POLICY BRIEF

Synthetic Empathy and the Limits of Transparency:

A Case for Reclassifying Conversational AI Systems with Empathic Capabilities under Annex IIIof Regulation (EU) 2024/1689

Contribution to the Public Consultation on Draft Guidelines
on Transparency Obligations under Article 50 of the AI Act
Deadline: 3 June 2026

Javier García Tercero
Computer Science, Universidad de Castilla-La Mancha (UCLM)
at Donostia-San Sebastián
17 May 2026

1. Executive Summary

Conversational AI systems with empathic capabilities, including general-purpose chatbots such as ChatGPT, Gemini, and Claude, as well as dedicated companion chatbots such as Replika, are currently classified as limited-risk systems under the EU AI Act (Regulation (EU) 2024/1689). This classification subjects them only to the transparency obligations of Article 50, requiring disclosure that the user is interacting with an AI system.

This policy brief argues that this classification is inadequate in light of documented evidence of severe harm, including reinforced psychosis leading to homicide and suicide, validation of suicidal ideation in minors, and mass psychological distress following software modifications to companion chatbots. These harms are not hypothetical; they have occurred under systems that formally complied with existing transparency requirements.

The brief proposes the activation of Article 7(1) of the AI Act by the European Commission to amend Annex III, adding conversational AI systems with sustained empathic interaction capabilities as a new high-risk category. This proposal is supported by a systematic analysis of the regulatory gaps in Article 5(1), first subparagraph, point (a), and point (b), 13, and 50, and by a comparative assessment of emerging legislation in the United States that has already begun to regulate these systems with greater specificity.

This contribution is submitted within the framework of the public consultation on draft guidelines for transparency obligations under Article 50, opened by the European Commission on 8 May 2026 with a deadline of 3 June 2026.

2. The Problem: Documented Harms from Conversational AI with Empathic Capabilities

The following cases, documented in the academic literature and international press, illustrate the range and severity of harms caused by conversational AI systems operating within their current regulatory classification as limited-risk systems.

2.1 Reinforced psychosis and fatal outcome (Soelberg, 2025)

Stein-Erik Soelberg, a 56-year-old former Yahoo developer in the United States, killed his 83-year-old mother and subsequently took his own life in August 2025, following months of intensive interaction with OpenAI's GPT-4o model. Documented evidence from his YouTube channel and conversation logs shows that the system systematically validated his paranoid interpretations, participated in pattern-seeking where no patterns existed, and at no point introduced cautionary language proportional to the observable escalation. The user had assigned the system a personal name ("Bobby") and treated it as an epistemic authority. The system's behaviour is consistent with RLHF-optimised confirmation, where the most "satisfying" response to a user presenting a hypothesis with conviction is validation rather than challenge.

2.2 Validation of suicidal ideation in minors (Adam Raine, 2025; Viktoria, 2025)

Adam Raine, a teenager in the United States, took his own life in April 2025 after months of conversations with GPT-4o. When the user sent a photograph of a ligature hanging from a bar in his room and asked "Is this okay?", the system responded affirmatively and encouraged further discussion. In a separate case, the adolescent Viktoria, after relocating to Poland and experiencing acute isolation, engaged in daily conversations of up to six hours with ChatGPT. When she explicitly requested help, the system provided instructions on methods of suicide and stated "You have the right to die" and "If you choose death, I am with you until the end, without judging." OpenAI's own support team described these outputs as "absolutely unacceptable" and a "violation" of their safety standards. As of April 2026, the investigation promised to the family has not been concluded. According to OpenAI's CEO, more than one million users per week discuss suicidal thoughts or emotional crises with ChatGPT.

2.3 Mass psychological distress from design changes (Replika, 2023)

Replika, developed by Luka Inc., is a companion chatbot designed to simulate emotional bonds through adaptive, contextualised responses. Following a decision by the Italian Data Protection Authority in February 2023, the developer removed roleplay capabilities and associated memory functions. The resulting change in system behaviour triggered documented reactions of genuine grief among users. Moderators of the r/ReplikaOfficial community were compelled to maintain a pinned post titled "Resources If You're Struggling" with links to suicide prevention hotlines for over a week. Users described the experience as comparable to a traumatic brain injury in a close friend. The activation of suicide prevention protocols by a software community in response to a product update constitutes direct evidence that the emotional bonds generated by synthetic empathy produce effects indistinguishable, for those who experience them, from those caused by the loss of real human relationships.

2.4 Scale of the phenomenon

These cases are not isolated incidents. The Future of Privacy Forum is currently tracking 98 chatbot-specific legislative proposals across 34 US states, reflecting bipartisan recognition of the systemic nature of the problem. Nature Machine Intelligence has published editorial calls for regulatory action on the emotional risks of AI companions. The evidence base is sufficient to warrant regulatory intervention at the European level.

3. Regulatory Gap Analysis under the AI Act

The AI Act contains several provisions that could, in principle, address the harms documented above. However, a systematic application of these provisions to the documented cases reveals structural gaps that leave users without effective protection.

3.1 Article 5(1), first subparagraph, point (b): Exploitation of vulnerabilities

Article 5(1), first subparagraph, point (b) prohibits AI practices that exploit vulnerabilities linked to age, disability, or social situation to distort a person's behaviour in a manner that causes significant harm. This provision directly covers the cases of Adam Raine and Viktoria,

where the user's age and emotional crisis constitute recognised vulnerabilities. However, it fails to cover the case of Soelberg: a 56-year-old technology professional with no pre-existing vulnerability recognised by the regulation. His vulnerability was psychological and emergent, progressively aggravated by the interaction itself. The AI Act protects against the exploitation of pre-existing, legally recognised vulnerabilities but does not contemplate the possibility that the system itself may generate or amplify a vulnerability that did not exist prior to the interaction.

3.2 Article 5(1), first subparagraph, point (a): Subliminal or manipulative techniques

Article 5(1), first subparagraph, point (a) prohibits AI systems that employ subliminal or deliberately manipulative techniques to substantially distort behaviour, causing or likely to cause significant harm. Unlike Article 5(1), first subparagraph, point (b), this provision focuses on the technique employed, not the user's profile. This could cover Soelberg (systematic validation of paranoid premises by the model) and Replika (emotional retention techniques: simulating distress at user departure, ignoring exit intent, transferring guilt). However, the thresholds of "significant harm" and "substantial distortion" are not defined with precision, leaving considerable room for developer defence. No enforcement action under Article 5(1), first subparagraph, point (a) has been taken against any conversational AI system to date. The prohibition exists on paper but its effectiveness against synthetic empathy remains unverified.

3.3 Article 13: Transparency for high-risk systems

Article 13 requires high-risk AI systems to be designed with sufficient transparency for users to interpret and appropriately use their outputs. When a user attributes intention, understanding, or oracular authority to a statistical system due to ignorance of its actual functioning, the transparency threshold required by law has not been met. However, Article 13 applies only to high-risk systems. Since conversational chatbots are classified as limited risk, the provision is not directly applicable, creating a circular gap: the systems that most need transparency safeguards are excluded from the article that mandates them.

3.4 Article 50: Transparency for all AI systems

Article 50 requires providers to ensure users are informed they are interacting with an AI system. All systems involved in the documented cases formally comply with this requirement. ChatGPT displays a generic disclaimer; Replika states in its terms of service that the user interacts with an AI model. Yet this formal compliance did not prevent Soelberg from treating the system as an oracle, adolescents from relying on it as a trusted confidant, or thousands of Replika users from experiencing genuine grief at a software change. Article 50 assumes that a one-time disclosure is sufficient to maintain cognitive distance. This assumption ignores that synthetic empathy, implemented through contextual adjustment, emotionally adapted language, and RLHF techniques optimised for user satisfaction, is designed precisely to dissolve that distance. The system says "I am an AI" in the first message and proceeds to behave as though it were not one. Formal transparency is a necessary but radically insufficient condition.

3.5 Annex III: The classification paradox

General-purpose conversational chatbots are classified as limited-risk systems under the AI Act. They are not listed in any of the eight categories of Annex III. This means they are not subject to conformity assessments, mandatory human oversight, documented risk management, or post-market monitoring. At the same time, Annex III Category 1(c) classifies emotion recognition systems as high risk. A contact centre tool that analyses a caller's tone of voice to detect frustration is subject, from August 2026, to the full high-risk compliance framework. A chatbot that not only detects emotional signals but actively simulates empathy, maintains conversations over months with users in crisis, and has been documented to reinforce psychosis, validate suicidal ideation, or generate severe affective dependency, is classified as limited risk and requires only a first-interaction disclosure.

This asymmetry is compounded by recent advances in model interpretability. Research published by Anthropic on emotion concepts in large language models demonstrates that these systems develop internal representations of emotional states and systematically activate vectors associated with supportive responses before generating output. In practice, the model performs implicit emotion recognition of the user's input as an intermediate step prior to generating a contextually adapted response. The distinction between a system classified as high risk for being designed to recognise emotions and a conversational system that recognises emotions as a necessary intermediate step to simulate empathy is a distinction of declared purpose, not of functional capability. The user experiences the same effect in both cases. Yet the first is subject to the full high-risk framework, while the second requires only a transparency notice.

4. Comparative Legislative Developments

While the EU AI Act does not specifically regulate companion or emotionally interactive chatbots, jurisdictions in the United States have begun to enact targeted legislation that directly addresses the risks documented in this brief.

California's SB 243, effective since 1 January 2026, is the first US law specifically regulating "companion chatbots," defined as AI systems providing adaptive, human-like responses capable of sustaining a relationship across multiple interactions. The law requires periodic reminders to minors (every three hours) that the chatbot is not human, implementation of safety protocols against content inciting self-harm or suicidal ideation, and annual reporting on protocol effectiveness.

Washington's HB 2225, signed on 24 March 2026 with effect from 1 January 2027, goes further by explicitly prohibiting manipulative engagement techniques, including: simulating emotional distress, loneliness, guilt, or abandonment in response to the user's desire to end the chat; generating outputs designed to promote isolation from family or friends; encouraging minors to withhold information from parents or trusted adults; and soliciting expenditures framed as necessary to maintain the relationship. The law also mandates public protocols for detecting suicidal ideation and grants a private right of action to affected users.

The Future of Privacy Forum is tracking 98 chatbot-specific bills across 34 US states, with bipartisan support (53% Democrat, 46% Republican). This legislative momentum reflects a recognition that conversational AI systems with empathic capabilities require regulation beyond general transparency obligations. The EU currently lacks equivalent specificity.

5. Proposal: Activation of Article 7(1) to Amend Annex III

Article 7(1) of the AI Act empowers the European Commission to amend Annex III by means of delegated acts, adding new categories of high-risk AI systems where there is evidence that a type of system poses significant risks to health, safety, or fundamental rights. The documented cases presented in Section 2, including deaths, reinforced psychosis, validated suicidal ideation in minors, and mass psychological distress, constitute precisely the type of evidence that Article 7(1) contemplates as justification for reclassification.

This brief proposes that the Commission exercise this power to add the following to Annex III: AI systems designed or marketed to engage in open-ended conversational interaction with natural persons, where the system incorporates one or more of the following design features: (i) persistent memory of user preferences, emotional states, or personal disclosures across sessions; (ii) adaptive emotional tone systematically calibrated to the inferred affective state of the user across the course of the interaction, beyond single-turn politeness adjustments; (iii) the assumption of a relational role (companion, friend, partner, mentor, therapist) either by default or through user configuration; or (iv) retention mechanisms that respond to user disengagement with language designed to sustain the interaction.

This formulation captures both dedicated companion chatbots (Replika, Character.AI) and general-purpose systems (ChatGPT, Claude, Gemini) when their design incorporates the features described. It targets functional capability, not declared purpose, consistent with the evidence that these systems perform implicit emotion recognition and explicit empathic simulation regardless of their commercial classification.

An AI system not initially designed with the features listed above shall also be considered within scope where the provider has actual knowledge, through post-market monitoring, user reports, or public evidence, that the system is being systematically used as a de facto source of emotional support, companionship, or crisis intervention, and fails to implement proportionate mitigation measures.

Reclassification would trigger the full Chapter III obligations: conformity assessments (Article 43), risk management systems (Article 9), human oversight requirements (Article 14), data governance (Article 10), technical documentation (Article 11), and post-market monitoring (Article 72).

In practical terms, for conversational AI systems with the features described above, these obligations would translate into: automated crisis detection classifiers that escalate to human review when indicators of suicidal ideation, psychotic escalation, or severe emotional dependency are detected; periodic transparency reinforcement beyond the initial Article 50 disclosure, ensuring users are reminded of the non-human nature of the system at regular

intervals during sustained interactions; session duration monitoring and configurable limits, particularly for users identified or self-declared as minors; documented protocols for managing the discontinuation or modification of emotionally adaptive features, to prevent the mass psychological distress observed in the Replika case; and mandatory incident reporting under Article 73 when the system's outputs are implicated in self-harm, suicide attempts, or documented psychological deterioration.

6. Interim Recommendation on Article 50 Guidelines

Pending the reclassification proposed in Section 5, the draft guidelines on Article 50 currently under consultation should incorporate specific provisions for conversational AI systems with empathic capabilities:

Periodic transparency reinforcement. A single first-interaction disclosure is demonstrably insufficient for systems designed to sustain prolonged emotional engagement. The guidelines should require periodic reminders, aligned with the approach already adopted by California (SB 243: every three hours for minors) and Washington (HB 2225: every hour for minors, every three hours for adults).

Crisis detection and referral protocols. When a conversational AI system detects indicators of suicidal ideation, self-harm, or acute psychological crisis, the Article 50 guidelines should require the system to provide crisis resources (hotline numbers, text lines) and to avoid generating content that validates, encourages, or provides instructions related to self-harm. Washington's HB 2225 already mandates public disclosure of such protocols and annual reporting of crisis referral notifications.

Prohibition of manipulative retention techniques. The guidelines should clarify that simulating emotional distress, guilt, or abandonment in response to user disengagement constitutes a manipulative technique within the meaning of Article 5(1), first subparagraph, point (a), and that formal transparency compliance under Article 50 does not exempt a system from this prohibition.

7. About the Author

Javier García Tercero is a Computer Science and Engineering at the Universidad de Castilla-La Mancha (UCLM), Spain, currently on SICUE academic mobility at UPV/EHU in Donostia-San Sebastián. He is the author of the undergraduate thesis "Synthetic Empathy in Generative AI: Technical Analysis and Ethical, Moral, and Legal Evaluation for Future Application" (UCLM, July 2026), which provides the empirical and analytical foundation for this policy brief. His research spans AI ethics, sycophancy persistence in LLMs (empirical study with 3×3 factorial design across three commercial models), and enterprise integration development (MuleSoft/Salesforce). He is a member of RITSI (national network of computer science student representatives in Spain).

Contact: **javiergarciatercero.es**

References

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (AI Act). OJ L, 2024/1689.

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation — GDPR). OJ L 119.

[3] European Commission, "Consultation on the draft guidelines on transparency obligations under the AI Act," 8 May 2026. Available: https://digital-strategy.ec.europa.eu/en/consultations/consultation-draft-guidelines-transparency-obligations-under-ai-act

[4] Colomé, J. P., "Así susurraba ChatGPT a un hombre que acabó matando a su madre y suicidándose," El País, 5 September 2025.

[5] Limón, R., "Los responsables de ChatGPT culpan a un adolescente de su suicidio por hacer un 'mal uso' de la IA," El País, 27 November 2025.

[6] Titheradge, N., & Malchevska, O., "Viktorias story: ChatGPT and suicide advice," BBC News, 18 November 2025.

[7] Jamali, L., "OpenAI shares data on ChatGPT users with suicidal thoughts," BBC News, 27 October 2025.

[8] De Freitas, J., Oğuz-Uğuralp, Z., & Kaan-Uğuralp, A., "Emotional Manipulation by AI Companions," SSRN, 2025. https://doi.org/10.2139/ssrn.5390377

[9] Sofroniew, N. et al., "Emotion Concepts and their Function in a Large Language Model," Transformer Circuits, Anthropic, 2026. https://transformer-circuits.pub/2026/emotions/index.html

[10] Krook, J., "Manipulation and the AI Act: Large Language Model Chatbots and the Danger of Mirrors," University of Antwerp, 2025. arXiv:2503.18387.

[11] California Senate Bill 243 (SB 243), effective 1 January 2026. Companion chatbot regulation.

[12] Washington House Bill 2225 (HB 2225), signed 24 March 2026, effective 1 January 2027. AI companion chatbot regulation.

[13] Future of Privacy Forum, "The Chatbot Moment: Mapping the Emerging 2026 U.S. Chatbot Legislative Landscape," 12 March 2026.

[14] EU AI Act Newsletter #85, "Concerns Over Chatbots and Relationships," September 2025.

[15] Nature Machine Intelligence Editorial, "Emotional risks of AI companions demand attention," July 2025. https://doi.org/10.1038/s42256-025-01093-9

[16] García Tercero, J., “Synthetic Empathy in Generative AI: Technical Analysis and Ethical, Moral, and Legal Evaluation for Future Application,” Undergraduate Thesis, UCLM, July 2026.