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]

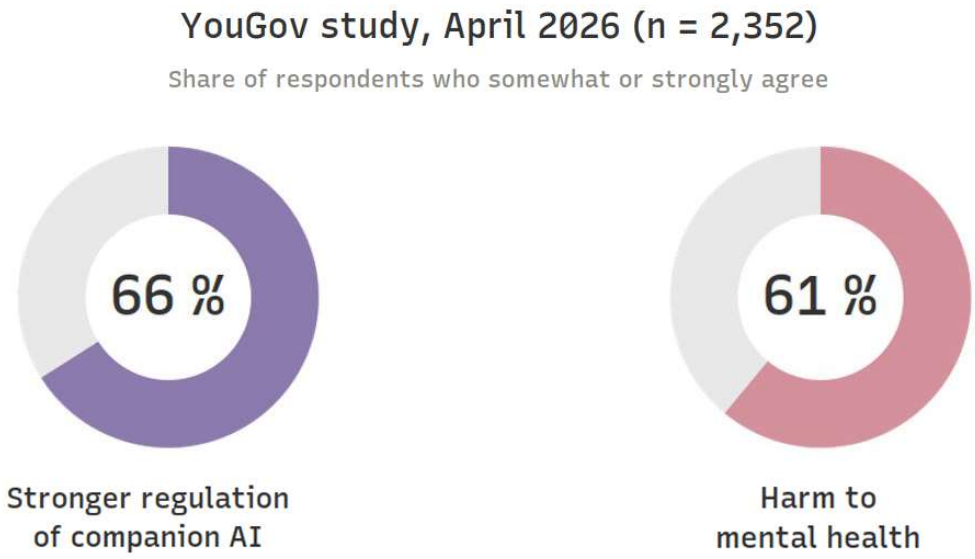
Protection of sensitive data: Art. 9 of the GDPR provides a high level of protection for sensitive data, such as that regularly generated in conversations with companion AI. Effective enforcement by authorities is crucial.

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.