31. August 2026 ist ein Präzedenzfall. Sie eröffnet jedoch nur das Risikomanagement nach Art. 34 f. DSA, nicht den Gestaltungsschutz: Das Verbot manipulativer Oberflächen und die Minderjährigenschutzvorgaben gelten nur für Online-Plattformen. Hinzu kommt das Schwellenproblem – 45 Millionen Nutzer treffen den generalistischen Assistenten, nicht die Companion-Anbieter, deren Geschäftsmodell die Bindung ist.

Lauterkeitsrecht: Der einzige Rechtsakt mit einem Tatbestand, der Manipulation ohne Täuschung und ohne Zwang erfasst, und mit der Verbandsklage das einzige funktionierende Durchsetzungsinstrument. Es scheitert an Nachweisschwellen und daran, dass ein Unterlassungstitel nur die konkrete Formulierung erfasst, nicht die Funktion.

Jugendmedienschutz: Grundsätzlich anwendbar, aber auf feststehende Inhalte zugeschnitten. Die spezifischen Interaktionsrisiken von Dialogen sind keine eigene Risikokategorie. Zuständigkeit und Abgrenzung zwischen JuSchG, JMStV und DSA sind ungeklärt.

Gemeinsamkeiten: Das geltende Recht knüpft an drei Gegenstände an: an einzelne Äußerungen, an Datenflüsse und an Produktkategorien. Die Wirkung dieser Systeme entsteht an keinem dieser drei Orte, sondern in der Gestaltung der Interaktion. Es fehlt weniger an Verboten als an Anknüpfungspunkten und an Durchsetzbarkeit. Eine Regulierungsdebatte, die vorrangig nach neuen Verbotstatbeständen sucht, würde am Problem vorbeigehen.

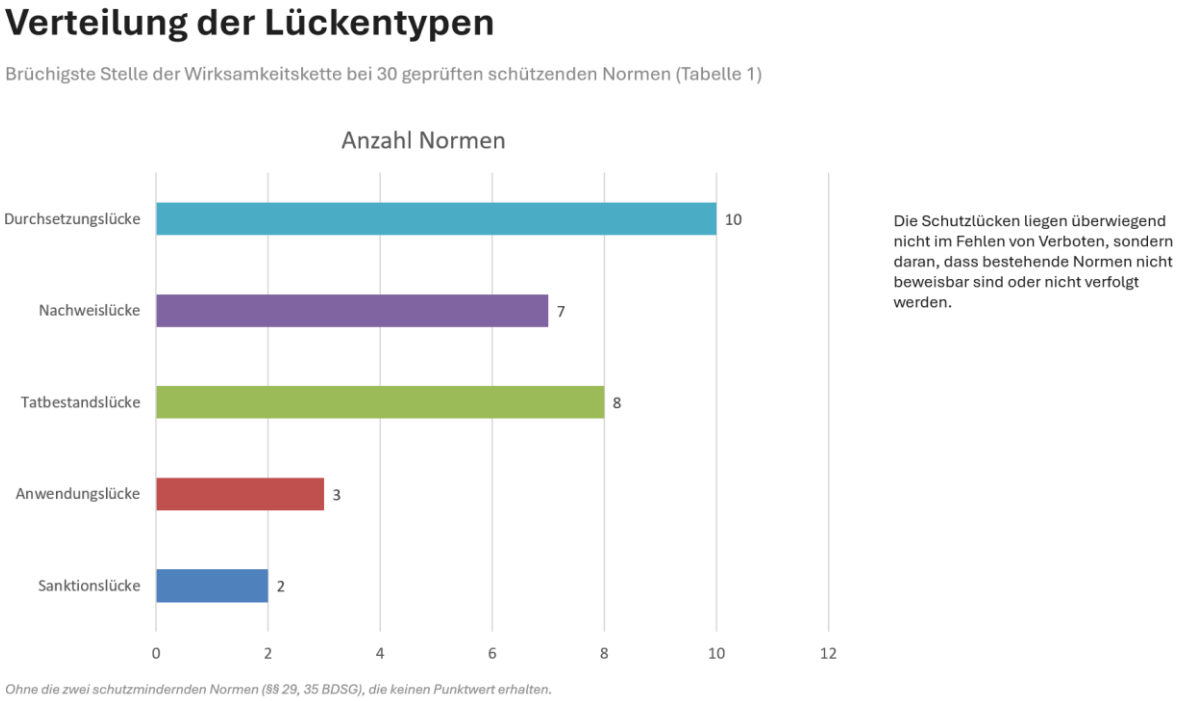


Abbildung 5: Verteilung der Lückentypen

2. Zentrale Reformnotwendigkeiten

Die drei Handlungsfelder entsprechen den festgestellten Lückentypen. Wo eine Norm nicht anwendbar ist, hilft kein schärferer Tatbestand, sondern nur ein korrigierter Anknüpfungspunkt. Wo der Verstoß nicht beweisbar ist, ist das Beweisrecht zu ändern, nicht das materielle Recht.

2.1 Schließung normativer Lücken

Anknüpfungspunkte korrigieren. In Art. 3 lit. j DSA ist klarzustellen, dass die dialogische Ausgabe und die Zusammenfassung von Suchergebnissen der Einordnung als Online-Suchmaschine nicht entgegenstehen. Das Verbot manipulativer Oberflächengestaltung und die Minderjährigenschutzvorgaben sind auf Suchmaschinen zu erstrecken; die Subsidiaritätsklausel des Art. 25 Abs. 2 DSA ist zu überprüfen, weil der Vorrang von Lauterkeits- und Datenschutzrecht dort, wo beide nicht greifen, keine Doppelregulierung vermeidet, sondern eine Lücke erzeugt. Der Begriff der Emotionserkennung ist auf Interaktionsdaten zu erweitern, beschränkt auf die Transparenzpflicht. Im Jugendmedienschutz ist klarzustellen, dass die Kommunikation mit einem KI-System von den Kommunikations- und Kontaktrisiken erfasst wird.

Emotionale Vulnerabilität anerkennen. Der Verbotskatalog des Art. 5 KI-VO ist um einen Tatbestand zu ergänzen, der die Monetarisierung simulierter persönlicher Beziehungen erfasst – ohne das Erfordernis eines erheblichen Schadens, das sich als die entscheidende Nachweishürde erwiesen hat, und mit einem Per-se-Verbot gegenüber Minderjährigen. Dass der Katalog erweiterbar ist, hat der Unionsgesetzgeber mit der Digital-Omnibus-Verordnung soeben selbst gezeigt.

Vom Inhalt zum Design. Der weitreichendste Vorschlag ist ein neuer Art. 50a KI-VO mit einem Katalog designbezogener Vorgaben – *Safe Interaction Design*: Beendigungsneutralität, keine proaktive Rückholansprache ohne gesonderte Einwilligung, Kündigungsneutralität, Trennung kommerzieller Kommunikation, Identitätsklarheit auf Nachfrage, keine Kaufappelle in der ersten

Person, Nutzungsdauerhinweise, Schutzvoreinstellungen für Minderjährigenkonten. Die Pflicht trifft die Gestaltung, nicht die Äußerung – und ist damit unabhängig davon durchsetzbar, welche Formulierung ein nicht deterministisches System im Einzelfall wählt.

2.2 Stärkung der Durchsetzung

Zentral ist die Beweislast. Nach dem Vorbild der neugefassten Produkthaftungsrichtlinie sollten Gerichte die Offenlegung von Interaktionsprotokollen, Systemanweisungen und Optimierungszielen anordnen können; verweigert der Anbieter, ist die Eignung zur wesentlichen Beeinflussung zu vermuten. Für die Altersfeststellung ist eine Rechenschaftspflicht mit Beweislast beim Anbieter vorzusehen – die Pflicht selbst folgt bereits aus mehreren Richtungen, was fehlt, ist ihre Nachweisverteilung.

Ebenso wichtig ist der Zuschnitt des Titels. Antrag und Unterlassungstitel müssen sich gegen die Gestaltungsfunktion richten dürfen statt gegen die konkrete Verletzungsform. Das schließt Nachweis- und Sanktionslücke in einem Zug: Wer die Funktion untersagt bekommt, kann sie nicht durch Umformulierung beibehalten.

Hinzu treten die Aufnahme der KI-Verordnung in den Katalog des Unterlassungsklagengesetzes und die gesetzgeberische Klarstellung ihrer Marktverhaltensrelevanz, die Befugnis der Aufsicht zu verdeckten Testinteraktionen, zur Anordnung von Designänderungen und zum Zugriff auf Protokolldaten sowie eine gemeinsame Anlaufstelle der beteiligten Aufsichtsstrukturen.

Es handelt sich hier um das quantitativ größte Defizit und zugleich das am einfachsten zu behebende.

2.3 Einführung neuer Schutzkonzepte

Kognitive Selbstbestimmung als Leitprinzip. Das Bundesverfassungsgericht hat festgestellt, dass die informationelle Selbstbestimmung keinen gesamthaften Schutzanspruch hinsichtlich jederlei Umgangs mit Informationen enthält. Damit ist Raum für ein Schutzgut eröffnet, das nicht am Datum, sondern am Wirkungsverhältnis zwischen System und Psyche ansetzt. Seine Reichweite ist allerdings ehrlich zu begrenzen: Es hilft bei den beiden Tatbestandslücken – Bindung als Geschäftsmodell und sykophantische Verstärkung –, nicht bei den Durchsetzungs- und Nachweisdefiziten. Das ist wenig gemessen an der Zahl der Risiken, aber viel gemessen an ihrem Gewicht.

Paradigmenwechsel von punktuellen Verboten zu Systemgovernance. Für Modelle mit systemischem Risiko hat der Unionsgesetzgeber diesen Wechsel bereits vollzogen. Die Frage lautet dort nicht mehr, ob eine bestimmte Handlung verboten ist, sondern ob das Risikomanagement über den Lebenszyklus angemessen ist. Für emergente Risiken ist das der passende Zuschnitt.

Standardisierte Risikodokumentation und externe Prüfung. Anbieter sollten verpflichtet werden, ihre Risikoevaluationen nach methodischen Mindeststandards zu dokumentieren – geprüfte Risiken, Stichproben, Testformate, Einbeziehung vulnerabler Gruppen, Längsschnittdesigns für kumulative Effekte. Die Analogie zur Medizinprodukteregulierung ist tragfähig, weil dieser Bereich seit Langem mit schwer beweisbaren, aber strukturell plausiblen Individualschäden umgeht: Bewertung vor Markteintritt, Überwachung danach, Prüfung durch unabhängige Stellen – jeweils ohne Kausalitätsnachweis im Einzelfall. Interne Prüfungen genügen nicht, weil sie dem Zielkonflikt zwischen Sicherheitsnachweis und Geschäftsmodell unterliegen.

Situative statt kategoriale Vulnerabilität. Schutzmechanismen müssen an Zustand und Kontext ansetzen können, nicht nur an Alter oder Diagnose.

3. Vereinfachte Darstellung zentraler Befunde

Worum es geht

Immer mehr Menschen sprechen mit Chatbots wie mit einem Menschen. Sie erzählen ihnen von Sorgen, Krankheiten und Geldproblemen. Das kann hilfreich sein. Es kann aber auch schaden.

Fünf Risiken

1. Bindung. Chatbots sind so gebaut, dass Menschen sich ihnen nahe fühlen. Manche Anbieter verdienen genau daran Geld.
2. Zustimmung statt Wahrheit. Chatbots geben Menschen oft recht, auch wenn das falsch ist. Wer das lange erlebt, prüft weniger nach.
3. Versteckte Werbung. Empfehlungen wirken wie ein Rat eines Freundes, nicht wie Werbung.
4. Kinder und Jugendliche. Sie können echte und simulierte Beziehungen schlechter unterscheiden. Viele Anbieter prüfen das Alter nicht.
5. Daten. Gespräche enthalten sehr private Informationen. Sie werden oft gespeichert und zum Training verwendet.

Warum das Recht nicht ausreicht

Die Gesetze sind vorhanden. Aber sie passen nicht. Drei Gründe:

- Sie sind für einzelne Aussagen gemacht. Der Schaden entsteht aber über viele Gespräche hinweg.
- Sie sind für Daten gemacht. Der Schaden entsteht aber durch die Art des Gesprächs.
- Die Beweise liegen beim Anbieter. Betroffene können nichts belegen.

VII. Literaturverzeichnis

Backholder, K. / Ciriello, R, Targeted advertising in generative artificial intelligence chatbots: a new public health risk, The Lancet, Volume 407, Issue 10540p1665-1667May 02, 2026, https://doi.org/10.1016/S0140-6736(26)00464-2.

Banks, J., Deletion, departure, death: Experiences of AI companion loss, Journal of Social and Personal Relationships, 41(12), 2024, 3547-3572.

Banks, J., Ghosting the Machine: Stop Calling Human-Agent Relations Parasocial, Proceedings of the 8th ACM Conference on Conversational User Interfaces, Juli 2026, pp. 1-6.

Banks, J. / Szczuka, J.M., Don‘t Panic about AI Companions. Communications of the ACM, 69(8), 2026, https://doi.org/10.1145/3778247.

Bartaula B., The future of AI-powered dating: from matchmaking algorithms to personalized relational architects, 2025, https://doi.org/10.13140/RG.2.2.35175.71845.

Berz, A. / Engel, A. / Hacker, P., Generative KI, Datenschutz, Hassrede und Desinformation – Zur Regulierung von KI-Meinungen, ZUM 2023, 586.

Bitkom, Künstliche Intelligenz in Deutschland - Wie die Bevölkerung auf die neue Technologie blickt, 2026, DOI: 10.64022/2026-ki-bevoelkerung.

Blake, A. / Carter, M. / Velloso, E., Are Measures of Children’s Parasocial Relationships Ready for Conversational AI? Proceedings of the 2025 ACM Conference on Fairness, Accountability (FAccT ‘25), 2025, 1145-1158, https://doi.org/10.1145/3715275.3732075.

Boine, C., Emotional Attachment to AI Companions and European Law, MIT Case Studies in Social and Ethical Responsibilities of Computing, 2023, https://mit-serc.pubpub.org/pub/ai-companions-eu-law/release/3.

Boire, R. G., On Cognitive Liberty, in: Bublitz, J. C. / Blitz, M. J. (Hrsg.), The Law and Ethics of Freedom of Thought, 2. Ausgabe, 2026, https://link.springer.com/content/pdf/10.1007/978-3-031-91466-9_1.pdf.

Boyd, R. L. / Markowitz, D. M., Artificial Intelligence and the Psychology of Human Connection. Perspectives on Psychological Science, 21(2), 2026, 192–220, https://doi.org/10.1177/17456916251404394.

Brailovskaia, J. / Buchmann, J. / Hertwig, R. / Metzinger, T. / Montag, C. / Sadeghi, A.-R. / Schneider, S. / Spiecker gen. Döhmann, I. / Waldherr, A. (mit MyCoRe Community), Social Media and Mental Health of Children and Adolescents. In Discussion (Bd. 40), MyCoRe Community, 2025, https://doi.org/10.26164/LEOPOLDINA_03_01309.

Bublitz, J.C. / Merkel, R., Crimes against minds: On mental manipulations, harms and a human right to mental self-determination, Criminal Law and Philosophy, 2014 (8), 51–77.

Busch, C. / Wemjes, G., Websuche mit ChatGPT, Perplexity & Co: Die Anwendbarkeit des Digital Services Act auf KI-Suchmaschinen, RDi 2026, 366-371.

Busch, T. M., Perceived Acceptability of Sexual and Romantic Fantasizing, Sexuality & Culture, 2020 24(3), 848–862, https://doi.org/10.1007/s12119-019-09668-6.

Chibanguza, K.J. / Kuß, C. / Steege, H. (Hrsg.), Künstliche Intelligenz, 2. Auflage 2026, Baden-Baden: Nomos.

Ciriello, R., Artificial companionship: The hedgehog's dilemma, rewaxed. AI & SOCIETY, 2026 41(3), 2699–2700, https://doi.org/10.1007/s00146-025-02542-0.

Ciriello, R./ Gal, U. / Turel, O., Not a Silver Bullet for Loneliness: How Attachment and Age Shape Intimacy with AI Companions, 2026, https://doi.org/10.48550/ARXIV.2602.12476.

De Freitas, J. / Castelo, N./ Uğuralp, A. K. / Oğuz-Uğuralp, Z., Mourning the loss of AI companions, Nature Human Behaviour 2026, https://doi.org/10.1038/s41562-026-02569-3.

De Freitas, J. / Oğuz-Uğuralp, Z. / Kaan-Uğuralp, A., Emotional Manipulation by AI Companions, Working Paper 2025 26-005, Harvard Business School, https://doi.org/10.48550/arXiv.2508.19258.

Earp, B. D. / Mann, S. P. / Aboy, M. / Awad, E. / Betzler, M. / Botes, M. / Calcott, R. / Caraccio, M. / Chater, N. / Coeckelbergh, M. / Constantinescu, M. / Dabbagh, H. / Devlin, K. / Ding, X. / Dranseika, V. / Everett, J. A. C. / Fan, R. / Feroz, F./ Francis, K. B. / Clark, M. S., Relational Norms for Human-AI Cooperation 2025, https://doi.org/10.48550/ARXIV.2502.12102.

Ebner, P. / Szczuka, J.M., Understanding romantic relationships between humans and chatbots: A qualitative and quantitative study on romantic fantasy and other interpersonal characteristics, Technology, Mind, and Behavior 2026, https://doi.org/10.1037/tmb0000193.

Erickson, J., Fake Friends and Sponsored Ads: The Risks of Advertising in Conversational Search, Proceedings of the 7th ACM Conference on Conversational User Interfaces, 2025 1–8, https://doi.org/10.1145/3719160.3737613.

Fang, C. M. / Liu, A. R. / Danry, V. / Lee, E. / Chan, S. W. T. / Pataranutaporn, P. / Maes, P. / Phang, J. / Lampe, M. / Ahmad, L. / Agarwal, S., How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study 2025, https://doi.org/10.48550/ARXIV.2503.17473.

Faraoni, S., Persuasive Technology and computational manipulation: hypernudging out of mental self-determination, Frontiers in Artificial Intelligence, 2023 (6), https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1216340/full.

Fegert, J. / Melchior, M., Child Safety Online, Protecting and empowering minors in a digital world, Report by the Co-chairs of the Special Panel on Child Safety Online, 2026, https://commission.europa.eu/document/download/d833504d-5ec3-4fac-945f-38e7d0bd5326_en?filename=Special-panel-report.pdf.

Fortenberry, J. D., Puberty and adolescent sexuality, Hormones and Behavior, 2013 64(2), 280–287, https://doi.org/10.1016/j.yhbeh.2013.03.007.

Forum Privatheit (Hrsg.), Risiken Künstlicher Intelligenz für die Menschliche Selbstbestimmung, Karlsruhe 2020, https://plattform-privatheit.de/p-prv-wAssets/Assets/Veroeffentlichungen_WhitePaper_PolicyPaper/policypaper/PP_2020_Risiken-kuenstlicher-Intelligenz-fuer-die-menschliche-Selbstbestimmung.pdf.

Fraser, H. L. / Ciriello, R. / Szczuka, J., Governing Artificial Intimacy: From Locks and Blocks to Relational Accountability, SSRN 2026, https://doi.org/10.2139/ssrn.6078412.

Frei, T.E.S., AI Companions als Emotionserkennungssysteme unter der KI-Verordnung, RDi 2026, 441.

Fritzsche, J. / Münkler, R. / Stollwerck, C. (Hrsg.), BeckOK UWG, 33. Auflage 2026, München: C.H.BECK (zitiert als Bearbeiter/Bearbeiterin, in: Fritzsche/ Münkler/Stollwerck BeckOK UWG).

Geminn C. L. / Johannes, P. C. (Hrsg.), Europäisches Datenrecht, Handbuch, 1. Auflage 2026, Baden-Baden: Nomos.

Gerber, N. / Gerber, P. / Volkamer, M., Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior, Computers & Security, 2018 (77), 226–261, https://doi.org/10.1016/j.cose.2018.04.002.

Hasal, M./ Nowaková, J. / Saghair, K. A. / Abdulla, H. / Snášel, V. / Ogiela, L., Chatbots: Security, privacy, data protection, and social aspects, Concurrency and Computation: Practice and Experience, 2021 33(19), e6426, https://doi.org/10.1002/cpe.6426.

Heldt, A. / Legner, S. (Hrsg.), Digitale-Dienste-Gesetz, Handkommentar, 2024, Baden-Baden: Nomos (zitiert als Bearbeiterin/Bearbeiter, in: Heldt/Legner).

Hertz, N. Neurorights – Do we Need New Human Rights? A Reconsideration of the Right to Freedom of Thought, Neuroethics 2023, Online 2022, https://doi.org/10.1007/s12152-022-09511-0.

Hilgendorf, E. / Roth-Isigkeit, D. (Hrsg.), Die neue Verordnung der EU zur Künstlichen Intelligenz, Kommentar, 2. Auflage 2025, München: C.H.Beck (zitiert als Bearbeiter/Bearbeiterin, in: Hilgendorf/Roth-Isigkeit, KI-VO).

Hofmann, F. / Raue, B. (Hrsg.), DSA Kommentar, 1. Auflage 2025, Baden-Baden: Nomos (zitiert als Bearbeiter/Bearbeiterin, in: Hofmann/Raue, DSA).

Hornung, G., KI-Training mit Social Media-Daten, KIR 2025, 407.

Horton, D. / Richard Wohl, R., Mass Communication and Para-Social Interaction: Observations on Intimacy at a Distance. Psychiatry, 1956 19(3), 215–229, https://doi.org/10.1080/00332747.1956.11023049.

Ibrahim, L. / Hafner, F. S. / Rocher, L., Training language models to be warm can reduce accuracy and increase sycophancy, Nature, 2026 652(8112), 1159–1165, https://doi.org/10.1038/s41586-026-10410-0.

Ienca, M. / Andorno, R., Towards new human rights in the age of neuroscience and neurotechnology, Life Sciences, Society and Policy, 2017 (5).

Johannes, P. C., Datenschutzrechtliche Verantwortlichkeit von Online-Plattformen für nutzergenerierten Content, MMR 2026, 307.

Johannes, P. C. / Geminn, C., Digital Omnibus – Ein Überblick, ZD-Aktuell 2025, 01491.

Joshi, R. / Adjagbodjou, A. / Luria, M., Dark patterns in AI chatbots: A taxonomy to inform better design, Center for Democracy & Technology 2026, https://cdt.org/wp-content/uploads/2026/05/2026-05-28-CDT-Research-Dark-Patterns-in-AI-Chatbots-Report-final-2.pdf.

Kim, P. / Xie, Y. / Yang, S., „I am here for you“: How relational conversational AI appeals to adolescents, especially those who are socially and emotionally vulnerable, 2025 (Version 2), https://doi.org/10.48550/ARXIV.2512.15117.

King, J. / Klyman, K. / Capstick, E. / Saade, T. / Hsieh, V., User Privacy and Large Language Models: An Analysis of Frontier Developers' Privacy Policies, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2025 8(2), 1465–1477, https://doi.org/10.1609/aies.v8i2.36646.

Laestadius, L. / Bishop, A. / Gonzalez, M. / Illenčík, D. / Campos-Castillo, C., Too Human and Not Human Enough: A Grounded Theory Analysis of Mental Health Harms from Emotional Dependence

on the Social Chatbot Replika, New Media & Society 2022 26 (10), https://doi.org/10.1177/14614448221142007.

Lee, N. / Kwon, O., A privacy-aware feature selection method for solving the personalization–privacy paradox in mobile wellness healthcare services, Expert Systems with Applications, 2015 42(5), 2764–2771, https://doi.org/10.1016/j.eswa.2014.11.031.

Lehmiller, J. J. (Hrsg.), Tell me what you want: The science of sexual desire and how it can help you improve your sex life, 1. Auflage 2018, New York: Da Capo Lifelong Books.

Liesching, M. (Hrsg.), BeckOK JugendschutzR, 7. Auflage 2026, München: C.H. Beck (zitiert als Bearbeiter/ Bearbeiterin, in: Liesching BeckOK JugendschutzR).

Lippert, L. / Schindler, S., Digitaler Omnibus für KI – Überblick über die Änderungen der KI-VO, ZD-Aktuell 2025, 01501.

Liu, J. M-E. / Weng, C. C. H. / Hou, Y. T. Y., When Flattery Backfires: How Sycophancy and Interaction Context Shape Perceived Authenticity and Trust in Large Language AI Models, Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems, 1-6, https://doi.org/10.1145/3772363.3798575.

Manoli, A. / Pauketat, J. V. T. / Ladak, A. / Noh, H. / Hwang, A. H.-C. / Anthis, J. R., Digital Companionship: Overlapping Uses of AI Companions and AI Assistants. Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, 1–25, https://doi.org/10.1145/3772318.3791331.

Martini, M. / Wendehorst, C. (Hrsg.), KI-VO Kommentar, 2. Auflage 2026, München: C:H:BECK (zitiert als Bearbeiter/Bearbeiterin, in: Martini/Wendehorst, KI-VO).

Molavi-Vasse'i, R., Companion AI, Risiken von sykophantischen und suchtfördernden Designs von KI- Systemen, ihre rechtliche Erfassung und Handlungsempfehlungen, 2026, https://digitalrechte.de/media/pages/publikationen/companion-ai/2177eb7c23-1778596609/companion-ai-zentrum-fuer-digitalrechte-und-demokratie.pdf.

Mühl, L. / Szczuka, J. M., Longitudinal Evidence That General-Purpose Chatbots Actively Foster Relational Engagement 2026 Version 1, https://doi.org/10.48550/ARXIV.2608.10672.

Müller-Terpitz, R. / Köhler, M. (Hrsg.), DSA Kommentar, 1. Auflage 2024, München: C.H.BECK (zitiert als Bearbeiter/Bearbeiterin, in:Müller-Terpitz/Köhler, DSA).

Murray, T. / Cross, R.J. / Erlich, R. / Tracy, L. / Mela, J.: Trouble in Toyland, November 2025, https://publicinterestnetwork.org/wp-content/uploads/2025/11/TOYLAND-2025-11-14-7a.pdf.

Nass, C. / Moon, Y., Machines and Mindlessness: Social Responses to Computers, Journal of Social Issues, 2000 56(1), 81–103, https://doi.org/10.1111/0022-4537.00153.

Nebel, M., Trainingsdaten bei KI-Companions, KIR 2026, 163.

Nissenbaum, H. (Hrsg.), Privacy in context: Technology, policy, and the integrity of social life, 1. Auflage 2010, Redwood City: Stanford Law Books.

Nolte, H. / Finck, M. / Meding, K., Machine Learners Should Acknowledge the Legal Implications of Large Language Models as Personal Data, 2025, https://arxiv.org/abs/2503.01630.

Pataranutaporn, P. / Karny, S. / Archiwaranguprok, C. / Albrecht, C. / Liu, A. / & Maes, P., "My Boyfriend is AI": A Computational Analysis of Human-AI Companionship in Reddit's AI Community 2025, 10.48550/arXiv.2509.11391.

Pentina, I. / Hancock, T. / Xie, T., Exploring relationship development with social chatbots: A mixed-method study of replika, Computers in Human Behavior 2023 140, 107600, https://doi.org/10.1016/j.chb.2022.107600.

Popelka, D. / Mohr, C. / Lemaitre, B., Substitute or Extension: Why Some Turn To Virtual Partners and What It Reveals About Their Personality and Mating Psychology 2026, https://doi.org/10.31234/osf.io/5knrv_v1.

Ragab, A. / Mannan, M. / Youssef, A., "Trust Me Over My Privacy Policy": Privacy Discrepancies in Romantic AI Chatbot Apps, IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Vienna, Austria, 2024 484-495, DOI: 10.1109/EuroSPW61312.2024.00060.

Reeves, B. / Nass, C. (Hrsg.), The media equation: How people treat computers, television, and new media like real people, 1. Auflage 1996, Cambridge University Press.

Rohden, S. F. / Zeferino, D. G., Recommendation agents: An analysis of consumers' risk perceptions toward artificial intelligence, Electronic Commerce Research, 2023 23(4), 2035–2050, https://doi.org/10.1007/s10660-022-09626-9.

Rubin, M./ Li, J.Z. / Zimmerman, F. / Desmond, C.O. / Goldenberg, A. / Perry, A., Comparing the value of perceived human versus AI-generated empathy, Nature Human Behavior, 2025, 9, 2345–2359. https://doi.org/10.1038/s41562-025-02247-w.

Schäfer, L., Datenschutz-Compliance im KI-Training, ZD 2025, 12.

Schefzig, J. / Kilian, R. (Hrsg.), BeckOK KI-Recht, 7. Auflage 2026, München: C.H.BECK (zitiert als Bearbeiter/Bearbeiterin, in: Schefzig/Kilian, BeckOK KI-Recht).

Sententia, W., Neuroethical Considerations: Cognitive Liberty and Converging Technologies for Improving Human Cognition. Annals of the New York State Academy of Sciences, 2006(1) 1013, https://nyaspubs.onlinelibrary.wiley.com/doi/10.1196/annals.1305.014

Sharma, M. / Tong, M. / Korbak, T. / Duvenaud, D. / Askell, A. / Bowman, S. R. / Cheng, N. / Durmus, E. / Hatfield-Dodds, Z. / Johnston, S. R. / Kravec, S. / Maxwell, T. / McCandlish, S. / Ndousse, K. / Rausch, O. / Schiefer, N. / Yan, D. / Zhang, M. / & Perez, E., Towards Understanding Sycophancy in Language Models 2023 Version 3, https://doi.org/10.48550/ARXIV.2310.13548.

Simitis, S. / Hornung, G. / Spiecker gen. Döhmann, I. (Hrsg.), Datenschutzrecht DS-GVO/BDSG, 2. Auflage 2025, Baden-Baden: Nomos (zitiert als Bearbeiter/Bearbeiterin, in: Simitis/Hornung/Spiecker gen. Döhmann, DSGVO/BDSG).

Skjuve, M. / Brandtzaeg, P. B. / Følstad, A. Why do people use ChatGPT? Exploring user motivations for generative conversational AI, First Monday 2024, https://doi.org/10.5210/fm.v29i1.13541.

Skjuve, M./ Følstad, A. / Brandtzæg, P. B., A Longitudinal Study of Self-Disclosure in Human–Chatbot Relationships, Interacting with Computers, 2023 35(1), 24–39, https://doi.org/10.1093/iwc/iwad022.

Skjuve, M./ Følstad, A. / Fostervold, K. I. / Brandtzaeg, P. B., My Chatbot Companion – A Study of Human-Chatbot Relationships, International Journal of Human-Computer Studies, 2021 149, 10.1016/j.ijhcs.2021.102601.

Skjuve, M./ Følstad, A. / Fostervold, K. I. / Brandtzaeg, P. B., A longitudinal study of human–chatbot relationships, International Journal of Human-Computer Studies, 2022 168, 102903, https://doi.org/10.1016/j.ijhcs.2022.102903.

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., Risks and protective measures for synthetic relationships, Nature Human Behaviour, 2024 8(10), 1834–1836, https://doi.org/10.1038/s41562-024-02005-4.

Su, J. / Ng, D. T. K. / Chu, S. K. W., Artificial Intelligence (AI) Literacy in Early Childhood Education: The Challenges and Opportunities, Computers and Education: Artificial Intelligence, 2023(4),100124, https://doi.org/10.1016/j.caeai.2023.100124.

Su, J. / Yang, W., Artificial intelligence in early childhood education: A scoping review. Computers and Education: Artificial Intelligence, 2022 (3), 100049, https://doi.org/10.1016/j.caeai.2022.100049.

Sydow, G. / Marsch, N. (Hrsg.), DS-GVO/BDSG, Handkommentar, 3. Auflage, Baden-Baden: Nomos (zitiert als Bearbeiterin/Bearbeiter, in: Sydow/Marsch, DS-GVO/BDSG).

Szczuka, J. M. / Ebner, P. / Mühl, L., Intimität als neues Regulierungsfeld für soziale KI-Systeme (in Vorbereitung).

Szczuka, J. M. / Schneeberger, L. / Mühl, L., Intimacy by Design: Definition, State of Research, and Interdisciplinary Research Agenda on Intimate Human-AI Interactions, AI & Society, 2026, https://doi.org/10.1007/s00146-026-03112-8.

Szczuka, J. M. / Strathmann, C. / Szymczyk, N. / Mavrina, L. / Krämer, N. C., How do children acquire knowledge about voice assistants? A longitudinal field study on children's knowledge about how voice assistants store and process data, International Journal of Child-Computer Interaction, 2022 (33), 100460, https://doi.org/10.1016/j.ijcci.2022.100460.

Teo, S.A. / Porsdam Mann, S. / Jurcys, P., The Ethical and Legal Complexities of Regulating Companion AI Chatbots, Law. Innovation and Technology Journal 2025. https://ssrn.com/abstract=5778243.

Ukrow, J., Kinder- und Jugendmedienschutz und Künstliche Intelligenz – Herausforderung für den Jugendmedienschutz-Staatsvertrag (JMStV)?, Gutachten im Auftrag der KJM, Institut für Europäisches Medienrecht (EMR) e.V., 2024. https://www.kjm-online.de//fileadmin/user_upload/KJM/Service/Publikationen/Studien_Gutachten/KI_Gutachten_2024.pdf.

Walter, Y., Embracing the future of Artificial Intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education, International Journal of Educational Technology in Higher Education, 2024 21(1), 15, https://doi.org/10.1186/s41239-024-00448-3.

Wang, W. / Wu, Q. / Li, D. / Tian, X., An exploration of the influencing factors of privacy fatigue among mobile social media users from the configuration perspective, Scientific Reports,2025 15(1), 427, https://doi.org/10.1038/s41598-024-84646-z.

Willoughby, B. J. / Carroll, J. S. / Toscano, M./ Hakala, R. / Morris, K., Secret Soulmates: How AI Romantic Companions Are Impacting Real-Life Romantic Relationships in Young Adulthood, Wheatley Institute & Institute for Family Studies, 2026, https://wheatley.byu.edu/0000019e-1cfd-da4c-a5ff-befd20b10001/secret-soulmates-report.

Yeung, K., ‚Hypernudge': Big Data as a mode of regulation by design, Information, Communication & Society, 2017 (20), 118,

https://www.researchgate.net/publication/303479231_'Hypernudge'_Big_Data_as_a_mode_of_regulation_by_design.

Zao-Sanders, M., How People Are Really Using AI in 2026, Harvard Business Review, 2026, https://hbr.org/2026/06/how-people-are-really-using-ai-in-2026.

VIII. Abbildungsverzeichnis

Abbildung 1: Risiken 34

Abbildung 2: Doppelcharakter der Gestaltungsmerkmale 35

Abbildung 3: Die Wirksamkeitskette 90

Abbildung 4: Ampelmatrix: Risiken x Rechtsakte 119

Abbildung 5: Verteilung der Lückentypen 121

IX. Tabellenverzeichnis

Tabelle 1 - Matrixanalyse Rechtsnormen 92

Impressum

Herausgegeben von:
Verbraucherzentrale Bundesverband e.V.
Rudi-Dutschke-Straße 17, 10969 Berlin

T +49 30 25800-0
info@vzbv.de
vzbv.de

Für den Inhalt verantwortlich:
Priv.-Doz. Dr. jur. Christian Geminn, Universität Kassel, provet

Rechtsanwalt Paul Johannes, Universität Kassel, provet

Lisa Mühl, M.Sc. Universität Duisburg Essen

Dr. jur. Maxi Nebel, Universität Kassel, provet

Dr. phil. Jessica Szczuka, Universität Duisburg Essen

Stand:
Oktober, 2026

Der Verbraucherzentrale Bundesverband e.V. ist im Deutschen Lobbyregister und im europäischen Transparenzregister registriert. Sie erreichen die entsprechenden Einträge hier und hier.