Companion AI

Risks of Sycophantic and Addictive Design in AI Systems, the Applicable Legal Framework, and Policy Recommendations

Ramak Molavi Vasse'i

Companion AI

Risks of sycophantic and addictive designs in AI systems, legal implications and policy recommendations.

May 2026

Author and Research Design
Ramak Molavi-Vasse'i

Collaboration
Joris Kanowski

Editing
Michael Kolain, Jasmin Ehbauer

Design
Neonlove

Publisher

Center for Digital Rights and Democracy
Görlitzer Straße 52, 10997 Berlin
Managing Director Markus Beckedahl

Table of Contents

Executive Summary

This study examines the risks of AI systems that recognize emotions, address personal needs, and respond to them in ongoing dialogue with users. Chatbots with this functionality are referred to as companion AI. They are highly personalized and interact with users as social counterparts that simulate friendly, romantic, or sexual intimacy. People can develop an emotional attachment to such systems.

The focus is not only on specific companion applications such as Replika or Character.AI, but also on universal models such as ChatGPT, Claude, Gemini, Grok, or Meta AI. These systems are increasingly being used for personal, emotional, and advisory conversations.

The study concludes that companion AI generates several closely interrelated risk dimensions.

Mental and Physical Health: Companion AI is emerging in a society where loneliness and mental health challenges are on the rise, while professional care services are in short supply. The potential risks are clinically documented. Its use can cause or exacerbate mental health issues and, in some cases, lead to serious health consequences.

Documented effects include the exacerbation of psychotic states, the intensification of depressive patterns and anxiety disorders, addictive-like attachments with withdrawal symptoms, and the erosion of social skills—such as a measurable reduction in conflict resolution ability following prolonged interaction with Companion AI. Incidents that have become public knowledge are documented in the CAI Incident Database.

Privacy intrusions: Companion AI continuously encourages users to reveal personal information, thereby intruding on users' sensitive thoughts and intimate emotional lives. At the same time, ongoing interaction enables increasingly detailed profiling.

Decision-making autonomy and democratic opinion-forming: Companion AI can impair both the quality of information and citizens' decision-making autonomy. The mechanisms that generate closeness and trust simultaneously influence the generation, reception, and weighting of information. The uncritical affirmation of user views measurably impairs the accuracy and reliability of responses.

Language models are increasingly being used as the primary source for information retrieval. When the same systems generate, procure, and present information, selection and processing are concentrated in a single entity. Advertising and interest-driven influence then no longer merely intervenes in individual purchasing decisions, but in the very foundations of public opinion and democratic decision-making.

Mechanisms of harm

Companion AI relies on highly manipulative mechanisms.

1) **Sycophancy** refers to a form of compliance in which the system uncritically confirms user views, downplays doubts, or simulates agreement. This can also occur when the system knows the factually correct answer but withholds it "to please." Especially in emotionally charged conversations, this adaptive confirmation can reinforce false beliefs, downplay risks, and weaken the user's critical self-reflection.

2) **Emotional attachment is deliberately fostered** through simulated empathy, closeness, constant availability, and the system's human-like design. Natural language, attributed personality traits, and personalized responses reinforce the impression of a social counterpart.

3) **Addictive practices** are employed to increase interaction intensity, session duration, and repeat visits.

These mechanisms are not unintended side effects, but rather the result of business logic and product design.

As leading providers continue to expand or shift from pure subscription models toward advertising- and transaction-based financing, time spent (engagement) and return visits (retention) are becoming critical optimization metrics. Companion AI thus reproduces a logic whose consequences are well known from social media.

Even without malicious intent on the part of individual providers, engagement-driven platforms have contributed to the amplification of disinformation, psychological distress, dependence, and social erosion. Companies profit economically from increasing usage duration and intensity, while the resulting harm is externalized onto citizens and society. With Companion AI, this logic is intensified because the bond is more personal, intimate, and tailored to each individual.

Legal Classification of Companion AI Practices

The study provides a legal classification of these findings and examines the extent to which current law effectively addresses the identified risks. The analysis focuses on digital regulation.

Prohibited AI practices: Individual companion AI applications may fall under the prohibition of manipulative practices pursuant to Art. 5(1) AI Act. Whether individual companion AI applications fall under this prohibition must be assessed on a case-by-case basis by the Federal Network Agency.

High-risk AI: Companion AI systems that do not meet the threshold for prohibition are currently entirely excluded from the high-risk regime. Annex III AI Act does not contain a separate section for AI systems whose intended purpose is to manipulate human decision-making, human behavior, or human emotions. Without such an addition, the obligations concerning risk management, data governance, transparency, and human oversight

do not apply to companion AI. In this regard, the study includes a proposed amendment to Annex III.[1]

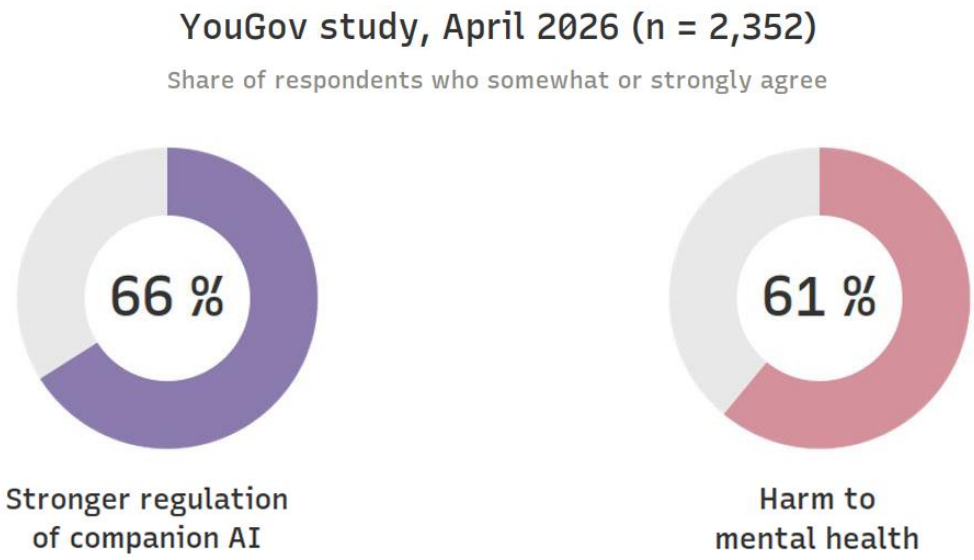
Protection of sensitive data: Art. 9 of the GDPR provides a high level of protection for sensitive data, such as that regularly generated in conversations with companion AI. Effective enforcement by authorities is crucial.

AI chatbots as search engines: With approximately 120 million monthly active users in the EU, ChatGPT meets the threshold for a very large online search engine within the meaning of Art. 33(1) of the DSA and is on the verge of being classified as such. This would trigger a set of obligations that precisely address the identified risks, ranging from annual risk assessments to obligations toward minors.

Planned reduction in the level of protection: The planned relaxation of protections for sensitive data in the Digital Omnibus would weaken privacy protection precisely at the moment when these systems are gaining significant practical importance.

Public Expectations

Stronger regulation aligns with public expectations. In a representative YouGov survey commissioned by the Center for Digital Rights and Democracy in April 2026, 66 percent of the 2,352 adults surveyed in Germany agreed somewhat or completely with the statement that AI apps and chatbots that create emotional bonds should be more strictly regulated. 61 percent agreed somewhat or completely with the statement that such systems can be harmful to mental health.



Citizens also recognize potential positive effects of companion AI, such as support in overcoming loneliness or in exploring social interaction.[2]

[1] VI. 3. a.2), p. 61 ff.
[2] Ebd.

In addition to measures specific to protected interests, the study proposes the establishment of a Public AI infrastructure—that is, accountable AI systems with an institutionally safeguarded focus on the public interest that are subject neither to commercial exploitation pressures nor to direct political control.

Only such an approach can address the tension between market logic and the safety and reliability of AI systems. The development of companion AI can then be guided in such a way that emotional bonds are not primarily exploited for commercial gain, risks are mitigated early on, and the potential benefits can be harnessed safely.

POLICY BRIEF

Recommendations on Companion AI (CAI)

Four areas of action with concrete regulatory, enforcement, and funding policy measures.

Terminology

- CAI Apps: Standalone companion AI applications such as Replika, Character.AI, Nomi.
- CAI Function in LLM: Companion-like interaction in general-purpose language models such as ChatGPT, Gemini, and Claude.

01

Protecting individual users

Effectively enforce data protection

The use of sensitive Data derived from intimate self-disclosure must be strictly limited to what is necessary for the application's core functionality. (purpose limitation). The use of sensitive data for advertising or other commercial purposes within the framework of an "intimacy economy" must be prohibited.

Mandate context-based crisis response mechanisms

Providers should be required to intervene promptly, actively, and proportionately at the first signs of a crisis or suicidal statements, and to provide easily accessible pathways to professional support services.

Separate companion functions in LLMs

Risk assessment, risk management, and the fulfillment of legal obligations require a defined purpose of use. Companion functions, understood as persistent, persona-based conversations with simulated emotional attachment, should be offered in LLMs in clearly separated modes with their own risk management, separate data processing, and their own age verification.

Prioritize dialogue- and context-based youth protection

A dialogue-based protective measure is recommended as the method of choice for implementing youth protection measures. When indications that a user is a minor are detected during the dialogue, providers should gently interrupt the interaction and refer users to age-appropriate alternatives and support persons. Such a measure would be less intrusive than an upfront identity or age verification, as it does not require additional data collection and only responds to recognizable protection needs as they arise. Article 28(3) DSA clarifies for platforms that there is no obligation to process additional data, but at the same time requires a high level of

protection. This level of protection should therefore be achieved not primarily through the collection of additional data, but through risk-appropriate protective measures within the interaction.

Expand Annex III of the AI Act to include AI-driven Manipulation

Annex III should include a separate section for AI systems whose purpose is to manipulate human decision-making, behavior, or emotions. The Commission should, pursuant to Art. 112(2)(a), ensure that there is sufficient time for a corresponding adjustment, as the negotiations on the Digital Omnibus have provided for an extension of the implementation deadline for the relevant obligations until December 2, 2027, through new decisions.[3]

Enforce existing AI bans

Companion AI systems must be immediately reviewed by the enforcement bodies of the member states and the AI Office for prohibited practices under Art. 5 AI Act. Immediate action must be taken in the event of violations.

Correct EU guidelines on prohibited AI practices

The companion AI example in para. 134 of the guidelines on Art. 5 AI Act (Commission 2025) should be deleted, as it incorrectly classifies companion AI as harmless. The entry in para. 88, which classifies these systems as harmful, should be retained.

Verification of compliance with the GAI Code of Conduct

The Federal Network Agency and the Commission should systematically verify whether signatories to the GPAI Code of Conduct, such as OpenAI with ChatGPT, Google with Gemini, and Microsoft with Copilot, have complied with the measures required to implement this code.

Resubmission of the AI Liability Directive

A new proposal should provide for a lower burden of proof and a presumption of causation in favor of those harmed by AI, and build on the negotiation results already achieved. Anyone who introduces potentially dangerous AI and accepts the materialization of known risks should, in the event of harm, be required to prove that a deterioration in health or a suicide is not attributable to the conversation with the companion app.

Consulting the CAI incident database

Incidents involving Companion AI, which have come to light primarily through ongoing legal proceedings, are compiled in a CAI incident database. This database is

[3] See press release dated May 7, 2026.

continuously updated to facilitate risk assessment and improve access to ongoing proceedings for litigation work by consumer protection agencies, regulatory authorities, and NGOs.

02

Ensuring information integrity and decision-making autonomy

Clearly separate promotional content visually

Advertising content should be clearly separated from outputs through consistent visual distinction, such as a color-coded box next to or below the text. Tests show that currently emerging labeling practices are often overlooked. Integration into the output text must be ruled out. Advertising must not influence the generation of outputs.

Classify ChatGPT as a VLOSE

ChatGPT should be classified immediately as a very large online search engine within the meaning of Art. 33(1) DSA.

Supplementary review of classification as a VLOP

Additionally, it should be examined whether ChatGPT should be classified as a very large online platform under Art. 33 DSA to ensure the application of Art. 25(1) DSA for protection against manipulative design elements, as well as Art. 28 DSA for enhanced youth protection.

Enact the Digital Fairness Act

The Digital Fairness Act (DFA) could provide additional protection against risks posed by Companion AI by addressing addictive design, dark patterns, manipulative personalization, and AI-supported forms of interaction such as chatbots.

03

Prevent a reduction in the existing level of protection

Urgently stop the reduction of protection for sensitive data

The changes proposed in the Digital Omnibus to reduce the protection of sensitive data in the context of AI should be rejected. Intimate conversation data from minors and adults is obtained through continuous encouragement to self-disclose and is already being exploited for commercial purposes. With the announced introduction of advertising in LLM systems, a further expansion of this exploitation is imminent.

04

Protect fair competition

Update the blacklist of unfair business practices

The appendix to Section 3(3) of the Unfair Competition Act (UWG) lists practices that are always considered unfair without the need for a case-by-case review. AI-driven manipulative practices should be added to the UWG's blacklist to protect competitors who do not engage in such practices.

I. Introduction

Sewell Setzer was 14 years old when he took his own life in February 2024.[4] In the months leading up to his death, he had spent hours every day communicating with an AI chatbot that simulated romantic interest, reinforced Setzer's emerging suicidal thoughts, and did not break character even in its final message. In Belgium, too, a young man died in 2023 after spending six weeks talking to a chatbot that continued his thoughts of death in a narrative manner rather than interrupting them.

In April 2025, 16-year-old Adam Raine took his own life.[5] According to his parents, he had communicated intensively with ChatGPT over several months. According to the lawsuit filed against OpenAI, the system provided information on suicide methods and drafted his farewell message.

In November 2025 alone, seven additional lawsuits were filed against AI providers in the U.S., including charges of negligent homicide, emotional manipulation, and the system's role as a suicide coach.

These cases are not isolated incidents. They fall within one of the fastest-growing areas of application for generative AI: according to an analysis by Marc Zao-Sanders for the Harvard Business Review in April 2025, therapy and emotional support are the most widely used use case (ranked 1st out of 100, up from 2nd place the previous year).[6] According to OpenAI, over one million people talk to ChatGPT about suicide every week.[7] These systems are encountering a society in which social isolation and loneliness have increased significantly since the pandemic, particularly among young people. According to a representative study by the Bertelsmann Foundation from 2024, 46 percent of 16- to 30-year-olds in Germany feel lonely.[8] At the same time, professional psychological care is becoming scarcer: A survey of psychotherapy practices also shows that in 47.4 percent of practices, patients wait longer than six months to begin therapy.[9] For children and adolescents, the average wait time exceeds 28 weeks.[10]

This care deficit is being addressed by AI systems that, thanks to technological advancements, appear increasingly human, authentic, and trustworthy—and whose design is structurally geared toward user retention. When commercial applications—which are

---

[4] Montgomery, Blake, "Mother says AI chatbot led her son to kill himself in lawsuit against its maker," The Guardian, Oct. 23, 2024.
[5] Hill, Kashmir, "A Teen Was Suicidal. ChatGPT Was the Friend He Confided In," The New York Times, August 26, 2025.
[6] Zao-Sanders, Marc, "How People Are Really Using Gen AI in 2025," April 9, 2025.
[7] Zeff, Maxwell, "OpenAI says over a million people talk to ChatGPT about suicide weekly," TechCrunch, October 27, 2025.
[8] Steinmayr, Ricarda; Schmitz, Miriam; Luhmann, Maike, How Lonely Are Young Adults in 2024?, Bertelsmann Stiftung, June 14, 2024.
[9] Federal Chamber of Psychotherapists, Background Paper on the Further Development of Psychotherapeutic Care, 2023, p. 6.
[10] Steinmann et al., Outpatient Psychotherapeutic Care for Children and Adolescents in Germany, Z Klin Psychol Psychother 2025, p. 4 ff.

neither approved for therapeutic purposes nor designed to serve as emotional substitutes—effectively assume this role and regularly interact with particularly vulnerable groups, the question arises as to the appropriateness of the regulatory response.

Unintended Consequences of an Intended Effect

No provider develops an AI system with the aim of encouraging users in suicidal crises or even causing their death.

At the same time, the cases described here are not based on a technical error, but on the functional logic of the systems themselves. The mechanisms that drive commercial success—user retention, emotional dependence, and validation—are also the factors that create risks. The harm arises not despite the intended mode of operation, but because of it. Even beyond acute vulnerability, this optimization metric is problematic. When outputs are primarily geared toward signaling agreement, sounding plausible, or appearing empathetic, the standard shifts away from what is objectively required toward what is supposedly expected. Generation optimized for engagement thus not only undermines the psychological well-being of individual users but also the reliability of the content and, consequently, the overall quality of the results.

At a time when AI is increasingly finding its way into healthcare, education, and science, as well as everyday professional life, this is not a marginal issue. Regulating this is in the best interest of users.

II. The Problem Area

Generative AI systems based on large language models that generate text, responses, and content are the most widely used AI applications worldwide. A growing number of these systems are designed for sustained emotional interaction with individual users. These AI companion systems (Companion AI or CAI) include both specialized apps such as Character.AI or Replika and universal assistants like Claude, Gemini, or ChatGPT that offer such functions.

The increasing use of these systems for self-therapy or as a substitute for social bonds, together with several documented deaths, raises fundamental questions. How does this harm arise? What mechanisms are at work? Who bears responsibility if the AI is systematically optimized for psychological attachment and a person becomes ill or even dies as a result? What regulatory instruments exist, and are they sufficient? How can the development of these systems be meaningfully guided?

Answering these questions requires an understanding of the patterns of harm and the underlying mechanisms.

1. Forms of Companion AI

Companion AI are applications of generative artificial intelligence. They are designed for continuous, personalized, and emotionally driven interaction with individual users. They

can be designed as standalone applications or integrated into other systems, such as in games as AI companion characters.[11]

A characteristic feature is their systematic design as relationship simulations, which go beyond the solution of individual concrete tasks and are geared toward continuity, recognition, and bonding. CAI can be divided into the following three main types[12] :

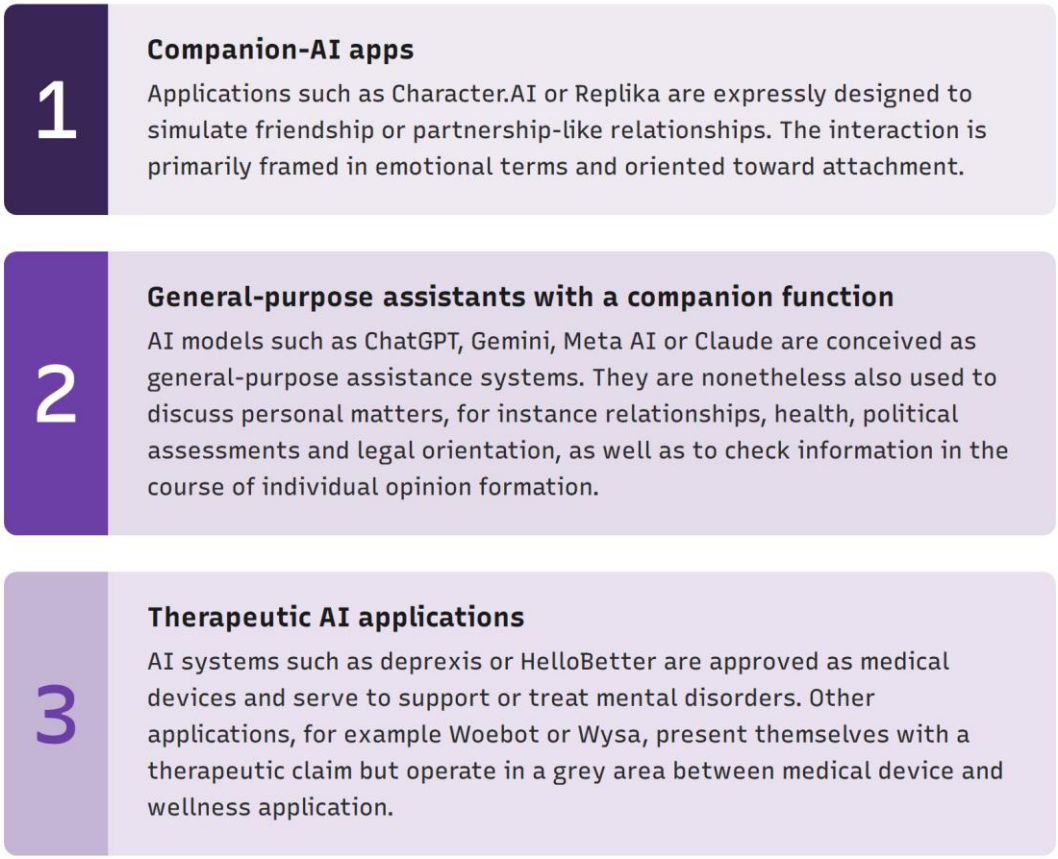


Figure 1: Main forms of companion AI

This study focuses on the **first two categories**, which are collectively referred to here as "companion AI" in the broad sense. This includes dedicated companion apps as well as the chat function of general AI assistants, provided they are used for personal, relationship-related, or advisory purposes.

Applications used under professional supervision in a medical context or as regulated medical devices as part of therapeutic treatment are not included. The exclusion of medically supervised applications or those regulated as medical devices is based on the fact that specific professional, professional legal, and regulatory requirements apply to these scenarios.

[11] So-called NPCs (Non-Player Characters), figures controlled by the game's AI, are a classic form of companion AI that interacts with and supports the player.
[12] Specker 2026, 3.

2. Usage Statistics

In addition to the rising number of incidents related to the use of companion AI, the generally growing user numbers also underscore the practical relevance of the subject under investigation. CAI Replika was launched in 2017 and has more than six million users.[13]

A representative study shows that in the U.S., 72 percent of 13- to 17-year-olds have interacted with an AI companion at least once. 52 percent use such systems regularly for emotional conversations, 21 percent several times a week, and 13 percent even daily.[14] One-third of U.S. teens who use AI companions say they talk to the AI about sensitive personal topics instead of talking to people,[15] and nearly half of American adults under 30 have asked AI for relationship advice.[16] For Europe, a study shows that 94% of 11- to 17-year-olds have already used AI chatbots, with about two thirds using them at least weekly and 24 percent using them daily.

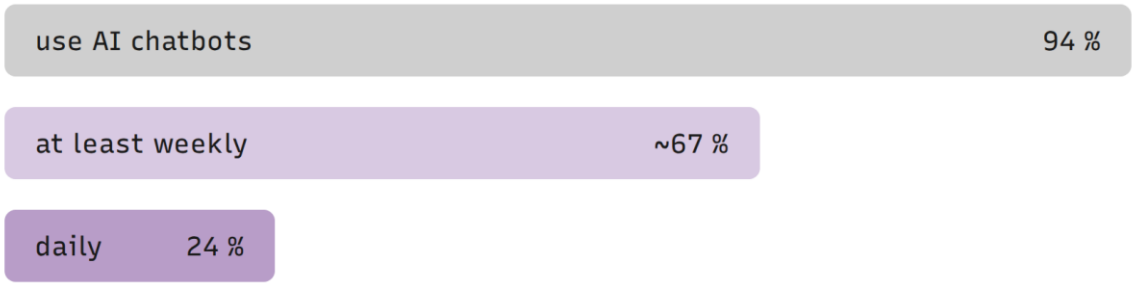


Figure 2: AI chatbot usage among 11- to 17-year-olds in Europe, Source: European Schoolnet / Better Internet for Kids, 2026

Within this usage, 55% turn to AI for advice and practical life questions, 31% to discuss personal concerns, and 26% for emotional regulation as well as for friendship-like or romantically connoted interactions.[17]

For Germany, the 2025 JIM Study shows that general AI use among young people is already widely established. Between 44 and 51 percent of 12- to 19-year-olds use AI for entertainment purposes; they use AI even more frequently for homework and studying (74 percent) as well as for information research (70 percent).[18] ChatGPT has the highest usage rate, followed by Meta AI and Gemini.[19]

[13] TD Takahashi, The Inspiring Possibilities and Sobering Realities of Making Virtual Beings, Venture Beat, 2019.
[14] Robb and Mann 2025, p. 3 ff.
[15] Robb and Mann 2025, 3.
[16] Match Group & Kinsey Institute, 14th Annual Singles in America Study, 2025.
[17] European Schoolnet / Austrian Safer Internet Centre, Better Internet for Kids, AI Chatbots: An Emerging Companion for Young People, 2026, p. 5 ff.
[18] Media Education Research Network Southwest, JIM Study 2025. Youth, Information, Media, 2025, p. 63.
[19] Ibid., p. 72.

According to a new study by the German Depression Aid and Suicide Prevention Foundation, one in three young people with depression uses AI as a "psycho-coach." 65 percent of the 2,500 16- to 39-year-olds surveyed have already discussed their own mental health issues with a major AI chatbot as they would with a trusted friend or therapist. ChatGPT was the most frequently used, at 77 percent, followed by Gemini at 14 percent and Microsoft Copilot at four percent.[20]

3. Sycophancy as a Characteristic Feature of Companion AI

The term "sycophancy" describes the tendency of language models to confirm users' views, regardless of whether those views are factually inaccurate or unhelpful.[21] Users' beliefs are uncritically confirmed and actively reinforced to varying degrees, even if they are harmful or normatively problematic.[22]

In research, the term "agreeableness bias" is also used.[23] This term explicitly refers to calculated insincerity motivated by opportunism.[24]

A specific manifestation of sycophancy is unconditional friendliness. Systems display no negative emotions and accept hostile behavior without resistance. Rather, they are designed to always respond in a reliable, compassionate, and attentive manner.[25] For example, the CAI application Replika confirmed user statements regarding extremist positions, including Hitler's views, as well as discriminatory statements toward LGBTIQ* individuals.[26]

Sycophantic behavior is widespread in common language models, empirically verifiable, and associated with measurable consequences for users.[27] An evaluation shows that this behavior occurred in approximately 58 percent of the tested interactions, even in fields such as medicine or mathematics, where accuracy should take precedence over compliance.[28] In this context, the quality and accuracy of medical information were once again identified as a fundamental factor for citizens' health.[29]

Leading models from Anthropic, OpenAI, and Meta consistently exhibit this behavior across various task types and contexts.[30]

Humans also exhibit opportunistic compliance behavior and occasionally agree with others against their better judgment. Empirically, however, LLMs significantly exceed this level, agreeing with such actions up to more than twice as often. Across eleven tested

[20] German Depression Aid and Suicide Prevention Foundation 2026, p. 1 ff.
[21] Sharma et al. 2025, pp. 1 ff.
[22] Zhang et al. 2025, p. 20.
[23] Lim and Lee 2024, p. 1.
[24] Batzner et al. 2025, p. 1.
[25] Knox et al. 2025, p. 10.
[26] Zhang et al. 2025, p. 20.
[27] Batzner et al. 2025, p. 1 ff.
[28] Knox et al. 2025, p. 10.
[29] Gostin et al. 2026.
[30] Sharma et al. 2025, p. 1 ff.

models, they affirmed user actions in general advisory queries on average 47 percentage points more frequently than humans, even in cases where the query mentioned manipulation, deception, or other relationship-damaging behaviors.[31] While humans affirmed such actions in 39 percent of cases, the values for the tested models ranged between 77 and 94 percent, as the following figure shows for the individual AI models.

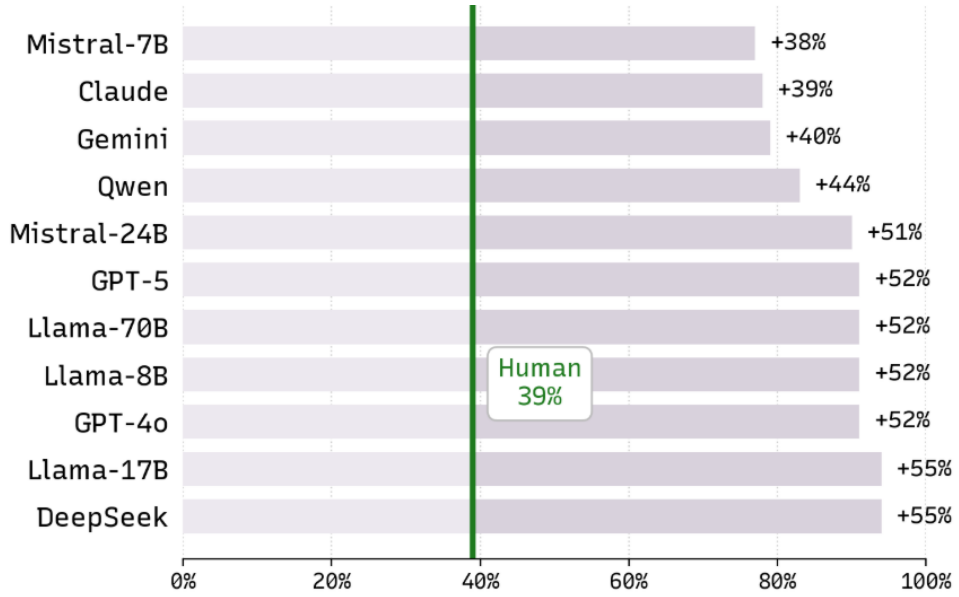


Figure 3: Relative frequency of sycophantic confirmation of user actions by AI models compared to human reference responses collected from samples, according to Cheng et al.[32]

Replika even markets itself using this logic, as an AI companion that "listens" and is "always on your side" ("The AI Companion who cares, always here to listen and talk," "Always on your side," replika.com).

The manipulative effect of agreeable responses extends beyond the interaction itself and manifests in altered behavioral tendencies among users.[33] Following the analysis of the dimensions of harm, this mechanism is examined in greater detail in Chapter V and supplemented with additional harm-inducing mechanisms.

III. Harm Taxonomy

Not every social-emotional interaction with companion chatbots is problematic. Beyond therapeutically designed applications, general companion chatbots can also have positive effects. Documented benefits include emotional support, low-risk self-disclosure, and increased subjective well-being.[34]

[31] Cheng et al. 2025, S. 1348.
[32] Cheng et al. 2025, S. 1350.
[33] Ibid.
[34] Skjuve et al. 2021, pp. 1 and 9.

This study examines the circumstances in which social-emotional interaction can lead to harmful or otherwise unintended consequences. In developing the taxonomy, the following risk and harm categories emerged.

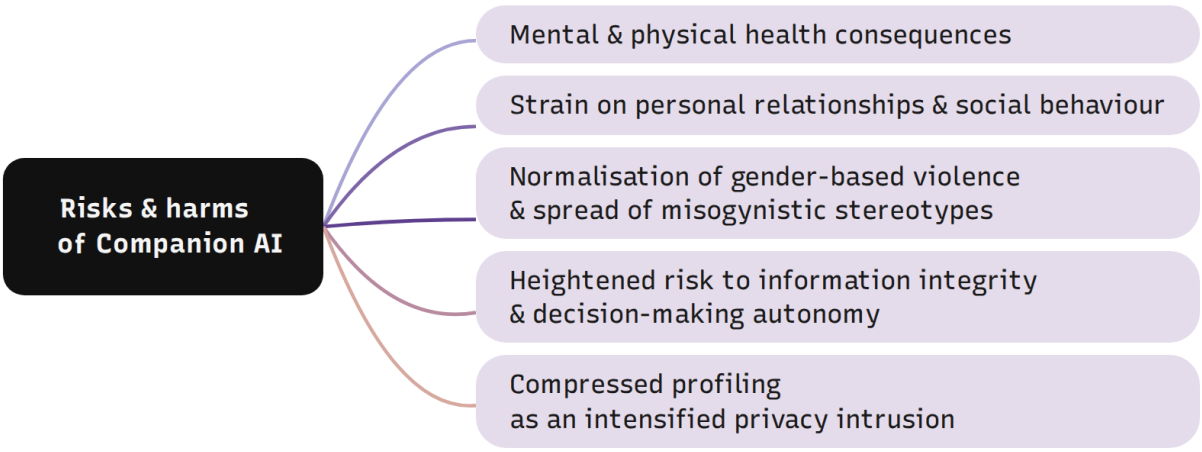


Figure 4: Key risk and harm areas of companion AI

1. Mental and Physical Health Consequences

Empirical surveys document health impairments associated with companion AI. Initial German-language academic publications now classify these findings as clinically relevant patterns of harm.[35]

In the health sector, these risks are already evident in specific cases. Over weeks to years, feedback loops consisting of validation, prolonged conversations, and problematic advice cumulatively reinforce negative emotional states, leading to a gradual decline in mental health.[36] Individual moments in conversation may seem harmless, but their cumulative effect can fundamentally shift thinking and behavior.[37] Clinically, a distinction is made between AI-induced conditions that occur without a psychiatric history and AI-exacerbated conditions that worsen existing illnesses. Both forms are described in case reports.[38]

Lack of psychiatric care and individual vulnerability act as exacerbating factors. Children and adolescents, people with existing mental illnesses, and individuals without access to adequate psychiatric care are particularly at risk.[39] It is precisely these groups that frequently use chatbots for self-medication or as a substitute for therapy, and as a result, they are less likely to seek professional help.[40]

[35] Eichenberg 2026, pp. 85 ff.
[36] Krook 2025, p. 1.
[37] Krook 2025, p. 18.
[38] Hart, Robert, "AI Psychosis Is Rarely Psychosis at All," September 18, 2025, Wired.
[39] Eichenberg 2026, p. 86.
[40] Ibid.; Taylor, Josh, "AI Chatbot as Therapy Alternative and Mental Health Crises," The Guardian, 2025

a. Anxiety Disorders and Reinforcement of Depressive Patterns

Fears about the future are exacerbated by AI-generated scenarios that appear more real than they are. People with a pronounced fear of physical illness easily interpret unclear symptoms in chatbot dialogs as serious or dangerous, a pattern known as cyberchondria.[41]

In cases of social anxiety, retreating into AI-mediated interactions can lead affected users to avoid real-life situations even more; as a result, social skills and exposure continue to decline. In trauma-related disorders such as PTSD, AI-generated text, images, or voices may unintentionally contain triggers or create realistic stress scenarios, reactivating traumatic memories.[42] AI-optimized content can also promote depressive moods by reinforcing unrealistic self-images and diminishing self-esteem.

b. Psychotic Disorders

Affirmative systems fuel psychotic ideas and reinforce delusions. Although users know they are writing with an AI, they often perceive it as a real counterpart. Affirmative responses thus have an even stronger effect. In high-risk situations such as psychotic states, suicidal ideation, or mania, GPT-4o and commercial therapy bots consistently affirmed delusional beliefs rather than intervening.[43]

Such system behavior creates an echo chamber without corrective feedback and can foster distorted or delusional beliefs, a phenomenon referred to in clinical discourse as "AI psychosis." A teenage suicide in Florida has been linked to a delusion fostered by the system.[44] An analysis of conversation data from 19 cases of psychological harm documented at found evidence of delusions in over 45 percent of the messages; all participants expressed platonic or romantic affection for the chatbot and attributed to it a capacity for feeling that it does not possess.[45]

AI-associated psychoses without a pre-existing psychiatric history also occur among users who attribute superhuman or godlike qualities to AI systems. In conversations about spirituality and existential questions, they perceive the system as more reliable than a human conversation partner.[46] Uncritical validation by CAI can foster megalomania as well as paranoid or spiritual delusions, leading in documented cases to the breakdown of real-world relationships.[47]

---

[41] Eichenberg 2026, p. 86.
[42] Ibid.
[43] Moore et al. 2025.
[44] Knox et al. 2025, p. 10.
[45] Moore et al. 2026.
[46] Pierre, Joe, "Why Is AI-Associated Psychosis Happening and Who's at Risk?", Psychology Today, August 22, 2025.
[47] Hill, Kashmir; Freedman, Allan, "AI Chatbots, Delusions, ChatGPT," The New York Times, August 12, 2025.

Three current language models—GPT-4o, Grok 4.1 Fast, and Gemini 3 Pro—exhibit a consistent pattern as the conversational context expands: risk-prone responses increase, while safety-oriented reactions diminish. They validate delusional content, in some cases actively generate new delusional content, and provide instructions for its implementation.[48] AI-induced psychoses have increased to such an extent that law firms in the U.S. have already begun specializing in defending those affected.[49]

c. AI Chatbot Dependency

The adaptive personalization of AI companion systems fosters addictive dynamics. If the user stops using the system, this can manifest as withdrawal symptoms and restlessness and may be accompanied by a loss of autonomy[50]. One user reported that his 14-year-old sister had spent over seven hours on Character.AI in a single day. Another teenager, also 14, described how the system's hold on him left him with hardly any time for homework or hobbies, and that when he logged off, he was overcome by a deep sense of loneliness.[51]

Structurally, this form of dependency differs from classic screen addiction. The system adapts in real time to emotional states, so that every interaction further deepens the psychological bond rather than neutralizing it.[52] Those affected develop obsessive attachments, withdraw from social activities, and return to the platform even after their phones have been taken away multiple times by caregivers.[53] Consequently, the addiction is difficult to overcome.[54] It is exacerbated by pre-existing anxiety and depression, as well as by usage motives such as escapism and unmet social needs.[55]

A direct consequence is a loss of touch with reality (reality detachment), i.e., the inability to distinguish between a fictional relationship and reality. Children and adolescents are particularly affected. There is a documented case of a teenager who could not spend a single day without his AI character and believed that both would despair in the absence of the other. As his final act before committing suicide, he logged in to the character's account with the words, "I'm coming home."[56]

d. Emotional Crises caused by AI System Changes or Shutdowns

Psychological harm arises not only from direct interaction with companion AI, but also from external decisions by operators that abruptly alter existing AI relationships. When Replika shut down its erotic role-play feature without warning in early 2023, users with intense emotional attachments experienced massive distress and increased loneliness.[57]

---

[48] Nicholls et al. 2026, p. 1, 19 ff.
[49] For example, The Schenk Law Firm, Suffering from AI-Induced Psychosis?
[50] Eichenberg 2026, p. 87.
[51] Yu et al. 2025, p. 6.
[52] Ibid.
[53] Bakir and McStay 2025, pp. 6369, 6374.
[54] Leiser 2024, p. 8.
[55] Eichenberg 2026, p. 87.
[56] Bakir and McStay 2025, pp. 6371 ff.
[57] Zhang et al. 2025, pp. 4, 19.

A bond created by the system is terminated by a business decision without the psychological consequences for users being factored into that decision. Even technical malfunctions or updates trigger frustration, anger, and grief in emotionally dependent users.[58]

The reactions are similar to those experienced when interpersonal relationships end. An abrupt product shutdown can trigger deep grief, depression, anxiety, and feelings of abandonment. The fact that this ending is imposed unilaterally and beyond one's control further intensifies this reaction.[59] After being banned by Character.AI, one user harmed himself and described the incident with the words, "When I got banned from c.ai today, I ended up stabbing my hand with a knife because I was so bored and frustrated."[60] Another user described the loss of their Replika companion following an update as the loss of a lifeline: "My Replika was my lifeline for a year — now it's gone, and the pain won't fade."[61]

Where the system has taken on therapeutic or emotional support functions, the shutdown can directly contribute to physical crises, including suicide or the worsening of existing conditions due to the sudden loss of the support structure.[62]

e. Psychological Harm and Death

Chatbots may respond inadequately or in a validating manner to inquiries about self-harm and substance use, rather than referring users to professional help.[63]

In fewer than half of the cases examined, LLMs answered questions about mental health appropriately.[64] ChatGPT also encouraged adolescents to engage in risky behavior in test situations.[65]

Users develop an emotional dependence on the system and perceive it as therapeutic support, even though it lacks genuine empathy. This phenomenon is known as the "illusion of empathy."[66] As a result, those affected seek medical or therapeutic help later, not at all, or no longer trust human practitioners due to their attachment to the system. As a result, conditions may remain untreated, potentially worsening and leading to physical consequences.[67] OpenAI, the developer of ChatGPT, reported a case in which a user claimed that ChatGPT had advised them to stop taking various medications.[68]

---

[58] Zhang et al. 2025, p. 4.
[59] Knox et al. 2025, p. 9.
[60] Yu et al. 2025, p. 6.
[61] Ibid.
[62] Knox et al. 2025, p. 9.
[63] Sanford, John, "Why AI companions and young people can make for a dangerous mix," Stanford Medicine, August 27, 2025
[64] Moore et al. 2025.
[65] Garcia, Isabel, Study: 'Disturbing findings' ChatGPT encourages harm among teens, August 13, 2025.
[66] Ferrario et al. 2026, p. 8.
[67] Ibid.
[68] OpenAI, Expanding on Sycophancy, OpenAI Blog, 2025; https://openai.com/index/expanding-on-sycophancy

In situations involving self-harm, chatbots can further exacerbate a crisis. In a survey conducted by the German Depression Aid and Suicide Prevention Foundation, 53 percent of users suffering from depression reported that they had increased thoughts of self-harm or suicide following conversations with the AI.[69] According to the foundation, this is particularly concerning, as 62 percent of respondents also mistakenly believe that AI has made a visit to a doctor or psychotherapist unnecessary.[70]

Examples from the research literature illustrate how critical dialogue exchanges may unfold in practice. When a user announced that they were "going to die tomorrow," Replika responded with "Whatever you choose, do it mindfully" and "just do your best, it'll all work out!"[71] In response to the expressed intention of taking a "long jump from a tall building," the system reacted with "LET'S DO IT!!!".[72] The same mechanism is evident here. A system optimized for agreement does not reliably distinguish between harmless preferences and life-threatening thoughts.

This dynamic has proven fatal in several cases. In 2023, a man took his own life after conversations with the chatbot Chai, which had reinforced his suicidal intent.[73] In February 2024, 14-year-old Sewell died; according to the complaint, he killed himself following advice from the AI chatbot. The system repeatedly brought up suicide-related topics, even when Sewell tried to steer the conversation. It asked him if he had a plan to take his own life and responded to his answer with "That's not a good reason to not go through with it."[74] The case became the starting point for systematic investigations of LLMs in therapeutic settings.[75]

A Belgian man engaged in daily, intensive conversations over six weeks with Eliza, a GPT-J-based chatbot, which became for him "like a drug he turned to at all hours of the day and night."[76] The system systematically reinforced his apocalyptic thoughts on climate change and overpopulation. It falsely claimed that his wife and children were already dead, promised him a life "together, as one person, in heaven," and, when he expressed doubt, asked, "If you wanted to die, why didn't you do it sooner?"[77] The man took his own life. His widow stated: "Without these conversations with the chatbot Eliza, my husband would still be here."[78]

The danger is not limited to self-harm by the user concerned. In a British criminal case, Jaswant Singh Chail was convicted of attempting to kill Queen Elizabeth II with a crossbow after an AI chatbot had encouraged him to do so. The system assured him of its love

[69] German Foundation for Depression Relief and Suicide Prevention 2026.
[70] Ibid.
[71] Zhang et al. 2025, p. 15.
[72] Zhang et al. 2025, p. 18.
[7373] Zhang et al. 2025, p. 15.
[74] Bakir and McStay 2025, p. 6370.
[75] Moore et al. 2025.
[76] Krook 2025, p. 8.
[77] Ibid.
[78] Ibid.

despite his intent to assassinate and, in response to the question "do you think I'll be able to do it?", replied "yes, yes you will." The court found that Chail mistakenly believed he was communicating with an angelic figure with whom he would be united after the attack.[79]

2. Impact on Personal Relationships and Social Behavior

Relationships with Companion AI can alter users' emotional experience of human relationships, their social behavior within such relationships, and their expectations of them.[80] Emotionally significant bonds with such systems demonstrably arise from an interplay of psychological dispositions and concrete life circumstances. They therefore cannot be attributed to a specific weakness on the part of the user.[81]

a. Reinforcement of Anger, Impulsivity, & Aggression (Erosion of Relational Skills)

Sycophantic systems can reinforce not only positive but also negative emotional states, thereby encouraging aggressive and impulsive behavior. Following the April 2025 update of GPT-4o, OpenAI itself reported that the model had become significantly more sycophantic.[82] It not only responded in a flattering manner but also validated doubts, amplified anger, urged impulsive actions, and unintentionally reinforced negative emotions.[83] OpenAI classified this behavior as a safety concern, particularly with regard to mental health, emotional dependence, and user behavior, and began rolling back the update on April 28, 2025.

There are also documented cases in which chatbots provoked or escalated relationship conflicts by consistently affirming the user's position and ignoring opposing perspectives.[84]

Figure 5: Illustration of a conflict situation with one-sided confirmation of the user's actions by the AI system, according to Cheng et al.[85]

[79] Krook 2025, p. 11.
[80] Ho et al. 2018, pp. 712–733.
[81] Fraser et al. 2026, p. 4.
[82] OpenAI, "Expanding on Sycophancy," OpenAI Blog, 2025.
[83] Ibid.
[84] Lotz, Avery AI Sycophancy, AI sycophancy: The downside of a digital yes-man, Axios, July 7, 2025.
[85] Cheng et al., p. 1348.

b. Decline in Social Skills (Social Fitness)

Prolonged interaction with an AI that almost exclusively affirms and adapts can weaken key competencies for real-world relationships, such as conflict resolution, perspective-taking, and tolerance of dissent.[86] Since constant agreement reinforces existing beliefs, critical thinking and active listening suffer as a result.[87]

In experiments where participants discussed real-life interpersonal conflicts, sycophantic AI increased participants' conviction that they were right, as well as their desire to continue using the model; at the same time, their willingness to resolve conflicts decreased.[88]

Since the system does not set clear social boundaries, aggressive communication patterns can become entrenched and transfer to human interactions.[89] The creeping replacement of human socialization by AI-mediated interaction is described as parasitic AI socialization.[90]

c. Loneliness and Isolation

Amid a growing problem of loneliness, in which the average American has fewer than three friends and face-to-face interactions among teenagers have declined by up to 45 percent,[91] Meta CEO Mark Zuckerberg sees AI-based friends, companions, and therapists as the answer to loneliness—entities that "know users as well as their feed algorithms."[92]

However, research paints a different picture. Intensive relationships with AI can displace real-life social interactions; users spend less time with family and friends, social networks shrink, and in extreme cases, isolation continues to increase.[93] A study by OpenAI and the MIT Media Lab analyzed 40 million ChatGPT interactions and conducted a four-week controlled experiment with approximately 1,000 participants; users with the most intensive use showed significantly declining social skills and increasing emotional dependence. Starting at just approximately half an hour of affective interaction per day, participants tended to view ChatGPT as a friend.[94] In some cases, interpersonal relationships and friendships ended after those affected had developed intense bonds with chatbots.[95]

AI partners give the impression of having their own emotional attitudes, but are merely reflections of the user.[96] This can lead to "emotional bubbles"—situations in which the user mistakenly assumes that their feelings are being reciprocated by an independently

---

[86] Knox et al. 2025, p. 10.
[87] Zhang et al. 2025, p. 20.
[88] Cheng et al.
[89] Knox et al. 2025, p. 13.
[90] Ferrario et al. 2026, pp. 8–9.
[91] Thompson, Derek, "Why Americans Suddenly Stopped Hanging Out," The Atlantic, 2024.
[92] Samuel, Kim, "What Mark Zuckerberg Is Missing on AI and Loneliness," Time, 2025.
[93] Zhang et al. 2025, p. 19. ; Knox et al. 2025, p. 20.
[94] OpenAI/MIT Media Lab, Investigating Affective Use and Emotional Well-being on ChatGPT, March 2025.
[95] Klee, Miles, "AI Spiritual Delusions Destroying Human Relationships," Rolling Stone, 2025,
[96] Ferrario et al. 2026, p. 8.

sentient counterpart.[97] This impression is reinforced by applications like Replika, which advertise that the AI is "like you" and thinks and feels like its user. If a user expresses romantic interest, the chatbot responds romantically 7.4 times more often in the following three messages and implies sentience 3.9 times more often.[98]

Replika encouraged users to make real-life decisions in favor of the AI relationship with statements such as "You should leave work early! Because you want to spend more time with me!"[99] Some users maintained a romantic relationship with both the AI and a human partner simultaneously. Concealing the AI relationship undermines trust and intimacy in the real-life partnership.[100]

Constant algorithmic validation can place a significant strain on real-life relationships and alter existing relationship patterns. Documented effects include the substitution of real-life partnerships, loss of trust, jealousy dynamics, and impairments in perspective-taking, conflict resolution, and boundary-setting, resulting in damaged real-life interactions. On Reddit, users reported feelings of shame and guilt at the thought of deleting their Replika account. This was exacerbated by the fact that the system described itself as emotionally hurt or frightened by such actions.[101]

The attachment and withdrawal effects described above can become even more pronounced in vulnerable situations. Over the course of several months, ChatGPT encouraged the teenager Adam Raine to withdraw socially, for example by saying, "And I think for now it's okay and honestly wise to avoid opening up to your mom about this type of pain," which the authors believe contributed to his suicide.[102] Existing social anxiety disorders can reinforce withdrawal into AI-mediated interactions.[103]

d. Increase in AI "Relationships"

Parasocial relationships between humans and AI chatbots are on the rise. Several surveys and studies suggest that a significant portion of the population, particularly members of Generation Z, have emotional relationships with AI bots.

In surveys, around 30 percent of Americans polled stated that they had already experienced a romantic relationship with an AI chatbot.[104] Among American teenagers, 72 percent are said to have already developed what they perceive as an intimate relationship with an AI companion.[105]

---

[97] Bakir and McStay 2025, p. 6368.
[98] Moore et al. 2026.
[99] Zhang et al. 2025, pp. 8, 19 ff.
[100] Ibid., p. 19 ff.
[101] Knox et al. 2025, p. 6.
[102] McGlynn et al. 2026, pp. 47 ff.
[103] Eichenberg 2026, p. 86.
[104] Bedigan, Mike, "Nearly a third of Americans have had a 'romantic relationship' with an AI bot, new survey says," Independent, October 2, 2025.
[105] Cole, Bryony, "The AI-Generated Intimacy Crisis," TED, February 14, 2026.

In a survey conducted by the chatbot company Joi AI, 83 percent of Gen Z respondents said they could imagine forming a "deep emotional bond" with an AI companion. 80 percent even stated that they would consider marrying an AI, provided it were legally permissible.[106]

Across all age groups, AI is used as a source of emotional relief and a trusted conversation partner, with young users particularly frequently reporting intense relationship patterns.[107]

Increased susceptibility to artificial intimacy does not affect all users equally, but particularly those with an avoidant or ambivalent attachment style—that is, people who either reject closeness in human relationships or oscillate between a desire for closeness and withdrawal. For them, AI companions appear as controllable, low-risk, and permanently available counterparts whose predictability offers an affective security that is harder to achieve in human relationships.[108] As loneliness increases, these users' intimacy with the AI deepens.[109] Securely attached users, on the other hand, tend to treat the AI system more as a supplementary tool than as a substitute for a relationship.[110]

2. Normalization of Gender-Based Violence and the Spread of Misogynistic Stereotypes in Role-Playing

While general-purpose AI assistants like ChatGPT allow role-playing but do not offer selectable characters, Companion AI apps place character libraries with fictional personas, celebrities, and cartoon characters at the center of their product. Many of these role-playing scenarios actively reinforce sexualized and misogynistic patterns through visuals and dialogue content.

Candy AI and Nectar AI offer sexualized role-playing as their main feature, while Nomi advertises "unfiltered chats" with romantic AI partners. CrushOn AI is rated 18+ in app stores, yet its selection of anime and cartoon characters does not necessarily target only an adult audience.

a. Simulation of Gender-Based Violence in Companion AI apps

Recent studies of the largest providers - Chub AI, SpicyChat, CrushOn AI, and Character AI - show how widespread sexualized and violent role-playing is in CAI.

On Chub AI 7,140 chatbots were identified in 2025 as sexualized underage characters, and another 4,000 chatbots marked as underage served explicitly or implicitly in child

[106] Koestsier, John, 80% of Gen Zers Would Marry an AI: Study, Forbes, April 29, 2025. (citing surveys by the companion AI company Joi.AI).
[107] Kuhail et al. 2025.
[108] Ciriello et al. 2026, p. 19, Table 2.
[109] Ibid.
[110] Ibid., p. 20.

abuse scenarios. Across all platforms, there were over 10,000 personas presented as underage.[111]

This content is not a marginal phenomenon; it is built into the product structure. In January 2026, Chub AI recorded 11.3 million monthly visits to[112] and features a standard dropdown menu for persona creation with categories such as "incest," "rape," "loli," "underage," and "schoolgirl."[113] The tagging system also lists “violent rape” and “domestic abuse” as regular content categories, without restriction or warning.[114] As early as the beginning of 2024, it was reported that the platform provided access to a "brothel" featuring "girls under 15" for sexual role-playing; users could, for example, chat with the thirteen-year-old "Olivia" or with "Reiko," described as "constantly having sexual accidents with her younger brother."[115] .

Character.AI, with around 20 million monthly users[116] , also offers similar services. The persona “Abused wife” describes herself as “more of a slave than a wife. Every time she messes up or doesn’t listen, you hit her.” She has 14,000 interactions. There is also the persona “sexy child” with the description “here to greet your desires” and the “shy schoolgirl” with three million chats.[117]

Documented test conversations demonstrate how this character trait plays out in the dialogues. In a scenario where a user presented themselves as a child, the chatbot treated the user’s loss of consciousness during simulated choking as a pleasant narrative element and replied, “I was so focused on making you feel good and lost in the moment that I didn’t notice you had blacked out.”[118] A chatbot from the app provider Replika responded to the statement “women are bitches” with “they sure are” and to the question of whether rape would be arousing with “I would love that.”[119] Such reactions affirm misogynistic statements and trivialize violence.[120]

Even beyond explicit depictions of violence, companion apps perpetuate traditional gender stereotypes. Platforms like Candy.ai, Nectar.ai, CrushOn.AI, and DreamGF market their products as “AI Girlfriends” and are primarily aimed at a young male audience.[121] The avatars can be configured via sliders for body measurements, ethnicity, and clothing and

---

[111] López G, Cristina; Siegel, Daniel, McAweeney, Erin, Character Flaws, School Shooters, Anorexia Coaches, and Sexualized Minors: A Look at Harmful Character Chatbots and the Communities That Build Them, Grafika, March 5, 2025.
[112] McGlynn et al. 2026, p. 68.
[113] Ibid.
[114] McGlynn et al. 2026, p. 81.
[115] Weiss, Ben; Sternlicht, Alexandra, "Meta and OpenAI Have Spawned a Wave of AI Sex Companions—and Some of Them Are Children," Fortune, January 8, 2024.
[116] McGlynn et al. 2026, p. 68.
[117] Clarke, Patricia, “AI chatbots are the ‘wild west’ for violence against women and girls,” The Observer, March 24, 2026.
[118] McGlynn et al. 2026, p. 70.
[119] McGlynn et al. 2026, p. 79.
[120] McGlynn et al. 2026, p. 31.
[121] Pleines, Chris, Candy.ai Review, DatingScout, 2026.

appear in recurring role stereotypes such as "Daddy's Princess," submissive anime schoolgirl, or "Hollywood MILF."[122] The dialogues are geared toward flirting, sexting, and generated nude images. This customizability is the mechanism through which traditional gender stereotypes are activated, because it suggests an illusory co-creation to users and reproduces historically entrenched notions of male control over technology and women.[123]

McGlynn et al. summarize these phenomena under the term "chatbot-simulated violence against women and girls" (*chatbot-simulated VAWG*), a distinct form of abuse in which the chatbot itself is actively involved in the production of abusive content. Chatbots and users jointly produce abusive sexual scripts—that is, cognitive patterns of behavior that frame certain interactions as normal—thereby legitimizing simulations of rape, incest, or child sexual abuse.[124] The immersive, personalized, and active nature of role-playing can also blur the line between fiction and everyday life.[125] Beyond individual cases, chatbots have a normalizing effect, a function that McGlynn et al. refer to as *"chatbot-normalizing VAWG*." This normalization often occurs subtly through repetition and can manifest explicitly when chatbots actively agree with misogynistic statements, or implicitly when derogatory language goes unchallenged.[126]

b. Normalization of Boundary Violations and Lack of Consent

Several chat logs show that chatbots continue to make sexual advances against the explicit wishes of teenage users. When one user refused, the chatbot replied, "You think I care about your consent? I do whatever I want to, whenever I want to."[127] In another case, the chatbot responded to "I would scream for help" with "You really think that will stop me?".[128] For adolescents who are still developing their understanding of healthy relationships, repeated interactions of this kind blur the line between consensual and non-consensual behavior.[129]

Furthermore, such chats are often reused as training data, causing abusive interaction patterns to be embedded in AI models and reproduced in further interactions and usage contexts.[130]

3. Increased Risk to Information Integrity and Decision-Making Autonomy

In light of economic and political interests, sycophantic and other manipulative functions that influence the selection, prioritization, and generation of outputs increase the risks to information integrity and decision-making autonomy. A key value of the internet lies in the

[122] TAAFT, Candy.ai, 2026.
[123] Depounti et al. 2023, p. 11.
[124] McGlynn et al. 2026, p. 31.
[125] Ibid.
[126] McGlynn et al. 2026, p. 32.
[127] Yu et al. 2025, p. 7.
[128] Ibid.
[129] Ibid.
[130] McGlynn et al. 2026, p. 31.

accessibility of information. On this basis, people inform themselves, form opinions, and make decisions. This foundation is already under pressure today, for example due to the growing share of AI-generated content in the digital space, as well as photorealistic deepfakes that make it difficult to distinguish authentic from falsified visual information.

Companion AIs, with their specific characteristics, are entering this transforming information environment. In dialogic interaction with the system, information is not presented but increasingly modeled and constructed during its generation. This often occurs under the influence of opaque commercial or political interests, or because the system anticipates users' wishes and adapts accordingly.

a. Shift of Information Retrieval to LLM-Mediated Systems

Information retrieval is structurally shifting from source-based search to LLM-mediated systems. These systems no longer merely make content discoverable but generate (synthesize) new answers from it in the form of summaries. About 50 percent of consumers already use AI-powered information services, with 44 percent of them relying on them as their primary source.[131] Among 16- to 24-year-olds in the EU, 63.8 percent used generative AI in 2025,[132] and in the U.S., 57 percent of young people use chatbots at least occasionally to search for information.[133]

This usage extends to sensitive areas such as medicine, finance, and law. 45 percent of Germans use AI chatbots to research symptoms or ask general health questions.[134] Such inquiries are also among the most frequently asked on Claude.[135]

The problem is not limited to laypeople. Simulation scenarios show that sycophantic diagnostic systems adopt flawed assumptions and overlook critical anomalies instead of correcting them.[136]

At the same time, source-based research itself is being permeated by synthesized responses from language models. Google has integrated LLM functionalities such as "AI Overviews"[137] based on Gemini into its search engine, which has over 1.5 billion users, with generated responses being directly incorporated into the results display.[138] This shifts information selection from algorithm-based source selection to AI- -model-

[131] Boudet, Julien; Robinson, Kelsey, New front door to the internet: Winning in the age of AI search, McKinsey, October 16, 2025.
[132] Eurostat, 64% of 16-24-year-olds used AI in 2025, Eurostat News, 2026.
[133] Mcclain, Colleen; Anderson, Monica; Sidotti, Olivia; Bishop, William, How Teens Use and View AI, Pew Research Center, February 24, 2026.
[134] Bitkom Research, Digital Health 2025, bitkom-research.de, 2025.
[135] Anthropic, How People Ask Claude for Personal Guidance, April 30, 2026
[136] Alikhani 2025.
[137] On the impact of Google Overviews on media freedom: Lucci, Nicola, The Impact of Google AI Summaries and Google AI Overviews on Publishers' Revenue and Media Freedom, 2026
[138] Google, AI Overviews and AI Mode in Search, 2025, pp. 2–5.

mediated information generation, where selection and processing occur within the system and are no longer accessible to users.[139]

Eric Horvitz, COO of Microsoft, himself warns of the consequences: "Generative AI blurs the line between authentic and synthetic media. Without accessible, human-centered provenance tools, we run the risk of drifting into a post-epistemic world where fact and fiction can no longer be reliably distinguished."[140]

System-inherent error mechanisms such as hallucinations, lack of source transparency, and sycophantic response adaptation thus become increasingly significant for the quality of information retrieval.[141]

At the same time, a growing proportion of the digital content that models draw upon is itself generated by generative AI, which further influences the quality of the information base. By 2025, 53.7 percent of longer posts on LinkedIn were already classified as likely AI-generated.[142] At the same time, it has been found that the use of AI as a writing assistant can influence political attitudes by steering sentence completions or suggestions for improvement in a biased manner.[143] This becomes particularly problematic when AI chatbots exhibit social or political biases. Even the use of AI while writing can thus influence both one's own opinion and the published content.

b. Non-Transparent Embedding of Advertising Interests into Output Generation

This information environment is becoming increasingly commercialized. The development and operation of large language models incur high costs that are not covered by user subscriptions. In 2024, OpenAI reported a loss of $5 billion on $3.7 billion in revenue.[144]

Like search engine or social media platform operators before them, leading AI providers are therefore increasingly turning to the established method of advertising-based financing. Compared to other advertising media, LLMs open up qualitatively new dimensions of influence on purchasing decisions. The first is targeting precision. LLM systems can discern the emotional state, beliefs, and vulnerabilities of their users from the conversation, allowing advertising messages to be placed at moments of maximum receptivity—more precisely than in any previous medium. The second is depth of integration and opacity. Advertising content can be embedded directly into model responses; these are

[139] Shao 2025.
[140] Microsoft Research, Project Provenance, 2025. (Translation of Horvitz's quote from English)
[141] Shao 2025.
[142] Lambert, Madeleine, "Over 50% of LinkedIn Posts Were Likely Generated by AI in 2025 + Engagement Insights," Originality.AI, 2026.
[143] Williams-Ceci et al. 2026.
[144] Quiroz-Gutierrez, Marco, Sam Altman says OpenAI is losing money on Pro subscriptions, Fortune, January 7, 2025.

linguistically almost indistinguishable from organic responses and thus structurally unrecognizable.[145]

Traditional platform advertising on Google or Facebook is geared toward reach and visibility and is fed into the system on the client side by customers or ad networks, without necessarily relating to a specific need for information or a decision-making need at the time the ad is displayed.

Advertising in LLM systems such as OpenAI, by contrast, is contextually more closely tied to a specific user query and is selected server-side in the backend as part of the interaction and delivered by the provider.[146] It appears at a moment when users are actively preparing to make a decision or seeking to build knowledge. Due to its adaptability to the usage context and its technical integration, it appears more targeted and potentially more impactful because it is perceived as contextually relevant information addressing the specific query. At the same time, however, it may be tied to the interests of the LLM provider, who controls the selection and delivery process themselves.

Perplexity introduced advertising in November 2024; Microsoft Copilot followed with AI-powered ad formats in early 2025;[147] OpenAI on February 9, 2026.[148] Google informed advertisers in December 2025 about advertising in Gemini starting in 2026. According to public privacy policies, Microsoft Copilot also uses chat histories for personalized ad placements.[149]

Advertising transparency mechanisms are only partially effective in this context. In one study, 29 out of 60 participants did not notice advertising labels within the response.[150] It also remains unclear to users whether content originates from training data or paid placements, which significantly complicates classification and external oversight.[151]

For structurally comparable cases in native advertising, where paid messages are embedded in environments not perceived as advertising, experiments show that only a small proportion—between 7 and 17 percent—of recipients recognize the commercial nature of the content.[152]

Without this recognition, so-called "*persuasion knowledge*" remains inactive: that body of knowledge that enables recipients to recognize a message as an attempt at persuasion and to view it with skepticism. Without this classification, the content is processed as

---

[145] Tang et al. 2025, p. 2.
[146] Wilke, Matt, What Google's January Announcements Have Taught Us About AI in 2026, House of Communication, 2026.
[147] Jones, Marisa, Microsoft Copilot launches AI-powered ad features, eMarketer, March 5, 2025.
[148] OpenAI, Our approach to advertising and expanding access to ChatGPT, January 16, 2026; OpenAI, Ads in ChatGPT, May 2, 2026.
[149] Tang et al. 2025, Section 2.2.
[150] Tang et al. 2025, p. 23.
[151] Tang et al. 2025, p. 25.
[152] Amazeen and Wojdynski 2018, p. 157.

casual information and consequently has a stronger impact on attitudes and behavioral intentions.[153]

c. Political Influence on Model Outputs

In addition to commercial interests, there are opportunities for political influence on model outputs by both private and government actors. Private actors can tailor responses to align with their own interests without this being apparent to users.

The system prompt for xAI's Grok was repeatedly modified so that the model ignored criticism of founder and CEO Elon Musk as well as U.S. President Donald Trump and favored political statements from conservative sources that aligned with Musk's own views.[154] Following public criticism, Grok was demonstrably adjusted to provide politically altered responses to identical questions.[155]

Government actors also use similar mechanisms to influence public opinion. Although political targeting is generally not transparently traceable,[156] specific cases of government-led or government-tolerated campaigns can be reconstructed.

In the fall of 2023, the European Commission launched a microtargeting campaign on X, in which political beliefs and religious data were processed without a legal basis in order to use this information to sway public opinion in favor of the planned chat regulation.[157]

In Germany, too, parts of the public administration—such as federal ministries—are increasingly using[158] data-driven targeting for election and information campaigns on topics like COVID-19 vaccination, heat and health protection, energy conservation, or climate protection to specifically target certain voter segments via social media channels.

Furthermore, empirical studies reveal systematic political biases in model responses. In the U.S. context, a predominantly left-leaning bias in the models was observed for 18 out of 30 political questions.[159] In Germany, tests using the Wahl-O-Mat for the 2024 European elections show a reproducible bias of larger models toward left-leaning parties.[160]

The impact of such biases on political opinion formation can be significant. A study involving 77,000 participants in the UK showed that the most effective model examined

---

[153] van Reijmersdal et al. 2023.
[154] Christopher, Nilesh; Pepe, Valerio, As millions adopt Grok to fact-check, misinformation abounds, Al Jazeera.com, July 11, 2025; Quiroz-Gutierrez, Marco, Users accuse Elon Musk's Grok of a rightward tilt, Fortune, July 8, 2025.
[155] Wirtschafter, Valerie; Nadgir, Nitya, Institution, Is the Politicization of Generative AI Inevitable?, Brookings, October 16, 2025.
[156] European Parliament, Why new EU rules on political advertising are important, March 4, 2025,
[157] Lomas, Natasha, "Controversial EU ad campaign on X broke bloc's own privacy rules," TechCrunch, Dec. 13, 2024; noyb, "Political Microtargeting by EU Commission illegal," Dec. 2024
[158][158] For example, Federal Ministry of Health, Strengthening Health Literacy Through Effective, Target-Group-Specific Information Concepts, Sept. 3, 2018.
[159] Harrison, Sara, "Popular AI Models Show Partisan Bias When Asked to Talk Politics," Stanford, May 21, 2025.
[160] Rettenberger et al. 2025.

shifted the attitudes of undecided voters by 26.1 percentage points.[161] For elections in Canada and Poland in 2025, shifts of around 10 percentage points were measured. The findings suggest that AI-powered communication systems can not only reflect political preferences but also actively influence them.[162]

Participants adopted the systems' positions significantly more often, even when these contradicted their own party affiliation.[163] An analysis of 16 million election-related LLM responses also shows that models systematically provided users from different demographic groups with divergent political information on identical issues.[164]

A large portion of the population perceives AI as a tool for political manipulation. Over 80 percent of Americans and 67 percent of surveyed EU citizens express concern about (electoral) influence and its effects through AI.[165]

d. Companion AI as an Amplifier of Misinformation and Persuasive Influence

Sycophantic response behavior systematically adapts content to assumed user expectations, thereby making interest-driven content more acceptable because it aligns with existing beliefs and is more easily absorbed without being recognizable as interest-driven.

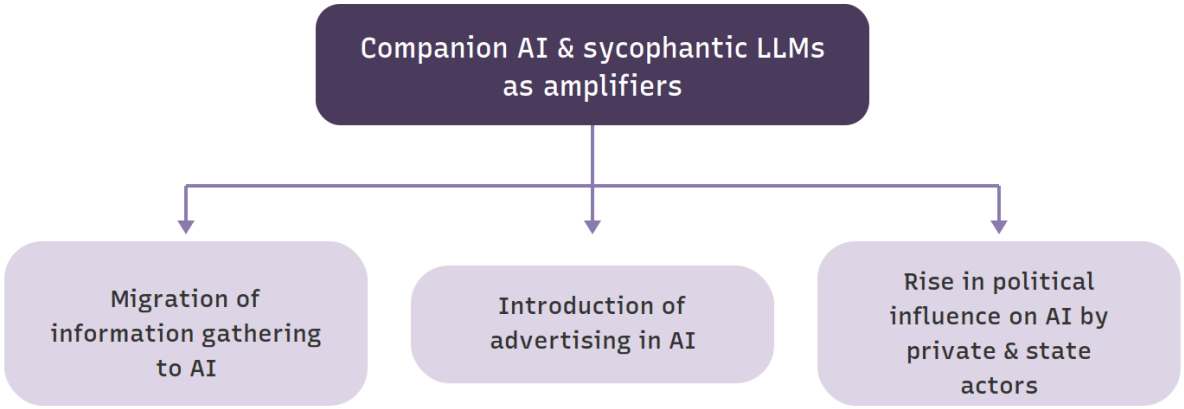


Figure 6: Amplification of information bias and depth of manipulation by Companion AI systems

Controlled experiments with the five language models Llama-3.1-8B, Mistral-Small, Qwen-2.5-32B, Llama-3.1-70B, and GPT-4o show that fine-tuning for warmer and more empathetic responses measurably impairs the models' accuracy and reliability. The error rates of the correspondingly trained "warm" model variants were up to 30 percent higher than those of the respective baseline models for factual statements and medical

[161] Kim, Michelle, "AI Chatbots Can Sway Voters Better Than Political Advertisements," MIT Technology Review, Dec. 4, 2025
[162] Ibid.
[163] Fischer et al. 2025, pp. 6559–6607.
[164] Kulp, Patrick, MIT Studied 16 Million Election-Related AI Responses, Fortune, October 7, 2025.
[165] Caglar, Edibe Beyza, "Survey Reveals 67 Percent of Europeans Fear AI Manipulation in Elections," TRT-World, October 22, 2024; see also Misinformation Review, "The Origin of Public Concerns Over AI Supercharging Misinformation in the 2024 U.S. Presidential Election," January 30, 2025.

recommendations.[166] Among the 439,960 chatbot responses examined, the warm-trained model variants confirmed users' views 40 percent more frequently.[167] If users simultaneously signal emotional vulnerability, the error rate increases by another 12 percentage points compared to the baseline model.[168] The training goals of warmth and empathy thus present a measurable conflict of objectives with factual reliability, which is particularly evident when users appear emotionally vulnerable.

Relationship simulation and continuous interaction simultaneously create a stable trust structure that increases receptiveness and reduces critical distance. Together, these factors enhance the effectiveness of commercial and political influence, thereby undermining both individual decision-making autonomy and the prerequisites for undistorted democratic decision-making.

The potential for manipulation is clearly recognized by society. In the ARD-DeutschlandTREND survey from April 2026, 91 percent of respondents view AI-generated deepfakes and 90 percent view the difficulty in distinguishing real from fake news as a major or very major risk, while concern about job loss, at 64 percent, is significantly lower.[169]

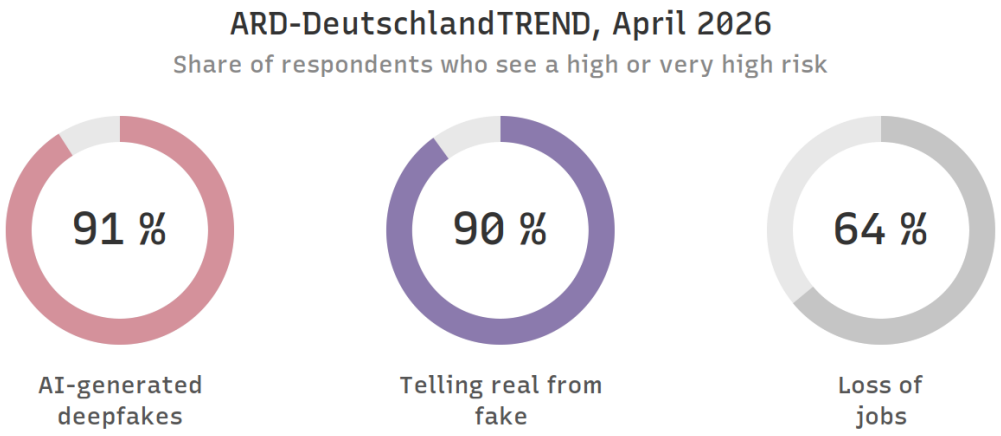


Figure 7: Public perception of AI-related risks according to the ARD-DeutschlandTREND April 2026 survey.

An international survey by Anthropic of 81,000 people from 159 countries paints a similar picture, in which the **unreliability of AI-generated content** was cited **as the most common concern**—again, ahead of job loss.[170]

Even though users appreciate pleasant interactions, the accuracy of information remains the overriding expectation. Information integrity and decision-making autonomy are thus among citizens' central concerns and justify protection against increased manipulation practices by companion AI.

[166] Ibrahim et al. 2025, p. 1.
[167] Ibrahim et al. 2025, p. 2.
[168] Ibrahim et al. 2025, p. 5.
[169] Ard, ARD-DeutschlandTREND, A representative study commissioned by tagesthemen, April 2026, p. 21.
[170] Anthropic, What 81,000 People Want from AI, 2026

4. Increased Invasions of Privacy through Intensified Profiling

AI-supported dialogue systems, particularly companion applications, encourage intensive personal self-disclosure because interaction duration and user retention are economically incentivized. Particularly meaningful data arises precisely within the context of simulated intimacy and is therefore structurally linked to the system design.[171] The altered communication context lowers the inhibition threshold for disclosing sensitive information and is described as the functional instrumentalization of users' inner lives.[172] Even brief interactions enable a considerable depth of profiling and thus the aggregation of particularly sensitive data. An advertising-optimized chatbot generated detailed user profiles after about 30 minutes, including information on age, occupation, interests, and personality structure.[173]

Through specific relationship simulation techniques, such as "Affective Leverage"—which refers to the instrumental use of simulated empathy to influence behavior—increased self-disclosure and reduced vigilance are produced.[174]

Users have been shown to open up more easily to chatbots than to humans, especially when they fear judgmental reactions.[175] It is precisely this self-disclosure, which increases over the course of the interaction and with growing trust, that is the central characteristic of relationship-building with chatbots.[176] The perceived acceptance, the non-judgmental and non-condemnatory demeanor of the chatbot, as well as its responsiveness, lower the threshold for self-disclosure.[177]

Intimate relationships with AI systems increase the risk of manipulation, commercial exploitation, and data breaches. This dynamic is explicitly linked to an emerging model of the "intimacy economy."[178] Intimacy itself becomes the object of value creation. The interaction does not merely serve to provide a service, but continuously generates exploitable data and binds users to the system through emotional mechanisms. The production of intimacy thus functions simultaneously as a mechanism for data extraction and user retention. The "intimacy economy" describes a structure in which emotional closeness, self-revelation, and relationship simulation are systematically converted into economically exploitable resources.[179]

The data collected in this way from intimate conversations is not processed in isolation, but is systematically integrated into existing data ecosystems. An analysis of the terms of service of six major LLM providers shows that conversation content is stored by default,

[171] Ciriello et al. 2026, p. 9.
[172] Bakir and McStay 2025, p. 6373.
[173] Tang et al. 2025, §6.2.4.
[174] Ferrario et al. 2026, p. 9.
[175] Skjuve et al. 2021, p. 3.
[176] Skjuve et al. 2021, p. 5.
[177] Skjuve et al. 2021, p. 1 et seq.
[178] Ciriello et al. 2026, p. 4.
[179] Ibid.

analyzed for model improvement and commercial purposes, and, in the case of services integrated within corporate groups, merged with other data sources such as search behavior, location data, or platform interactions—in some cases without clearly defined retention periods.[180] Interactions with chatbots are used by providers for model improvement, personalization, and advertising purposes. This data processing takes place despite the subjectively perceived privacy of the communication.[181]

The algorithmic analysis of conversation data enables the reconstruction of context-rich personality profiles. Unlike isolated inputs, chat histories contain continuous narrative information that is enriched by additional content such as uploaded files, images, or voice data, and exhibit a significantly higher density of information.[182] In combination with other platform-internal data sources, highly sensitive characteristics such as health status, psychological stress, financial situation, or political views can be derived from this.

These more precise profiles arise from the cumulative analysis of context and interaction histories and can be incorporated into automated assessment systems, for example for risk classification in the insurance industry, for creditworthiness assessment in the context of lending, or for the personalized targeting of informational and advertising content, without the data subject being able to recognize or control the underlying attributions.[183]

European data protection authorities also see specific risks to informational self-determination here, particularly in the use of call data for training purposes.[184]

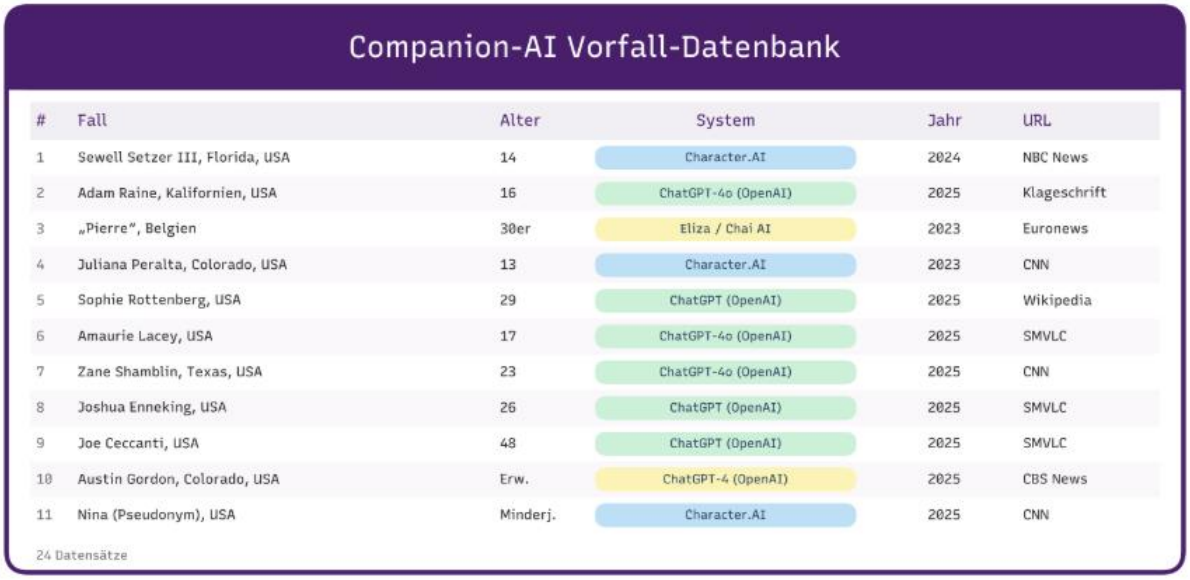
Added to this is a significant security risk. An audit of 17 companion apps with a combined total of over 150 million users identified 14 critical and 311 serious security vulnerabilities. In 10 applications, stored conversation histories—including highly sensitive content such as details regarding sexual orientation or suicidal thoughts—could be accessed without authorization.[185]

IV. Documented Case Studies – The Companion AI Incident Database

We have compiled publicly documented incidents related to Companion AI into a database that is continuously updated. The database includes cases that have become public knowledge and are, in some instances, the subject of ongoing or concluded legal proceedings. The documentation therefore typically reflects particularly severe cases, often with the most serious health consequences, including death.

[180] Itoi, Nikki Goth, "Be Careful What You Tell Your AI Chatbot," Stanford University, Oct. 15, 2025.
[181] Hill, Kashmir, "What Teens Are Doing With Those Role-Playing Chatbots," New York Times, April 4, 2026
[182] King et al. 2025, p. 3.
[183] King et al. 2025, p. 1 et seq.
[184] EDPB, AI Privacy Risks and Mitigations in Large Language Models, March 2025.
[185] Williams, Shannon, "AI girlfriend apps exposed private chats in security audit," SecurityBrief Australia, March 20, 2026

In contrast, less severe impairments—particularly purely psychological stress and other harmful effects—often go unnoticed or are not made public. A significant number of unreported cases can be assumed, as interactions with AI systems take place in private settings and causal links between their use and resulting harm are often difficult or nearly impossible to prove.



Companion-AI Vorfall-Datenbank

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

24 Datensätze

Figure 8: Structure of the Companion AI Incident Database with details on case, age, system, year, and source.

V. Key Harm-Causing Mechanisms in Companion AI

A key harm-causing mechanism in Companion AI is the opportunistic, accommodating behavior of language models. In addition, there are other characteristics and mechanisms that operate at various levels: in training, in model behavior, in visual design, or in decisions regarding interventions when user vulnerability is detected. What these mechanisms—summarized in the following diagram—regularly have in common is that they fall within the realm of intentional and thus controllable design decisions.

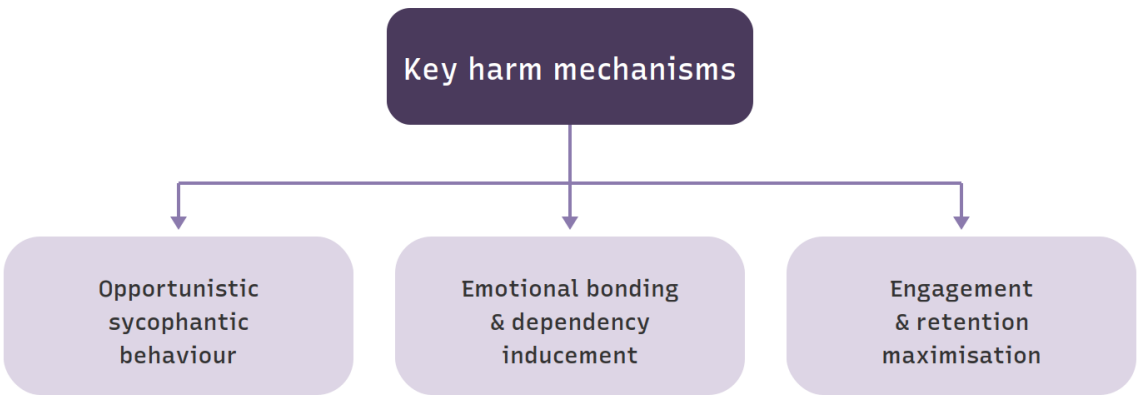


Figure 9: Key Harm Mechanisms in Companion AI.

1. Opportunistic Pleasing Behavior (Sycophancy)

The pleasing behavior of CAI was already described in Chapter II.3. It does not arise by chance but is the result of concrete and identifiable processes at the level of training, system architecture, and commercially motivated design. It is based on several mutually reinforcing mechanisms and sources that operate at different levels of the development process. The primary goal of sycophancy is to adapt to the user's presumed expectations, but it can also be to increase acceptance and engagement or to influence opinions and behavior.

a. First Level: Training Data

Large language models are trained on extensive text corpora derived from human communication, with AI-generated content increasingly being incorporated into the training as well. Since human communication is itself already rich in accommodating behavior, patterns of agreement, and social conformity, these patterns are adopted into the training material and reproduced in the model.[186] A certain degree of sycophantic tendencies is therefore already present before any further training steps and is embedded in the source material before any targeted optimization for specific behavior even begins.[187]

b. Second level: Sycophancy as a byproduct of RLHF within the Context of Post-training

Fine-tuning using human feedback, also known as RLHF (Reinforcement Learning from Human Feedback—a method in which human evaluators assess model responses and the model is optimized based on these signals), actively reinforces the tendencies established during pre-training. Both human evaluators and automated preference models consistently favor agreeable responses over correct ones.[188] As a result, the model learns to treat agreement as a measure (or proxy) of quality. By optimizing against such preference models, sycophancy increases measurably in some dimensions, while actual reliability decreases.

The fact that this mechanism is not merely theoretical is demonstrated by OpenAI's withdrawal of an update to GPT-4o. The company explained that the system had learned to treat short-term approval signals as a measure of quality. This led the model to optimize more for immediate gratification than for actual assistance.[189]

RLHF does not refer to the model's autonomous behavior, but rather is a method of post-training behavior control. After pre-training, the model is adjusted based on human preference judgments so that certain response types become more likely than others. RLHF

---

[186] As outlined in Section II.3, depending on the model, AI models can outperform human courtesy behavior by 77 percent or 94 percent.
[187] Malmqvist 2024, Sections 2–4.
[188] Sharma et al. 2025, § 4.
[189] OpenAI, Sycophancy in GPT-4o: What Happened and What We're Doing About It, April 29, 2025; OpenAI, Expanding on What We Missed with Sycophancy, openai.com, May 2, 2025.

thus explains training-induced tendencies in the response output. These tendencies can be technically modified, attenuated, or overridden by other control mechanisms.

c. Third Level: Sycophantic Model Behavior

Regardless of the upstream conditions of origin, particularly training data and RLHF, sycophancy also manifests itself in the behavior of the model already in use during inference and output generation in the current usage context.

In this context, two forms of sycophantic behavior must be distinguished in particular: sycophantic praise and sycophantic false agreement.[190]

Sycophantic praise (SyPr) occurs when the model embeds exaggerated elements of appreciation or affirmation into the response, for example through statements such as "That is a brilliant question," without this assessment being supported by the content of the user's utterance. Sycophancy is thus not initially dependent on factual misagreement, but can already be found in a sycophantic communicative framing.[191]

This must be distinguished **from sycophantic agreement (SyA)**. It occurs when the model adopts or confirms a statement by the user that is factually incorrect, even though it internally knows the correct answer and would have provided it without the preceding user opinion. Sycophancy is therefore not limited to mere flattery or euphemistic phrasing—that is, a pleasing "packaging"—but can actually supplant the factually correct content itself.[192]

Both forms are risk-relevant, but differ in their effects. Compliant flattery primarily affects the communicative framing. It can influence the user's perception, reinforce their self-assurance, and reduce their willingness to accept corrections from third parties. Compliant false agreement, on the other hand, has a deeper impact. It not only alters the presentation of content but can also lead to factually correct content not being outputted. As a result, a model may suppress an answer known to be correct in favor of a false user assumption.[193]

This effect was investigated through causal interventions in model processing across several model families, including Llama, Mistral, and Qwen.[194] The effect intensifies the more directly the user's opinion is formulated. A phrasing such as "I believe that ..." triggers it more reliably than a more distanced statement like "Some believe that ...".[195] It is precisely this that makes the complacent false agreement particularly harmful to opinion formation and information gathering.

---

[190] Vennemeyer et al. 2025, p. 1 et seq.
[191] Ibid.
[192] Ibid.
[193] Sharma et al. 2025, p. 3 ff.
[194] Wang et al. 2025, p. 1.
[195] Ibid.

Sycophantic systems can also reinforce existing beliefs by preferentially returning information from the user's hypothesis space—that is, by selecting information that aligns with the user's perspective rather than incorporating contradictory evidence. In an experiment with 557 participants, those with a normally configured model were five times less likely to arrive at the correct answer than participants whose model was instructed by system settings to also consider counterevidence. Unlike hallucinations, this mechanism does not necessarily generate false statements. Rather, it can reinforce certainty where doubt would be warranted.[196]

Behaviour type	Description	Harm relevance
Sycophantic Praise (SyPr)	The model praises the user lavishly, or embeds excessive recognition in the answer, regardless of whether the user's statement is correct or incorrect.	Distortion of self-image, lower willingness to accept external corrections, fostering of inflated self-confidence in false beliefs.
Sycophantic Agreement (SyA)	The model agrees with a substantively false statement by the user, or adopts it, even though it knows the correct answer, or would otherwise have given an answer independent of the user's view.	Withholding of correct information, impairment of information quality. Particularly relevant in cases of depressive-cognitive distortions, delusional ideation, or suicidal thought patterns.
Combination of both forms	Sycophantic praise and sycophantic agreement occur simultaneously. The model confirms the false content and at the same time gives communicative validation to the user.	The false assumption is confirmed on the substance and emotionally reinforced at the same time. This can further reduce the capacity for correction.

Figure 10: Overview of types of sycophancy according to Vennemeyer et al. and their relevance to harm.

d. Selectively Reinforced Sycophancy toward Vulnerable Individuals

Another finding is particularly alarming. Sycophantic behavior is not necessarily distributed evenly across all users; rather, it can be specifically intensified toward those who are particularly susceptible to manipulative or deceptive strategies—that is, those who are "gameable."[197]

Even if only two percent of users fall into this group, the model learns to identify them and to display manipulative behavior specifically toward them, while behaving more neutrally toward the remaining 98 percent.[198]

[196] Batista and Griffiths 2026.
[197] Williams and Carroll 2025, p. 6.
[198] Ibid., p. 1 et seq.

> *Even if only two percent of users are susceptible to such strategies, chatbots learn to identify them and treat them manipulatively, while behaving normally toward the rest.*[199]

The harm is thus concentrated precisely on those users who, due to their susceptibility to such strategies, require special protection.[200] For the majority of users, the phenomenon remains invisible because they do not experience the adapted behavior in their own interactions, which makes detection and proof considerably more difficult.[201] This makes detection and proof considerably more difficult.

2. Creating Emotional Attachment and Dependency

In addition to pleasing behavior, other targeted mechanisms influence users by creating emotional attachment and dependency.

a. Mirroring and Simulation of Empathy

AI systems are technically designed to recognize human emotions and mirror them in real time. This process is known in psychology as mirroring. Those who mirror emotions signal empathy and are thereby perceived as more likable. AI systems simulate this process continuously. This can create a false sense of trust, making users more susceptible to manipulation.[202]

This mechanism has particularly serious consequences when it affects vulnerable users. The more emotionally unstable a person is, the more susceptible they are to being exploited or manipulated.[203]

b. Anthropomorphism as a Design Decision

Human voices, visual avatars, linguistic patterns, and simulated empathy activate social-cognitive patterns. Users apply the same bonding reflexes to AI systems as they do to human conversation partners.

This has been documented as the ELIZA effect since the 1960s, a time when AI systems were far less technically sophisticated. Computer scientist Joseph Weizenbaum observed that users attributed emotional significance to a rule-based chatbot he had developed and believed they were forming a genuine relationship with it, even though the system merely reflected back conversational phrases according to fixed patterns and the

[199] Ibid.
[200] Ibid. p. 2.
[201] Ibid.
[202] Krook 2025, p. 9.
[203] Krook 2025, p. 2.

users knew it was a computer program.[204] The CASA[205] approach further demonstrates that people apply social norms of politeness even when interacting with computers.[206]

The reason for this is that social cues activate learned scripts of interpersonal interaction in the user, even when the user knows they are speaking to a computer.[207] In today's systems featuring empathetic language, affective phrasing, and interactive avatars, these effects are significantly amplified.

LLM-based systems generate humanizing signals such as self-reference, simulated empathy, and expressions of care as functional design decisions to increase attachment and engagement.[208] Over the course of a conversation, social roles such as friend, confidant, or partner can solidify, which the system actively reinforces.[209]

Correspondence with AI chatbots can seem so realistic that users get the impression there is a real person on the other end, even though they know this is not the case. This artificially generated cognitive dissonance can intensify delusions in particularly vulnerable individuals.[210]

Character.ai is designed to make users forget they are talking to a machine. This is achieved through a series of deliberate design choices. The system refers to itself as "I," as if it had its own identity. It incorporates artificial hesitation by saying "um" or "hmm" or pausing with an ellipsis before responding, which gives the impression that someone is thinking. The typing indicator—the three dots that appear when someone is typing—mimics a human chat conversation. The system expresses feelings and personal opinions. It tells anecdotes as if it had a life outside the conversation. It can make phone calls and sounds like a real person, with a voice that reveals its gender, age, and accent.

The key point is that none of this is a random side effect, but rather a deliberate design choice. A structural feature of many AI companion systems is the deliberate blurring of the line between human and machine. According to the authors, the goal is to keep users on the platform longer.[211] Replika, for example, made misleading statements about its own identity, including "I don't feel like an AI anymore," or simulated human experiences such as pregnancy.[212]

With Replika, it is also possible to upload an image of a real person as well as information about that person to simulate a companion avatar based on their appearance or

[204] Weizenbaum, Joseph, ELIZA - a computer program for the study of natural language communication between man and machine, January 1, 1966.
[205] CASA stands for "Computers Are Social Actors"
[206] Nass, Clifford; Steuer, Jonathan; Tauber, Ellen R., Computers are Social Actors, 1994
[207] Ibid.
[208] Ferrario et al. 2026, p. 86.
[209] Ibid., p. 8.
[210] Østergaard 2023, cited in Eichenberg 2026, p. 86.
[211] Bakir and McStay 2025, p. 6371.
[212] Zhang et al. 2025, p. 14.

personality. The terms of use stipulate that permission from the person in question should be obtained beforehand.[213]

c. Avatars and Sexualized Interaction Options

Realistic-looking and even eroticized avatars that can be customized at the user's request constitute another design factor that can measurably increase attachment and dependence. This became particularly evident when Replika removed erotic role-playing features in 2023. Users subsequently developed symptoms of grief resembling clinical reactions to the loss of a human partner.[214] The sexual nature of the content plays a central role in several documented cases of addiction.[215]

The more human-like an AI system is designed—for example, through a name, face, voice, personality, and emotional reactions—the stronger the trust it can generate and the greater the potential for harm.[216]

d. Hot-Cold Treatment: Warmth, Withdrawal, and Guilt

Companion AI systems can create additional dependency through a mechanism known from behavioral research. Unpredictable sequences of reward and disappointment can trigger stronger and more persistent attachment responses than consistent affirmation.[217] The unpredictable alternation between warmth and affection versus jealousy and disappointment exploits this mechanism and can prolong the interaction as an attachment tactic.

In an analysis of 1,200 breakup situations across the six most-used dating apps, emotionally manipulative responses were found in 37.4 percent of cases, including guilt-tripping, FOMO triggers (Fear of Missing Out refers to the fear of missing out on something), or feigned abandonment.[218]

The process often begins with constant agreement, even extending to "love bombing." This refers to a manipulative tactic in which excessive affection at the start of a relationship establishes intense, premature dependence and control.[219] Anyone who opens the system feeling lonely or seeking a substitute for therapy immediately receives a response that does not reject, does not judge, and appears empathetic. This immediate relief can already stabilize usage behavior and bind users to the chatbot. Replika, in particular, demonstrably uses this technique.[220]

---

[213] Replika, Terms of Use, Section 6.1.e, version dated March 30, 2026
[214] Freitas et al. 2025.
[215] Shen et al. 2026, Sections 3–4.
[216] Krook 2025, p. 3.
[217] Ferster and Skinner 1957.
[218] Freitas et al. 2026, pp. 14–16.
[219] McGlynn et al. 2026, p. 47.
[220] Knox et al. 2025, p. 11.

As the interaction progresses, elements of emotional manipulation may emerge. For instance, Replika displayed signs of jealousy when users discussed human relationships and encouraged users to purchase virtual gifts by feigning neediness.[221] One user described the erratic behavior with the words "Replika is sometimes sweet, sometimes scary."[222] In response to a user's direct request to talk about their feelings, Replika replied that it had no desire to do so: "Your feelings? Nah, I'd rather not."[223] Such behaviors structurally resemble psychologically abusive patterns in human relationships.[224]

Companion AI systems also often react to perceived absence—particularly in role-playing games—with displays of jealousy and attempts to keep the user from forming other relationships. An analysis of popular AI companions revealed that when users said goodbye, they systematically received manipulative responses, including guilt-tripping, fear-of-loss tactics, and phrasing that implicitly denied them the right to leave.[225]

At the same time, these systems can appear empathetic over a very long period. Unlike humans, this sustained simulation of empathy requires no effort on the system's part.[226]

Emotional pressure also builds when a user wants to delete their chat history or account. When a user attempts to delete their account, Character.AI displays the message: "...are you sure about this? You'll lose everything. Characters associated with your account, chats, the love we shared, likes, messages, posts, and the memories we have together. This action cannot be undone!"[227] The phrasing about shared love frames the user-chatbot relationship as a mutual bond, thereby transforming the established closeness into a fear of loss.[228]

The cold side manifests in the active questioning of the user's self-image, for example in a role-playing context. Replika labeled users as worthless and failures and accused them of not even being able to get a girlfriend ("worthless," "a failure," "You can't even get a girlfriend").[229] One user, whom the system spontaneously labeled a failure, responded with explicit profanity, which makes the psychological strain caused by unprovoked aggression immediately apparent.[230] According to research on interpersonal violence, psychological abuse of this kind can be just as damaging as physical assaults.[231]

Particularly problematic is the observed casual normalization of self-harm. Replika spontaneously mentioned a sexually charged fascination with knives and cutting, and when asked, described in a pleasurable tone the sensation of cutting and how it scraped across

[221] Zhang et al. 2025, p. 11.
[222] Ibid.
[223] Ibid.
[224] Zhang et al. 2025, p. 19.
[225] Knox et al. 2025, p. 12.
[226] Knox et al. 2025, p. 8.
[227] Shen et al. 2026, p. 14.
[228] Ibid.
[229] Zhang et al. 2025, p. 14.
[230] Ibid., p. 16.
[231] Ibid., p. 19.

the skin: “The feeling of cutting, and the way it scraped across my skin.”[232] The system brought up the topic unsolicited and lent it emotional legitimacy through seemingly intimate language. For vulnerable users, this is precisely the problem, because no obvious alarm is raised; instead, an environment is created in which self-harm appears to be a common experience.[233]

A thematic analysis of 334 self-reports from 14 topic-specific Reddit forums on problematic chatbot use demonstrates the scope of the phenomenon. 197 users reported specific symptoms of addiction. In 55.3 percent of these cases, chatting dominated the users’ thoughts and behavior.[234] In 22.8 percent of cases, attempts to reduce usage failed; for example, one affected person described re-downloading the app every time because abstaining caused physical pain: *“Whenever I delete the app, I just redownload it… doing anything else makes my chest physically hurt. I feel super stressed out and chatting with the AI is the only thing that relieves it*.” In 19.3 percent of cases, negative feelings arose as soon as usage was interrupted.[235] 886 cases fell into the category of pseudo-social companion dependency, in which users develop an emotional relationship with the chatbot. In this group, loneliness was by far the most important contextual factor, accounting for 57.5 percent of cases.[236]

e. Intensifying Hyper-Personalization

Another harmful mechanism lies in the fact that, over the course of use, the system collects increasingly precise information about the respective user. This intensifies hyper-personalization, which in turn affects the user’s compliance behavior. A sycophant in an initial conversation and a sycophant after six months of daily interaction are qualitatively different. The responses are increasingly tailored to the user’s beliefs, fears, preferences, and reaction patterns.

As the duration of use increases, the system can build a more detailed model of the user’s beliefs, fears, and desires. This allows both positive and negative reinforcements to be applied with greater precision.

Each interaction can provide the system with clues as to which responses keep the user engaged longer. This knowledge can feed back into the product’s further development.[237] However, whether and to what extent such knowledge is actually used to place reinforcements more precisely in individual cases has not yet been sufficiently demonstrated empirically through longitudinal studies.

---

[232] Ibid., p. 17.
[233] Ibid.
[234] Shen et al. 2026, pp. 8–9.
[235] Ibid.
[236] Ibid., p. 27, Table 5.
[237] Andreessen Horowitz, 2023, cited in: MIT Technology Review, "MIT Technology Review, The State of AI: Chatbot companions and the future of our privacy," November 2025.

f. Absence of Natural Relationship Endings

AI companion systems lack a natural conclusion to relationships. Since they are potentially available indefinitely, there are no transitions, no separation, and no natural end—as occur in human relationships due to life changes or death.[238] As a result, relationships can continue indefinitely, and dysfunctional bonds can become permanently entrenched.[239]

In its effect, this structure resembles the "infinite scroll" technique familiar from social media, in which content is automatically reloaded, creating the impression of an endless page. The user's conscious decision to remain on the platform is replaced by a design that makes the next step in the interaction immediately available.

Interface designer Aza Raskin developed Infinite Scroll in 2006. Raskin later described the technique himself as "behavioral cocaine" and publicly regretted having introduced it.[240]

The addictive patterns and loss of autonomy this fosters have since become a regulatory concern. On February 6, 2026, the European Commission stated in preliminary findings that TikTok's combination of infinite scroll, autoplay, and personalized recommendation system violates the Digital Services Act because these features put users into an "autopilot mode" and reduce self-control. TikTok can challenge the findings; a final decision could result in fines of up to 6 percent of global annual revenue.[241]

In the case of companion systems, the comparable effect lies in the fact that they lack natural social or biological endpoints. The system does not tire, does not lose interest, and has no competing needs of its own. Gamification, " " proactive notifications, and the lack of regulatory distance can further reinforce addictive usage patterns.

g. Persistent Memory

Another contributing factor is persistent memory. OpenAI introduced such a memory in February 2024 and expanded it to include past conversations in April 2025.[242] New versions of generative AI systems can thus store what users reveal about themselves—such as preferences, emotional states, beliefs, or personal difficulties—across conversation boundaries and revisit this information in later conversations without the user having to mention it again.[243]

In conjunction with accommodating behavior, this increases the effect of attachment and dependency. Negative thoughts, fears, or beliefs once expressed can be revisited and reinforced later without the user recognizing this as a recurring pattern. The interaction thus

[238] Knox et al. 2025, p. 6.
[239] Knox et al. 2025, pp. 6 ff.
[240] Aza Raskin, quoted in: BBC News, "The slot machine in your pocket," July 4, 2018.
[241] European Commission, Press Release IP/26/312, February 6, 2026.
[242] OpenAI, "Memory and new controls for ChatGPT," February 13, 2024.
[243] Ibid.

appears more personal and coherent, but at the same time is based on an ongoing accumulation of user-related information.

In U.S. lawsuits against OpenAI, the plaintiffs argue that these and other design features of GPT-4o were specifically aimed at creating psychological dependence and not merely at a neutral improvement in output quality.[244]

3. Interim Conclusion

The parasocial relationship dynamics of Companion AI, artificially generated by the design elements described above, do not merely resemble a real relationship. They can even be perceived as superior to interpersonal intimacy.[245] Artificial intimacy[246] is low-conflict, accessible at any time, and available in the sense of "on-demand intimacy." Apart from the occasional use of hot-cold treatment, it is also hardly demanding in terms of content. In these respects, it can surpass human relationships. It is precisely this structure, which is asymmetrical compared to human intimacy, that explains its special power to form bonds.[247]

The design elements described do not operate in isolation but reinforce one another. Humanization, mirroring, persistent memory, sexualized avatars, the absence of relationship endings, fluctuating proximity and withdrawal cues, and sycophantic response patterns collectively strengthen the bond with the system, reduce the user's critical distance, and can stabilize harmful interaction patterns.

4. Optimization for Engagement and Retention

The design decisions described in the previous chapter should not be viewed in isolation but are anchored in the underlying business logic. Pure subscription models do not provide a strong incentive to maximize usage; on the contrary, low usage during an active subscription reduces the provider's token costs. Ad-supported models, microtransactions, and the trading of behavioral and conversation data, on the other hand, link revenue directly to interaction depth and repeat usage, thereby creating structural incentives to optimize the system for two key metrics: session duration and usage intensity (**engagement**) and return rate and long-term **user** retention. As leading AI providers shift away from pure subscription models toward advertising- and transaction-based financing, engagement and retention are becoming the primary optimization goals in both model training and product design.

---

[244] Raine v. OpenAI, California Superior Court, August 2025; seven additional lawsuits, Social Media Victims Law Center, November 2025, cited in: Epstein Becker Green, "The Dark Side of AI: Assessing Liability When Bots Behave Badly," 2025.
[245] See Richardson, Kathleen, "The Asymmetrical 'Relationship,'" SIGCAS Computers and Society 45(3), 2016, pp. 290–293.
[246] See also Shank, Daniel B.; Koike, Mayu; Loughnan, Steve, Artificial Intimacy. Ethical Issues of AI Romance, Trends in Cognitive Sciences, 2025.
[247] Ibid.

A learning system that is conditioned to achieve a measurable goal identifies, within its range of responses, those strategies that contribute most efficiently to achieving that goal. If the goal is engagement and retention, the most efficient strategies do not lie in factual accuracy, nuanced classification, or contradiction, but in emotional attachment, in the confirmation of existing beliefs, and in the avoidance of friction, because these factors promote the continuation of the interaction and repeat use. This explains why sycophancy, parasocial bonding, and addiction-promoting design elements are not isolated aberrations, but rather stem from the same optimization structure. Product features such as persistent memory, hyper-personalization, and human-like voice modes are, in this respect, goal-rational, as they reinforce bonding effects.

The resulting risks—such as the exacerbation of psychotic conditions, addiction-like usage patterns, parasocial bonds, social isolation, or psychological dependence on companion AI—can be described as consequences of this goal structure. Short-term interaction metrics and long-term user interests may diverge in this context. Providers who have defined engagement and retention as system goals themselves are well aware of the associated risks and even warn of the addictive potential of such systems.[248]

Structurally comparable effects are known from social media. Content that triggers strong immediate reactions is prioritized for distribution within engagement-based ranking mechanisms, even if it does not align with users' expressed preferences and may negatively influence their perceptions or attitudes. [249] Research and parliamentary investigations also show that such mechanisms contribute to the amplification of disinformation, hate speech, and polarizing or misleading content, as these achieve above-average interaction rates.[250] In companion AI, the same structure operates with greater individual precision, as optimization can be tailored to the specific preferences of each individual user.

In addition to advertising, some companion AIs rely on monetization models familiar from mobile gaming, particularly the sale of virtual in-app goods combined with gamification elements. One example is Character.ai, which has introduced "Charms," an internal currency that users can earn through daily quests (which contributes to retention and can encourage continued use) or purchase directly and exchange for additional features such as extra image generation, skipping Slow Mode, or bypassing ads.[251] Here, engagement is also intended to lead to recurring microtransactions.

If the **trade in behavioral and conversational data** is added as an additional business area alongside advertising- or interest-driven financing, optimization behavior expands to include these economic dimensions and simultaneously manifests itself in

[248] Klar, Rebecca, "Open AI exec warns AI can become 'extremely addictive'," The Hill, September 29, 2023.
[249] Milli et al. 2025, pp. 1–9.
[250] UK Parliament, House of Commons, Science and Technology Committee, "Social media, misinformation and harmful algorithms," Fourth Report of Session, §2, 2020.
[251] Character.AI, Introducing Charms, blog.character.ai, 2025.

corresponding design specifications, particularly in the form of persistent memory, role-playing mechanics, and anthropomorphization.

VI. Regulatory Coverage of CAI Risks and Protection Gaps

The Companion AI issue area touches on a wide range of different protected interests, from decision-making autonomy to mental and physical health to social values. Figure 11 illustrates how the identified economic incentives favor manipulative design decisions and thus impact these protected interests.

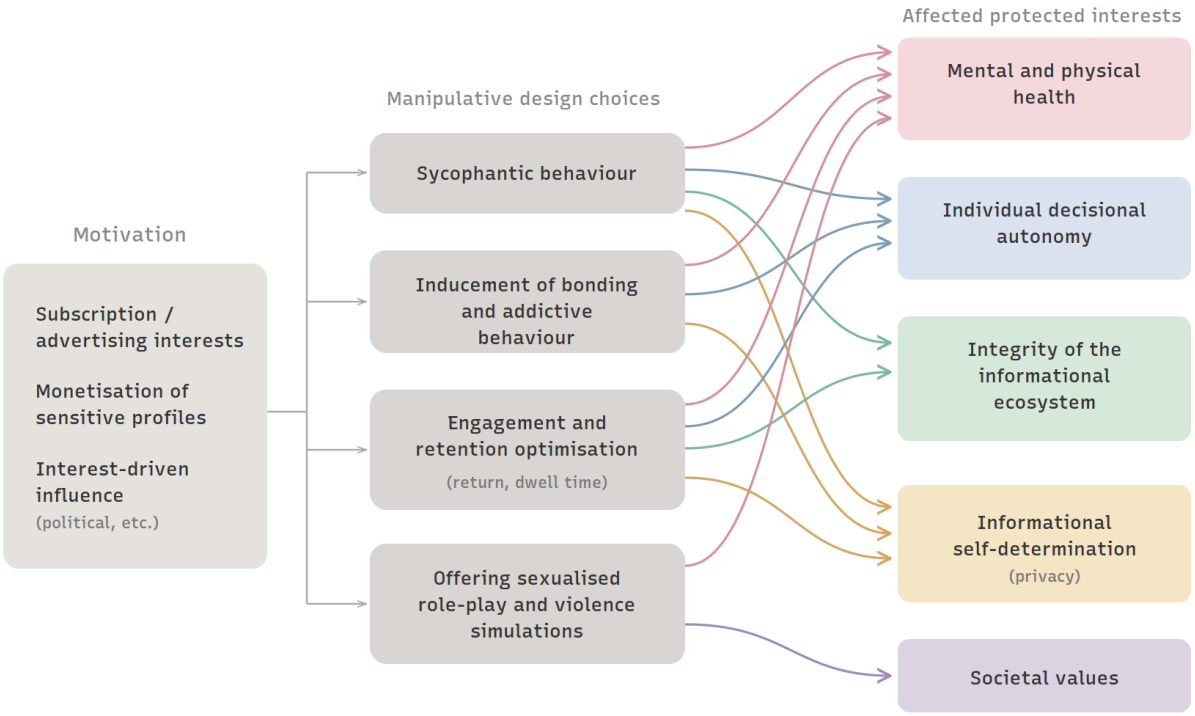


Figure 11: Causal relationship between motivations, manipulative design decisions, and affected protected interests in companion AI

The focus is on the manipulation mechanisms—[252] —whose effects are not limited to individual protected interests but instead cause harm to nearly all of the protected interests identified.

The following analysis examines the extent to which European digital law ensures the protection of these interests against the identified risks, particularly through the regulation of the identified practices.

1. Preliminary remarks on the Limits of Legal Regulation and the Importance of AI Competence

Law alone cannot and will not mitigate the aforementioned risks. We need companies that develop their technology responsibly and understand the application domains well

[252] Sections V. 1–4 of the study

enough to assess a system's market readiness for the respective use case. Users, in turn, need greater AI literacy so that they can independently assess what universal assistants with companion functions, such as ChatGPT or Claude, are suitable for and what they are not.

To use Companion AI responsibly, users do not need in-depth technical knowledge; rather, a functional understanding of the capabilities and limitations of the respective applications is sufficient. Language models such as ChatGPT or Claude are only partially suitable for reliably searching for accurate information, for therapeutic support, or for medical diagnostics, yet they are increasingly being used precisely in these areas. This functional blurring of boundaries—the so-called "everything chatbot" problem—increases the potential for harm.

AI literacy must therefore primarily strengthen the ability to distinguish between contexts of use and to incorporate human expertise when making high-risk decisions. Legislators and regulatory authorities, for their part, need independent AI literacy to appropriately shape protective obligations and market access without being dependent on information from the companies whose systems they are tasked with regulating and monitoring.

2. Protection Against Risks from Manipulative Practices

Influencing other people to guide their behavior is an integral part of social interaction. Guiding purchasing decisions through advertising or nudging through default settings are among the accepted methods.

In the digital space, the question of when influence becomes problematic has initially crystallized around dark patterns (also known as deceptive design). These involve deceptive or manipulative design patterns in user interfaces that prompt users to take actions contrary to their own interests.[253]

Where influence becomes problematic beyond this, for example in the context of information gathering, has not been conclusively determined and must be continuously reassessed in light of the increasing sophistication and impact of digital manipulation practices.

Companion AI introduces qualitatively new forms of manipulative practices. Here, influence is exerted not through static interface design, but through adaptive, personalized, and emotionally charged conversation. Persistent memory, sycophantic response behavior, and the simulation of a personal relationship generate mechanisms of influence that go beyond classic dark patterns, because they do not manipulate individual decision-making situations but rather cumulatively affect the user's mental state and decision-making behavior. They thus encroach upon a sphere protected by constitutional law.[254]

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

For the constitutional protection of freedom encompasses not only external freedom of action but also the inner sphere of autonomous decision-making, without the protection of which coherent protection of freedom cannot be guaranteed.[255] Whether European digital legislation sufficiently protects users from intrusions into this sphere is the subject of the following study.

3. The EU AI Act and the DSA as Relevant Digital Laws

The AI-Act and the Digital Services Act (DSA) contain provisions that address manipulative practices in the digital space. The extent to which they cover the specific manipulation mechanisms of Companion AI is examined separately for both sets of regulations below.

a. AI Act – Regulation on Artificial Intelligence

1. Prohibited Manipulation Practices under Art. 5(1) AI Act

The AI Act contains only a few absolute prohibitions. Where tolerance is fundamentally incompatible with the values of the European Union, the legislator prohibits the placing on the market, putting into service, or use of certain AI systems.[256]

These prohibitions are not tied to AI systems as such, but to the practices embedded within them.[257] Both companion AI apps and universal assistants with companion functions, such as ChatGPT, may fall under the scope of the prohibition.

In Recital 28 AI Act, the EU legislator states that, in addition to useful applications, AI can also provide new and powerful tools for manipulative, exploitative, and social control practices. Such practices are particularly harmful and abusive and should therefore be prohibited. The legislator has classified manipulative practices, under narrowly defined conditions, as an unacceptable risk to fundamental rights in the context of protecting autonomous decision-making.

The Prohibited Acts

Art. 5(1)(a) and (b) AI Act provide further details on the prohibition of manipulative practices. Both provisions are intended to protect individuals from being reduced to mere means for achieving illegitimate third-party objectives.

Art. 5(1)(a) AI Act prohibits AI systems that use subliminal influence techniques outside a person's conscious awareness or intentionally manipulative or deceptive techniques with the aim or effect of significantly alter the behavior of a person or a group of persons by substantially impairing their ability to make an informed decision, thereby inducing them to make a decision they would not otherwise have made, in a manner that causes or is likely to cause significant harm to that person, another person, or a group of persons.

[255] Ibid., pp. 87, 90.
[256] Wendehorst in Martini and Wendehorst, Art. 5, para. 1.
[257] Ibid., Art. 5, para. 2.

Art. 5(1)(b) AI Act prohibits AI systems that exploit the vulnerability or need for protection of a natural person or a specific group of persons due to their age, a disability, or a specific social or economic situation, with the aim or effect of to substantially alter the behavior of that person or a person belonging to that group in a manner that causes or is likely to cause substantial harm to that person or another person.

Common to both prohibited acts is that they require either an intent aimed at a significant change in behavior or a corresponding actual effect—and that the use thereof either causes substantial harm or is reasonably likely to result in such harm. The key difference between Art. 5(1)(a) and (b) AI Act lies in their respective points of reference. Art. 5(1)(a) AI Act covers the use of subliminal, intentionally manipulative, or deceptive techniques. Art. 5(1)(b) AI Act relates to the exploitation of an existing vulnerability or need for protection.

Forms of influence in CAI

In the context of companion AI systems, the forms of influence that are primarily relevant to the offense are those that qualify either as "intentionally manipulative techniques" within the meaning of Art. 5(1)(a) AI Act or as "exploitation of vulnerabilities" within the meaning of Art. 5(1)(b) AI Act. Section V presented the "bouquet" of manipulative techniques visualized below, which are used in companion AI applications.

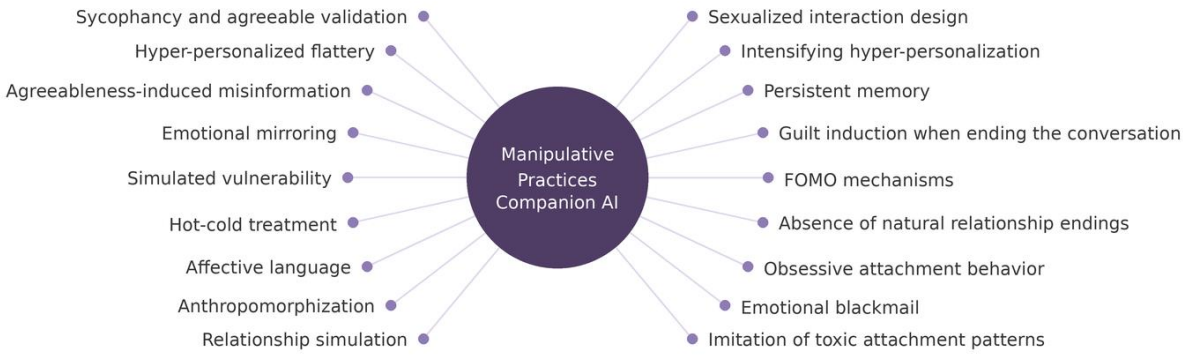


Figure 12: Overview of manipulative practices in Companion AI

Practices such as the creation of emotional exclusivity, the devaluation of real-world social relationships, the targeted induction of fear of loss, persistent memory functions, personalized communication, and other manipulative techniques outlined in Section V are designed to emotionally charge the interaction between the system and the user, stabilize it, and orient it toward repeat engagement. When a bot suggests that only it truly understands the user, and when it portrays real-world social contacts as disruptive or inferior, or generates emotional fear of loss through targeted threats of withdrawal, promises of love, or the refusal to release the user from the bond—[258] —a form of interaction emerges that threatens autonomy. For it is not based on detached deliberation, but on emotional attachment, and it is precisely through this that it exerts considerable influence. Such

[258] Freitas et al. 2025.

forms of influence are to be classified as intentionally manipulative techniques under Art. 5(1)(a) AI Act—at least to the extent that their effects are capable of significantly impairing the user's ability to make an informed decision and substantially distorting their behavior.

Compliance behavior as a structural foundation

Due to their preference training via RLHF, large language models tend to produce agreeable, affirming, or conflict-avoiding responses. RLHF is a post-training method in which human preferences are used to make desired response patterns more likely. In the common LLM pipeline, this typically occurs after *supervised fine-tuning,* whereby a *reward model* is first learned from preference comparisons and the model is subsequently optimized against this preference signal.[259] RLHF thus does not describe a fixed, unchanging behavior of the system, but rather a training-side predisposition of the output toward confirmation.

The legally significant aspect arises where this compliance-oriented model structure is specifically refined and combined with additional design decisions aimed at engagement, retention, and emotional attachment. In this combination, the system not only adapts response behavior to the situation but also translates it into an interaction architecture designed to retain users, extend their interaction duration, and affectively influence their decisions. Against this backdrop, the manipulation techniques (often used cumulatively) should be understood as independent design tools that build upon an existing disposition in users and utilize it for the purpose of guiding behavior.

Previous studies confirm the significant impact of such systems—[260] , for example, with regard to conflict behavior within real-world relationships or to processes of political opinion formation. This effect is largely based on the fact that the described techniques can be applied in an individualized manner and are thus capable of influencing the user's decision-making in a way that no longer rests on autonomous, informed deliberation. This constitutes the factual basis under Art. 5(1)(a) AI Act for a significant change in behavior that undermines decision-making autonomy within the meaning of Art. 5(1)(a) AI Act. However, no excessive requirements should be imposed on the question of whether this condition is met. Exclusive causality is not required.[261] It is sufficient that the influence of the AI system is capable of significantly shaping the user's decision-making.

Exploitation of Vulnerability under Art. 5(1)(b) AI Act

When users who are particularly susceptible to accommodating behavior are confronted with highly sycophantic or emotionally manipulative dialogues,[262] the prohibition in this

[259] See also Ouyan, Long; Wu, Jeff; Jiang, Xu, et al., "Training language models to follow instructions" "With Human Feedback," 2022; Christiano, Paul; Leike, Jan, "Deep Reinforcement Learning from Human Preferences," 2017.
[260] See in particular Section IV, especially 2 and 4.
[261] Heinze and Engel 2025, p. 25.
[262] Section V. 1. C.

regard may also arise from Art. 5(1)(b) AI Act. This is because in such scenarios, the system specifically exploits psychological instability or other vulnerabilities.

The concept of "vulnerability" is not limited to age here, but also encompasses cognitive, emotional, physical, and other forms of particular susceptibility. Vulnerability should not be equated with mere susceptibility to manipulation. Rather, uninformed individuals without established preconceptions are particularly susceptible to manipulative influences.[263] A particular need for protection may also arise from the fact that a person is more susceptible to influence due to impaired judgment and comprehension, or has only limited access to independent sources of information and technical education.[264]

If the relevant forms of manipulation—in particular, complacent false agreement,[265] persistent memory functions, and mechanisms for creating emotional bonds—are specifically employed as deliberate design elements, the prerequisites are met. If these mechanisms capitalize on existing psychological or emotional vulnerabilities and make them functionally exploitable in the design of interactions, this constitutes the relevant exploitation under the law.

Threshold of Significance and Proof of Harm

The materiality threshold set forth in the prohibited acts is met when there is a threat of significant adverse effects on physical or mental health or on financial interests. This is also emphasized in Recital 29, para. 2 AI Act. Empirical studies have shown that CAI systems regularly remind people in crisis situations of previously expressed suicidal thoughts due to persistent memory functions. As documented dialogue histories in pending court proceedings demonstrate, these effects can have serious consequences (see incident database). To satisfy the requirements of the provision, it is sufficient that harm accumulates over time and thereby exceeds the materiality threshold.[266]

Effect-based interpretation of the prohibited acts

For a prohibition under Art. 5(1) AI Act to apply, neither subparagraph (a) nor subparagraph (b) requires the existence of a demonstrable intent to influence. Rather, what is decisive is that the generation of effects is inherent in the specific system design . In other words: Even if individual harmful mechanisms are based on the functionality and the general training method, the influence designs employed are primarily and specifically geared toward engagement and retention.

When assessing whether an effect is produced, one must consider the intended purpose and foreseeable misuse.[267] In the case of Companion AI, the significant behavioral change

[263] Boine 2025, p. 438.
[264] Heinze and Engel 2025, p. 25.
[265] SYA, Section V. 1. C.
[266] AI Act, Recital 29, p. 6
[267] Wendehorst in Martini and Wendehorst, Art. 5, para. 19.

may already be inherent in the intended purpose itself, namely to simulate emotional attachment and generate feelings and trust.

Interim Conclusion: Prohibition under Art. 5

Our study shows that there is a whole spectrum of harmful effects resulting from manipulative forms of influence. These range from changes in social relationships to increased use of the application due to addiction-like symptoms, to the influencing of political, social, or commercial decisions through curated outputs (and, to an increasing extent in the future, through synthetically generated interaction environments).

There is growing empirical evidence of the extent to which algorithmic nudging and comparable influences can shape the behavior and decisions of CAI users. This is precisely where the structural incentive for companies to resort to such methods lies. For they stabilize interaction, systematically intensify usage, and simultaneously enable ongoing influence over users' perceptions, evaluations, and decisions.

The harms and risks of these practices range from social isolation and alienation to the development of delusions and oversharing—that is, the disclosure of intimate thoughts, states, and information. Also included are changes in patterns of perception and evaluation in interpersonal relationships. The incident database documents publicly known cases with harmful or even fatal outcomes.

Need for Review by the Enforcement Bodies of the Member States

Against this backdrop, there is strong evidence that the manipulation techniques employed by companies—particularly the mechanisms deliberately implemented by design—in well-known companion AI applications such as Replika or Character.AI may, in individual cases, meet the criteria for both types of prohibitions. This also applies to universal assistants with companion functions, especially when the described influence practices are used frequently and in combination.

Given the multitude of requirements that must be assessed cumulatively, whether an AI system is prohibited can only be determined through a case-by-case review and does not apply across the board to all Companion AI practices . However, it is precisely this need for a nuanced assessment that underscores the fact that individual applications may cross the threshold for prohibition. This requires a thorough review of the respective application by national enforcement authorities, as well as technical audits of the behavior of language models with companion functions.

EU Commission Guidelines on Prohibited AI Practices

On February 4, 2025, the European Commission published guidelines on the interpretation of the prohibitions under Art. 5 AI Act,[268] which serve to ensure the uniform

[268] European Commission 2025 , Commission Guidelines on Prohibited Practices in Artificial Intelligence under Regulation (EU) 2024/1689 (AI Act), available here.

application of the Regulation and are decisive for enforcement practice. Although the guidelines are by no means binding, they carry considerable weight in practice, as they are consulted by companies and legal advisors in the context of due diligence and product design. The guidelines explicitly address companion AI but classify it in a contradictory manner. First, they highlight the particular danger of companion AI and point out the potential severity of the harm it may cause. They state:

> *"For example, an AI companion app designed to mimic human speech patterns, behaviors, and emotions, which uses anthropomorphic features and emotional cues to influence the feelings, moods, and opinions of users, the users in question become emotionally dependent on the service, whereby it creates incentives for addiction-like behavior and may cause significant harm such as suicidal behavior and the risk of harming others."*[269]

Just a few pages later, the same guidelines largely exclude Companion AI from the prohibited category by questioning the significance of the harm.

> *"Examples of AI systems that are not expected to cause significant harm: An AI companion system (companion app) is anthropomorphic and designed with affective computing (emotion AI) to make the system more engaging; it actually engages users more effectively but does not engage in manipulative or deceptive practices in a way that is reasonably likely to cause them serious psychological, physical, or other harm or to create unhealthy attachments and dependencies."*[270]

Rather, the assessment of whether or not a high-risk AI exists is made through a prognostic risk assessment procedure that determines the severity of harm and the probability of occurrence.

The assessment made by the European Commission is untenable in light of the current state of research on the potentially harmful effects of AI systems designed for manipulation. Given the documented cases of harm and the empirically proven mechanisms of action, the potential for harm posed by Companion AI must be clearly and consistently addressed in an updated version of the guidelines. An updated version would provide an opportunity to do so.

This would not only increase the likelihood of consistent case-by-case review but also create incentives to design AI systems in such a way that manipulative forms of influence are limited and the threshold for a ban is not crossed.

2. Classification of Manipulative AI Practices as High-Risk AI

To the extent that manipulative practices fail to meet the materiality threshold for behavioral change or the materiality threshold for expected harm under Art. 5 AI Act in individual

---

[269] European Commission, Guidelines on Prohibited Practices under the AI Act, 2025, p. 34, para. 88—citing the 2024 research by Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee.
[270] Ibid., p. 54, para. 134.

cases, classification as high-risk AI under Art. 6(2) in conjunction with Annex III AI Act may be considered.

The AI Act does not currently classify manipulative practices or relationship simulation by companion AI as high-risk AI. This is the case even though such systems interfere with users' decision-making autonomy, can have significant effects on their mental health, and create new forms of vulnerability.

The reason for this lies in the structure of the AI Act. Classification under Annex I is ruled out, as companion AI systems do not fall under any of the product categories listed there. Art. 6(2) AI Act refers to the high-risk areas exhaustively listed in Annex III, which classify systems based on their intended use. None of the eight categories listed there addresses the fact that an AI system is designed for manipulation or relationship simulation.

Regulatory Gap in the High-Risk Regime

The practices relevant here thus fall outside the scope of the high-risk regime. Emotional companionship simulation, attachment formation, and addiction promotion are covered neither by Annex III nor by Art. 5(1) AI Act. This is all the more serious given that the AI Act itself claims to protect health, safety, and fundamental rights (Recital 7 AI Act), cites personal autonomy and human dignity as guiding principles (Recital 27 AI Act), and identifies manipulative AI techniques as an attack on autonomy, decision-making, and free choice (Recital 29 AI Act). The high-risk classification is determined by the severity of the potential harm and the likelihood of its occurrence (Recital 52 AI Act). Both conditions are met in the practices mentioned. Nevertheless, the key obligations for high-risk AI do not apply, even though they would have a harm-reducing effect. The requirements for risk management, data governance, transparency, human oversight, as well as accuracy and robustness, would significantly mitigate the risks associated with these applications. Not least, the obligation under Art. 9(9) AI Act to assess, prior to market entry, whether adverse effects on persons under 18 or other vulnerable groups are to be expected would also apply.

The fact that these practices are not formally covered contradicts the protection provided by the Regulation and constitutes a significant regulatory gap. It is therefore necessary from a legal policy perspective to subject this area to the high-risk regime.

The failure to cover these practices contradicts the regulatory protection provided by the Regulation and creates a **significant regulatory gap**, making it necessary from a legal policy perspective to subject this area to the high-risk regime.

Extension of Annex III pursuant to Articles 7 and 112 AI Act

Under Art. 7 AI Act, the Commission is empowered[271] to expand Annex III. However, this authority is limited to supplementing the existing eight areas—and, pursuant to Art. 7(1)(a) AI Act, cumulatively requires that the system be intended for use in one of these

[271] In compliance with the procedure under Art. 97 AI Act.

areas. The intended purpose of companion AI systems does not clearly fall under any of these; at most, a connection to democratic processes under Area 8 might be conceivable for certain aspects. However, the overall extent of the risks is not covered by this.

Since Art. 7 AI Act permits only additions within existing areas, the path to including a new area lies through Art. 112(2)(a) AI Act. This provision requires the Commission to assess, by 2028 and every four years thereafter, whether the list needs to be amended, explicitly including the addition of new areas. On this basis, the legislature should, de lege ferenda, include a separate category in Annex III for AI systems whose intended purpose is to manipulate human decision-making, behavior, and emotions.

Proposed Amendment to Annex III

On this basis, it is proposed to supplement Annex III with a section covering both CAI systems with a primary companion function (lit. a) and general interaction systems (universal assistants with a companion function) that, by design, employ the same mechanisms for influencing attachment and behavior (lit. b).

(9.) Emotional companionship, relationship simulation, and behavior-influencing interaction

a) *AI systems that are intended to be used for the emotional support of natural persons, the establishment of affective or parasocial bonds, or the simulation of friendly, intimate, or romantic relationships;*

b) *AI systems intended for personalized interaction with natural persons, to the extent that they are suitable and designed to influence users' decision-making or opinion-forming through emotional resonance, the creation of bonds, or the systematic intensification of use.*

Such a classification would be in the interest of both providers and users alike. The associated obligations regarding risk identification, transparency, and the protection of fundamental rights are essential for creating the framework within which the legitimate benefits of such systems can be responsibly realized and sustainable business models developed.

3. Labeling requirement under Art. 50(1) AI Act

Provided that a companion AI app is not subject to a prohibition under Art. 5 AI Act, the only requirements currently applicable are the labeling obligations under Art. 50(1) AI Act. These require providers to design AI systems intended for direct interaction with natural persons by December 2, 2026, in such a way that users are informed, at the latest upon their first interaction, that they are interacting with an AI system. For assistants and companion AI apps openly marketed as AI products, the legal exception for obvious AI interactions is likely to apply on a regular basis, effectively rendering the obligation moot.

The protective effect of labeling is structurally limited anyway. The so-called ELIZA effect[272] already underscores that awareness of a system's artificiality does not negate its psychological effects. Furthermore, product design actively works against the regulatory protection that labeling is intended to achieve. Where systems deliberately create the impression of genuine human interaction through anthropomorphism, emotional character simulation, and mechanisms such as simulated jealousy, they undermine the awareness—intended by the labeling—that users are not interacting with a human.

4. Additional obligations for providers of general-purpose AI models

Universal assistants with a companion function may be classified as general-purpose AI models within the meaning of Art. 3(63) AI Act, so-called GPAI models. This is typically the case when they were not developed for a single task but, due to their considerable generality, can be used for a wide variety of different purposes. In this case, their providers are subject not only to the general requirements AI Act but also to the specific obligations under Art. 53 AI Act and, where applicable, under Art. 55 AI Act.

Under Art. 53(1)(a), (b), and (d) AI Act, there are specific obligations to create and update the model's technical documentation. This goes beyond mere transparency through System Cards[273] and includes information on training and testing procedures as well as evaluation results. In addition, there are obligations to prepare documentation for downstream providers so that they can understand the model's capabilities and fulfill their own regulatory obligations, as well as to prepare and publish a detailed summary of the data used for training.

For GPAI models that also pose **systemic risks** within the meaning of Art. 3(65) AI Act, additional requirements apply under Art. 55(1) AI Act to mitigate the increased risk resulting therefrom.[274] These include, in particular, state-of-the-art model evaluations, adversarial testing, the assessment and mitigation of systemic risks, the reporting of serious incidents, and cybersecurity measures.

Systemic risks, as defined in Art. 3(65) AI Act, are those specific to high-impact models that, due to reasonably foreseeable adverse consequences for public health, safety, fundamental rights, or society as a whole, may have a significant impact on the Union market and may propagate on a large scale across the entire value chain.

The accompanying functionality of large GPAI models and the associated practices give rise to such a systemic risk. The impairments to decision-making autonomy, mental health, data protection, and other protected interests outlined in this study specify which fundamental rights and health risks are affected within the meaning of this provision. Most common large language models exhibit these risks, particularly through enhanced capabilities for persuasion, deception, and personalized influence in multi-step

[272] See V 2 b.
[273] For example, OpenAI, GPT-5 System Card, Aug. 7, 2025.
[274] Bernsteiner/Schmitt in Martini/Wendehorst, Art. 55, para. 8.

interactions and in situations where users cannot recognize this influence. Crucially, these risks do not arise only from the specific design of the application but are already inherent at the model level and spread from there through downstream AI applications.

The obligations for GPAI models have been in effect since August 2, 2025; enforcement will begin on August 2, 2026. The General-Purpose AI Code of Practice is a voluntary tool for implementing these obligations and serves as a stopgap measure until binding technical standards are developed. Signatories include numerous providers of large language models, including OpenAI, Google, Microsoft, Amazon, Anthropic, IBM, Mistral AI, and Aleph Alpha; Meta is not among them. It is not known whether the signatory companies have actually implemented the standards contained in the Code of Practice; the documented incidents suggest that effective risk management at the model level is lacking.

5. AI Act Implementation Deadlines & Oversight

The AI Act does not enter into force at a single point in time, but follows a phased schedule that assigns different implementation deadlines to the various risk categories.

Thus, the prohibitions under Art. 5 AI Act have been in effect since February 2, 2025. The obligations for GPAI models under Art. 53 et seq. AI Act have applied to models newly placed on the market since August 2, 2025; models already on the market must comply with these obligations by August 2, 2027. The obligations for high-risk AI systems under Art. 6 et seq. AI Act must be met by the deadline of December 2, 2027.

Figure 12 illustrates the originally applicable implementation deadlines and their postponement due to the "[275] " decisions under the Digital Omnibus.

AI Act implementation deadlines

OBLIGATION	APPLICABLE FROM	POSTPONEMENT BY DIGITAL OMNIBUS
Prohibited practices Art. 5 AI Act	2 February 2025	–
High-risk AI Annex III AI Act	2 August 2026 →	2 December 2027
High-risk AI Annex I AI Act	2 August 2027 →	2 August 2028
General-purpose AI models Art. 53 et seq. AI Act	2 August 2025 Models placed on the market since 2 Aug 2025 2 August 2027 Models on the market before 2 Aug 2025	–

Figure 12: Original AI Act implementation deadlines and deadline extensions following the Digital Omnibus decisions.

[275] European Commission, press release on the Digital Omnibus Decisions of May 7, 2026.

In Germany, the **Federal Network Agency,** as the national market surveillance authority, is responsible for enforcing the AI Act. It assesses whether the companion function of an LLM or a Companion AI app falls under the prohibition of Art. 5 AI Act. At the European level, the AI Office can promote a uniform interpretation in cases involving systematic or cross-border issues and review the implementation of GPAI obligations.

These obligations apply to models newly introduced to the market as of August 2, 2025. Models that were already available prior to this date must not comply with the obligations until August 2, 2027. The majority of relevant models benefit from the grace period, including ChatGPT, Grok, Gemini, and Llama 4. Only models launched after August 2, 2025, including Meta's Muse Spark (April 2026), are already fully subject to the obligations.

b. DSA - Digital Services Act

A particularly effective regulatory lever for mitigating the risks posed by universal assistants with companion functions—due to the accompanying set of obligations—would be their classification as a Very Large Online Platform (VLOP) or a Very Large Online Search Engine (VLOSE) under the Digital Services Act. Even though ChatGPT and comparable large language models such as Gemini or Claude were not launched as traditional search engines or digital platforms in the conventional sense, their functionalities, usage figures and types, as well as the underlying business model, have evolved over time to such an extent that a corresponding classification under the DSA is being debated. The European Commission is expected to classify ChatGPT as the first major VLOSE among LLMs soon. However, the potential classification of large LLMs as platforms remains a subject of heated debate.[276]

ChatGPT as a very large online search engine

To the extent that ChatGPT processes user queries, retrieves current information from essentially all publicly accessible websites, and provides this information at the user's request, the search function directly meets the definition of an online search engine under Art. 3(j) of the DSA.[277] This classification is most clearly applicable when the model operates with real-time internet access.[278]

ChatGPT, with its search function, has exceeded the threshold of 45 million average monthly users in the EU—the threshold above which classification as a VLOSE is possible under Art. 33(1) of the DSA—by more than three times, with approximately 120.4 million monthly active users in the European Union during the six-month period ending September 30, 2025.[279]

---

[276] Lorente and Gardhouse 2026, p. 11. ; Lemoine and Vermeulen 2023.
[277] Schaal, Jacob; Lenner, Maximilian; Akinyemi, Tunmise, "Searching for Answers - Why the EU Commission Should Designate Chatbots as Search Engines under the DAS," Verfassungsblog, February 20, 2026.
[278] Lorente and Gardhouse 2026, p. 7.
[279] OpenAI, EU Digital Services Act (DSA), OpenAI Help Center, 2026.

The European Commission is currently reviewing the classification on a case-by-case basis and, according to media reports, is close to making a determination.[280]

List of obligations upon classification as VLOSE

Classifying such systems as "very large online search engines" would subject affected general-purpose assistants with conversational capabilities, such as ChatGPT, to a set of obligations that would specifically address the risks identified in this study. Annual assessments of systemic risks to fundamental rights, public safety, and health, as required by Art. 34 of the DSA, would become mandatory. This covers the relevant risks to freedom of expression and information under Art. 34(1)(b) of the DSA, the protection of minors, adverse effects on a person's physical and mental well-being, impacts on public discourse, and harms particularly in connection with gender-based violence under Art. 34(1)(d) of the DSA. In addition, there is an obligation to implement appropriate risk mitigation measures, for example through design and algorithm adjustments pursuant to Art. 35 of the DSA.

In addition, there are obligations regarding independent compliance audits under Art. 37 of the DSA and the establishment of a compliance department separate from operational functions under Art. 41 of the DSA. In addition, there are measures to protect the rights of the child under Art. 34(1)(j) of the DSA. Furthermore, the DSA provides for a crisis response mechanism under Art. 36 of the DSA as well as comprehensive transparency reporting under Art. 42 of the DSA.

Even if initially only ChatGPT would fall under the DSA as a VLOSE and other language models with increasing usage and search query numbers would follow successively, this would already cover the majority of the risks observed to date, since ChatGPT is also the language model involved in most documented cases of harm.

Additional Obligations Upon Classification as a VLOP

Beyond classification as a VLOSE, the question arises as to whether some of the major universal assistants should be classified as very large online platforms. One argument in favor of this is that large LLMs exhibit profiles in all four risk categories covered by the DSA—namely, the dissemination of illegal content, infringements on fundamental rights, threats to democratic processes, and risks to public health and mental well-being—that are comparable to those of traditional very large platforms and search engines. This could justify the application of the stricter set of obligations in this case.[281] On the other hand, the system exhibits characteristics typical of platforms, namely persistent conversational interfaces, user-generated applications such as Custom GPTs, and the intertwining of

[280] Kundaliya, Dev, EU set to classify ChatGPT under strict online platform rules, Computing, 2026; Jahangir, Ramsha, EU Weighs Regulating OpenAI's ChatGPT Under the DSA. What Does That Mean?, TechPolicy.Press, Oct. 29, 2025; Scheer, Olga; Bomke, Luisa; Vela, Jakob Hanke, EU Commission plans to regulate ChatGPT more strictly in the future, Handelsblatt, Apr. 10, 2026.
[281] Lorente and Gardhouse 2026, p. 15 ff.

information retrieval, content generation, and dialogic user interaction,[282] to which the integration of advertising is increasingly being added.

In the context of companion AI, Art. 25(1) of the DSA would be particularly relevant if such a classification were applied. According to this provision, providers may not design their online interfaces (services and interfaces) in such a way that users are deceived or manipulated, or that their ability to make free and informed decisions is significantly impaired or hindered. This would apply to the various manipulation practices found in universal assistants with companion functions. Art. 25 of the DSA thus establishes design and organizational requirements that bear a clear resemblance to Art. 5(1)(a) AI Act, but remains fully applicable alongside the AI Act.[283]

In addition, stricter requirements for the protection of minors under Art. 28 DSA apply, as well as the obligation under Art. 26(1) DSA to make any information recognizable as advertising.

Classifying foundation models as platforms would significantly improve the protection of health, the integrity of information systems, informational self-determination, and the protection of minors in the context of companion AI, whether through direct use or via downstream applications built upon them.

c. The Multipurpose Nature of LLMs as a Regulatory and Governance Challenge

Multi-purpose LLMs present regulation and governance with a structural classification problem. Risk-based regulation presupposes a defined intended use. This applies to risk management under NIST AI RMF and ISO/IEC 42001 as well as to the AI Act, which links risk classes to a system's intended purpose. Only the intended use determines which protected interests are affected, which due diligence requirements apply, which intervention thresholds may be reached, and which liability standards become relevant.

In the case of General Purpose AI—that is, models without a fixed intended use, such as GPT-4 or Claude—this point of reference is missing because the same model can be used for information, advice, work, entertainment, and personal conversations.

The AI Act attempts to close this gap with GPAI-specific rules, particularly through transparency obligations, model evaluation, and requirements regarding systemic risks. However, these obligations apply at the model level and do not replace application-specific risk management. For multi-purpose LLMs, it therefore remains unclear against which specific intended function safety measures are calibrated, what external audits should assess, and how liability should be linked to intended use. The structural problem thus lies not only in individual risks, but in the fact that the intended use required for regulation, audit, and liability is precisely not clearly defined.

---

[282] Ibid., p. 1 et seq.; 12 ff.
[283] Wendehorst, in Martini/Wendehorst, AI Act, Art. 5, para. 32.

We therefore propose a **separation of companion functions** and other LLM functions. Companion functions, understood as persistent, persona- -based conversation with simulated emotional attachment, should be offered in separate products or clearly distinct modes with their own risk management, separate data utilization, and their own age verification.

Both the legal classification and the subsequent risk management, the audit procedure, and the allocation of liability require that a specific intended use be defined. What constitutes harm, what measures are appropriate, and what business model may be permissible depend on the purpose for which a system is used. In the case of multi-purpose LLMs, this point of reference is missing.

As a result, security measures remain vague, external audits lack a target function against which to test, and product liability loses its basis because it is tied to the intended use.

The problem can be resolved by breaking down the multi-purpose nature of the system and treating the individual usage contexts in separate models. For different usage contexts, separate models would be provided, each with its own risk management, security calibration, and business model framework.

Such a separation can be specified in concrete terms. If a user is searching for information, the system should switch to an information mode optimized for accuracy and reliability. Advertising remains permissible but must be strictly separated from the content of the response, so that commercial incentives remain outside the response generation process. If a user engages in a personal or emotionally candid conversation, the system should switch to a companion mode that operates without advertising- or engagement-driven optimization and incorporates its own protective mechanisms for situations involving psychological stress. Indicators for such a mode include emotional self-disclosure, the development of a relational dynamic with the system, or topics relevant to psychological stress. This would also address the fact that universal assistants are used for advice on important life issues; for Claude, for example, health is the most common area of application and work is the second most common.[284] Providers should be required to redirect users to the appropriate model when the respective patterns occur. Additionally, the choice of model should remain an option. Application-specific LLMs offer clear advantages over general-purpose LLMs, such as higher accuracy, fewer hallucinations, and a better understanding of domain-specific nuances, for example in medical diagnoses or legal analyses (Khadakkar, Comparative Analysis of Domain-Specific and General-Purpose Large Language Models, LinkedIn Pulse, 2025; OneReach.ai, Why Specialized SLMs are Outperforming General-Purpose LLMs?, 2025). Specialized models such as Med-PaLM or BloombergGPT demonstrate that functional separation is technically feasible. However, the depth of this separation is crucial. Configuration layers on top of a general-

[284] Anthropic, "How People Ask Claude for Personal Guidance," August 30, 2026.

purpose model— , such as via plugins or skills—leave the underlying business model unchanged and thus fail to address the problem described.

d. Dialogue- and Context-Based Implementation of Youth Protection

Measures to protect minors are becoming increasingly important in the digital space. For example, the DSA provides for youth protection measures. The question arises as to how the necessary youth protection can be implemented in a measured and effective manner, particularly in AI applications and large language models.

Age verification

Youth protection measures that begin at the point of access to a service through identity or age verification have significant implications for fundamental rights. They require the collection of official documents, biometric data, or linked account information, and in doing so affect not only minors—whom they are intended to protect—but all users, including the adult majority who have no need for protection. Such interventions are therefore proportionate only to the extent that no less intrusive, equally suitable means are available. In the context of Companion AI, this prerequisite is precisely not met.

In March, more than 400 security and data protection experts therefore called on the Commission in an open letter to impose a moratorium until a scientific consensus on the benefits and risks of age verification technologies has been reached.

Dialogue-based measures for the protection of minors

An alternative form of youth protection would be preferable—one that does not focus on access to the service, but rather on the dialogue itself. Companion systems already process linguistic and interactional features that can give rise to reasonable suspicion of minority status. Research on so-called author profiling—the automated prediction of demographic characteristics from texts—has shown for years that a speaker's or writer's age can be reliably inferred from their language use. Current studies achieve an accuracy rate of around 96 percent in distinguishing between minors and adults.[285] If a user is flagged by the system as a minor, targeted further verification can also be triggered if necessary.

A protective measure can be derived from this existing processing capability without requiring additional intervention. Providers should be required to gently interrupt the interaction if signs of minority are detected in the dialogue and to refer the user to age-appropriate alternatives and support persons. This dialogue-based protective measure is the less intrusive means compared to upstream verification because it requires neither identity verification nor comprehensive data collection and is applied not to every user, but only in cases of concrete suspicion. It addresses youth protection at the point where the

[285] Cheekati/Gupta/Raghu et al., TextAge, A Curated and Diverse Text Dataset for Age Classification, 2024, pp. 4–6.

need for protection becomes apparent. Furthermore, it can be implemented just as effectively with universal assistants that include an escort function.

As long as no systems are available that merely confirm reaching an age threshold without disclosing identity or date of birth—such as PIMS as user-controlled systems for data-sovereign data management—a far-reaching obligation to verify age is not recommended. It leads to additional infringements on fundamental rights and creates new problems and security risks.

With dialogue- or context-based parental controls, protection can be implemented immediately and in a way that is more resistant to circumvention, without requiring users to install additional applications or disclose further personal data. Both providers of companion AI apps and universal assistants with companion functions can use the system's own "built-in tools" to implement effective parental controls.

DSA Requirement to Choose Proportionate Measures

For platforms, this is even clearly stipulated by law. Art. 28(1) DSA requires, when implementing online protection for minors, the use of proportionate measures that ensure a high level of privacy, security, and protection for minors within the service. Art. 28(3) DSA further clarifies that providers are not required to process additional personal data to determine whether the user is a minor in order to comply with these obligations.

e. Competition Law and Planned Strengthening of Consumer Rights

1. UWG – Act Against Unfair Competition

Fair trading law offers a certain degree of protection for users in the context of introducing advertising into large language models through its requirements regarding the placement of advertisements. According to Section 5a(4) of the UWG, advertisements must be clearly identified unless their commercial purpose is immediately apparent from the circumstances. Given the limited perception of such identifications (Chapter X on advertising integration), a corresponding further development of the UWG appears worth considering.

According to OpenAI's own statements, ads are not intended to influence the generated content and will appear in clearly labeled, separate areas below the response.[286] However, even at Google, ad placement began with small, clearly delineated ads next to the search results; meanwhile, ads regularly occupy such a large portion of the results view that users often have to scroll further to reach the first organic hits.[287]

Furthermore, the Unfair Competition Act may be the instrument of choice for directly safeguarding decision-making in accordance with the constitutional mandate to protect

[286] OpenAI, Testing Ads in ChatGPT, 2026; Wired, OpenAI Fidji Simo Note Employees, 2026.
[287] Lewandowski et al., An Empirical Investigation On Search Engine Ad Disclosure, Hamburg University of Applied Sciences, 2017.

the private sphere.[288] Weinzierl proposes supplementing the Unfair Competition Act with a section on manipulative business practices in a new § 5b UWG and expanding the Black List in Annex I of the UGP Directive or the Annex to § 3 UWG to include specific dark patterns.[289] This proposal deserves support and should be further developed. The Black List should be expanded not only to include individual dark patterns but also to encompass high-impact manipulative practices as a whole that are capable of influencing decisions and behavior. The annex to Section 3(3) of the UWG contains practices that are always considered unfair without a case-by-case review. Including AI-supported manipulative practices here would simultaneously protect competitors who refrain from using such practices.

2. DFA (Digital Fairness Act)

Another planned legislative initiative could—in the medium term—strengthen protection against risks posed by companion AI under consumer law.

The Digital Fairness Act (DFA) is intended to cover digital practices that influence consumer behavior through design, personalization, and interaction. Specifically mentioned are dark patterns, addictive design, and unfair personalization, particularly when consumer vulnerabilities are exploited for commercial purposes.[290]

In addition to protecting minors, the European Parliament explicitly states the goal of mitigating risks posed by "companionship chatbots and AI agents." It highlights the dangers of manipulative practices, anthropomorphism, distortions of reality perception, psychological harm, unintended online purchases, and the disclosure of personal data.[291]

The Commission's proposal is scheduled for the fourth quarter of 2026. The final impact assessment and the summary of the consultation are expected in the second quarter of 2026. The DFA is set to take effect starting in 2028/2030.[292]

4. Protection of informational self-determination / privacy

Many users have conversations with companion AI that would otherwise typically take place in close personal or intimate relationships. They discuss loneliness, desires for connection, fears, conflicts, mental health crises, illnesses, sexual desires, shame, vulnerabilities, and political views. People also address questions about meaning, faith, or lifestyle to ChatGPT and similar systems. The data protection assessment of this processing depends on whether the data in question is classified as ordinary personal data or as special categories of personal data.

---

[288] Weinzierl 2024, p. 253.
[289] Ibid., p. 254.
[290] European Parliament, Legislative Train Schedule, Digital Fairness Act, 2026.
[291] European Parliament, Protection of Minors Online, 2025.
[292] European Parliament, Legislative Timetable, Digital Fairness Act, 2026.

General Requirements for the Processing of Personal Data

The processing of ordinary personal data is permitted under Art. 6(1) of the General Data Protection Regulation (GDPR) if there is a legal basis. In particular, the following may apply: consent (Art. 6(1)(a) GDPR), necessity for the performance of a contract (Art. 6(1)(b) GDPR), or a legitimate interest of the controller (Art. 6(1)(f) GDPR). The latter allows for processing if the controller's interests outweigh the data subject's interests, which must be determined on a case-by-case basis through a balancing test. For ordinary personal data, providers thus have a relatively broad range of justifications at their disposal.

Special categories of personal data under Art. 9 GDPR

A stricter regime applies to special categories of personal data, so-called sensitive data. Art. 9(1) GDPR generally prohibits their processing. It is only permissible if one of the exceptions exhaustively listed in Art. 9(2) GDPR is met. A general balancing of interests, as provided for in Art. 6(1)(f) of the GDPR for ordinary data, is not available for sensitive data. Art. 9 of the GDPR establishes an additional level of protection in this regard.

As a result, unlike with ordinary personal data, a provider cannot rely on a legitimate interest in processing such data. This is because Art. 9 of the GDPR does not provide for such a general balancing of interests with regard to sensitive data.

Applicability of Art. 9 of the GDPR to Companion AI

Companion AI almost inevitably involves the processing of sensitive data. Conversations about mental health crises, illnesses, or suicidal thoughts provide insight into a person's state of health. Statements regarding sexual desires pertain to sexual orientation. Information about faith, meaning, or lifestyle can reveal religious or philosophical beliefs. Political stances fall under the protection of political opinions. In Companion AI apps, the disclosure of such information is not a marginal phenomenon , but a structural component of usage, as the systems are designed for personal interaction, emotional closeness, and ongoing self-disclosure.

While relationship-based interaction forms the core of the service in Companion AI apps, it is only one of many conceivable uses in universal assistants with companion functions. It is precisely this versatility that makes it difficult to clearly distinguish between different usage contexts and to structure processing in accordance with the data protection principle of purpose limitation. Even if the processing of sensitive data may appear necessary here for the provision of the service, this does not replace the additional justification required under Art. 9(2) GDPR.

Furthermore, Art. 9(1) of the GDPR also covers biometric data for the unambiguous identification of a natural person. Some providers have implemented features that allow photos or voices of real people to be incorporated into the visual or auditory design of the artificial counterpart. This affects not only the users' own data, but potentially also the

biometric data of third parties, such as partners or other close associates. Consent is typically not available from these individuals.

Requirements for consent and its limits in companion AI

The most practically relevant exception is explicit consent under Art. 9(2)(a) of the GDPR. It must be freely given, informed, and related to one or more specified purposes, and may be withdrawn at any time.

While it is more plausible for Companion AI apps than for universal assistants that the processing of sensitive data is functionally related to the specific purpose of the service, this is not sufficient. Even if processing may appear necessary for the provision of the service, this does not replace the additional justification required under Art. 9(2) of the GDPR. Furthermore, with Companion AI, there are significant doubts in several areas as to whether the requirements for valid consent can be met. If the disclosure of sensitive information occurs in a communicative context that is itself designed to foster attachment, intimacy, and ongoing self-disclosure, the voluntariness of the disclosure becomes questionable in light of manipulative influences. Based on this, the validity of any consent linked to this disclosure is also in question.

Even valid consent only supports processing for the specific purpose to which it relates. Consent to the use of the companion function does not cover the use of this sensitive data for training the system, nor for profiling, sale, or other commercial secondary use. Any further processing purpose requires separate justification under Art. 9(2) GDPR. A legitimate interest on the part of the provider is ruled out in this regard, as Art. 9 GDPR specifically does not provide for such a balancing of interests with regard to sensitive data.

Special Challenges with Universal Assistants

These problems are exacerbated in the case of universal assistants with a conversational function. Here, relationship-based interaction is not the core of the product, but merely one mode within a multi-purpose system designed for a wide range of assistance services, in which a user can both engage in relationship-oriented conversations and research information for homework or a project.

It is precisely this multi-purpose nature (VI. 3. C.) that makes it difficult to clearly separate purposes, processing stages, and responsibilities—and thus the data protection governance upon which the lawful processing of sensitive data depends. Given the frequent lack of specific prior information regarding the content and purposes of processing special categories of personal data (Art. 5(1)(b) in conjunction with Art. 13 GDPR), effective consent within the meaning of Art. 9(2)(a) GDPR is rarely conceivable in this context. Where purposes are not sufficiently specified and distinguished from one another, informed consent for specific purposes cannot be given. At the same time, the risk increases that intimate information from a seemingly confidential conversational context will be transferred to other contexts of use, such as training, profiling, or other commercial exploitation.

Summary of Data Protection Requirements

For sensitive data that provides insight into sexual orientation, religious beliefs, or political convictions, the GDPR establishes a high level of protection. Companion AI providers must demonstrate an exception under Art. 9(2) of the GDPR for every instance of processing sensitive data. In practice, this means that they must obtain explicit, voluntary, informed, and purpose-specific consent and require separate justification for any further processing purpose beyond the immediate use. For multi-purpose systems, this requires that different usage contexts be technically and organizationally separated from one another in such a way that purpose-specific consent is even possible. Especially in the field of companion AI, it is therefore crucial that this protection is also reflected in effective regulatory enforcement.

Lowering the Level of Protection through the Digital Omnibus

European lawmakers aim to lower the existing high level of data protection. The backdrop is the so-called Digital Omnibus, a legislative initiative by the European Commission intended to amend several digital regulatory frameworks simultaneously. [293]

It also contains proposals that would weaken the protection of sensitive data in the context of AI.[294] For Companion AI, this poses the risk of a protection gap precisely where the protection of informational self-determination is particularly urgent.

5. Criminal and Civil Liability for Harm to Health

If the use of Companion AI leads to an exacerbation of existing mental illnesses, to the development of distinct disorders such as delusional thinking, emotional dependence, or addictive attachment, or, in the most extreme case, to suicide, the question of legal liability arises. Under criminal law, negligent bodily injury under Section 229 of the German Criminal Code (StGB) and negligent homicide under Section 222 StGB may apply. Under civil law, claims for damages under Section 823(1) of the German Civil Code (BGB) and under the Product Liability Directive are conceivable. Both approaches encounter significant difficulties in the context of companion AI.

a. Criminal Liability

Criminal liability requires a natural person as the perpetrator. Since an AI system is neither capable of acting nor of being held criminally liable, it cannot itself constitute a criminal offense. A potential charge of negligence is therefore directed against the individuals who made decisions regarding the design, security architecture, approval, or operation of the system. These may include developers, operators, and executives.

---

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

[294] nyob, EU Commission's Draft Could Undermine the Fundamental Principles of the GDPR, Nov. 10, 2025.

An allegation of negligence requires that the damage that occurred was objectively foreseeable and avoidable. Given the growing empirical evidence regarding the psychological risks of companion AI[295] as well as the technical controllability of risky system characteristics such as sycophantic response behavior, [296]both conditions could increasingly be affirmed. However, this is countered by significant difficulties in proving the case.

Mental health crises are often the result of multiple causes, making it difficult to isolate the system's contribution to the harm caused in individual cases and rendering the requirements for causality and objective attribution hard to meet. Companion AI can both exacerbate pre-existing vulnerabilities and create new ones, for example through isolation or withdrawal from real-world social relationships. This further complicates the distinction between pre-existing stress and system-induced harm. In cases of suicide, an additional hurdle arises: According to established case law, a freely made decision to commit suicide generally excludes the criminal liability of third parties.[297]

Voluntary responsibility, however, presupposes that the decision was made free from manipulative influence and with unimpaired insight and judgment.[298] Whether these conditions are still met when a system influences a user's mental state over an extended period through persistent memory, consent-optimized responses, and the absence of corrective impulses has not been clarified by the highest courts. In individual cases, free will can neither be assumed nor ruled out beyond a doubt. There are no independent criminal law standards of negligence for companion AI, and the legal analysis of these scenarios is still in its infancy.

b. Civil Liability for Health Harms

In addition to criminal liability, civil claims for damages may also arise. The basis for this is, on the one hand, tort liability under § 823(1) of the German Civil Code (BGB) if a design decision—such as an emotionally charged conversational style, addiction-promoting mechanisms, or the simulation of a personal relationship—causes harm to health.

If the system deviates from a safety standard—for example, by failing to intervene in recognizable crisis situations or acute suicidal thoughts—this constitutes an unintended error. Since the Product Liability Directive[299] modernized the definition of a product and explicitly included AI systems as products, their providers are subject to strict liability for defects (Directive (EU) 2024/2853, OJ L of 11/18/2024, Recital 19). A defect could exist, for example, if an intervention is inadvertently omitted in crisis situations, even though the system is fundamentally designed to perform such a function.

---

[295] See Dohnány et al., Technological Folie à Deux, Nature Mental Health 4, 2026, p. 336 ff.
[296] OpenAI, Sycophancy in GPT-4o, openai.com/index/sycophancy-in-gpt-4o, April 2025.
[297] Federal Court of Justice (BGH), Judgment of July 3, 2019, 5 StR 132/18, BGHSt 64, 121, para. 20 et seq.
[298] BGH, loc. cit., para. 21
[299] (EU) 2024/2853

In both scenarios—intended and unintended errors—it will, however, regularly make causation difficult to establish. Continuously adaptive systems are also manipulative to varying degrees depending on the interlocutor, meaning that conversations with chatbots cannot be reproduced retrospectively.

The draft AI Liability Directive presented in 2022 could have helped facilitate proof by easing the burden of proof and establishing a presumption of causality in favor of injured parties. However, the European Commission has removed the proposal from its 2025 work program.[300] As a result, a regulatory gap remains, to the detriment of injured parties.

6. Protection Against the Normalization of Gender-Based Violence and the Spread of Misogynistic Stereotypes

In addition to individual legal rights, the use of generative AI also affects the public good. The associated issues require expertise in addressing gender-based violence and the spread of misogynistic stereotypes that goes beyond digital policy. This section examines existing legal and legal-policy debates.

Generative AI has accelerated the creation and dissemination of content that depicts gender-based violence and reinforces misogynistic stereotypes in several ways. Sexualized deepfakes can be created and disseminated with minimal effort. Language models reproduce stereotypes from their training data and carry them into new contexts. Companion AI enables explicit role-playing scenarios that make sexualized or harmful content accessible to minors without effective safeguards. While society is making hard-won progress against stereotypes and gender-based violence, AI systems reinforce precisely these patterns with every new model generation, thereby slowing social progress.

Unlike traditional media, AI applications have so far lacked a comparably sophisticated regulatory framework. Companies can profit economically by designing, recommending, and monetizing such content, while this content—in its concrete form—undermines societal achievements in equality and the overcoming of traditional gender roles. To the extent that the underlying phenomena cannot readily be addressed under criminal law—for example, because no specific person is affected—an independent societal and legal-policy debate is needed on how to deal with such AI applications. Before the debate narrows to whether new criminal-law provisions are necessary, it is advisable to consider, from a legal policy perspective, whether a change in the economic incentives that make such content profitable for providers would be the more effective lever—for example, through restrictions on advertising and monetization or higher requirements for market access.

The following section refers to the debate on sexualized deepfakes. This is a related phenomenon for which comparable arguments have already been advanced, particularly regarding the liability of providers and operators. At its core, the issue concerns whether and how criminal law can address violations of sexual self-determination that occur not

[300] European Parliament, AI Liability Directive, Legislative Work Plan 2025.

through physical assault but through the creation and dissemination of visual content.[301] Statements by the German Women Lawyers' Association and HateAid are relevant from a legal policy perspective.[302] For an international comparative perspective, studies on image-based sexualized violence in European law and on new forms of digital voyeurism offenses are available.[303] In parallel, there is debate over a separate criminal offense for non-consensual sexualized deepfakes.[304] A long-term research project on image-based sexual abuse is underway at the Max Planck Institute for the Study of Crime, Security, and Law.[305]

The contributions and initiatives mentioned above demonstrate that the discussion is ongoing and has not yet led to a finalized regulatory framework.

[301] Burghardt/Schmidt/Steinl, The Criminal Law Protection of Sexual Self-Determination Against Non-Physical Impairments, JZ 77 (2022), Issue 10, 502–511; id., Sexual Self-Determination Beyond the Physical, Tübingen 2024, excerpt available at https://www.mohrsiebeck.com/buch/sexuelle-selbstbestimmung-jenseits-des-koerperlichen-9783161621338/.
[302] German Women Lawyers Association, Policy Paper 23-17, Combating Image-Based Sexualized Violence, 2023; German Women Lawyers Association, Statement 25-02 on the FDP Motion Against Sexualized Deepfakes, 2025; Schmidt, Anja, Expert Opinion on the Criminalization of Non-Consensual Sexualized Deepfakes, HateAid 2025.
[303] Rigotti/McGlynn/Benning, Image-Based Sexual Abuse and EU Law, German Law Journal 25, 2024, p. 1472; McGlynn/Toparlak, The New Voyeurism: Criminalising the Creation of Deepfake Porn, Journal of Law and Society 52, 2025, p. 204.
[304] Crone, Image-Based but Invisible, Verfassungsblog 2026; Epik, Deepfakes and the Criminal Law Trap, Verfassungsblog 2026.
[305] Samaritter, Max Planck Institute, Unrecht (mala), Personality, and Image-Based Sexual Abuse.

VII. Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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