Ethical challenges of AI companions

Closing the governance gap in human-AI relationships

DATE
01 July 2026

AUTHORS
Horizon Europe Project AIOLIA (Operationalising AI Ethics for Learning and Practice: A Global Approach)
Grant Agreement N° 101187937

Executive summary

With an increasing number of human-AI interactions taking place through general-purpose tools and dedicated companion apps, people are gradually forming relationships with AI systems. Current EU regulation addresses this situation primarily through transparency obligations. This is insufficient to respond to risks that emerge gradually, such as dependency and manipulation. Drawing on AIOLIA findings, this policy brief advances a set of actionable recommendations for addressing the challenges of AI companions in practice. It identifies four key areas of intervention: 1) identifying and protecting vulnerable users, 2) implementing longitudinal oversight mechanisms, 3) introducing friction as a safety-by-design measure, and 4) developing forms of transparency that extend beyond simple disclosure requirements.

What are AI companions?

AI companions are AI systems that engage users in intimate conversations, offering emotional support, entertainment, human-like relationships, or personal coaching. This behaviour may be deliberately designed for or emerge without the designer's intent. It gradually induces a cognitive and behavioural shift in the user that the human does not always consciously register. AI companionship can arise both through dedicated AI products or through the use of general-purpose models.

Key tension

The features that make AI companions valuable, such as constant availability, personalised responsiveness, and affective language, also pose important risks of emotional dependency, manipulation, erosion of privacy, and erosion of meaningful consent. Regulation cannot simply prohibit these features, because they are essential parts of the technology. Instead, layered safeguards must be established and calibrated to individual user vulnerability and relational depth.

Empirical findings

01

Personalisation creates a tension

between usefulness and data protection that makes consent an ongoing challenge, not a one-time event.

02

Frictionless conversations are often harmful to humans

as risks arise from cumulative engagement, not discrete incidents, making frictionless design itself a potential source of harm.

03

Vulnerability is dynamic and contextual

requiring the detection of shifting risks rather than reliance on predefined categories.

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

1. Problem and policy context

Context

Millions of people are turning to AI systems for advice, emotional support, and personal guidance. Through increased and repeated interaction, users may develop relationships with AI characterized by trust, emotional attachment, and reliance. In the process, they may disclose sensitive personal information or their behaviour may become increasingly influenced by the AI system.

These risks are reinforced by design features such as persistent memory, personalisation, embodied avatars, human-like voices, and emotionally adaptive communication, which systematically encourage anthropomorphic projections and can further strengthen the user's emotional engagement with AI. Providers face strong commercial incentives to optimise for engagement, retention, and continued interaction.

This development poses a challenge for existing approaches to AI governance. Most current regulatory frameworks were designed for AI systems understood primarily as tools used for specific purposes, for example content generation, recommendation, or decision support. Ethically speaking, they focus primarily on accuracy, transparency, and disclosure. While these concerns remain important, they may not fully capture the specific cognitive and behavioural effects of AI companions.

Two AIOLIA examples

AIOLIA's findings stem from detailed discussions with industrial companies developing AI companions, who are confronted with their challenges in real-world contexts.

First, AIOLIA studied a personalised AI character platform, allowing users to create and interact with customised AI personas in open-ended private conversations. The product's value lies precisely in the flexibility and depth of personalisation of the synthetic interaction partner.

Second, AIOLIA studied AI-powered habit formation assistants, designed to guide users towards healthier behaviours and habits through sustained interaction. While the first use case was primarily oriented towards entertainment and self-expression and the second towards personal development, both relied on ongoing conversational engagement and extensive personalisation.

Policy gap

Under current European regulation, AI companions are subject to a single obligation: informing users that they are interacting with an AI system. This is important but insufficient. **Disclosure alone cannot address the distinctive risks associated with the use of AI companions**. Users may be fully aware of this, while nevertheless developing emotional attachment, disclosing sensitive personal information, or becoming increasingly influenced by the system over time.

Currently, the AI Act seeks to ensure human health, safety, and fundamental rights through human oversight mechanisms, which are primarily designed for discrete, high-stakes decisions. However, **AI companions create risks at scale**, which emerge through interaction trajectories rather than individual outputs, requiring oversight that is longitudinal rather than episodic.

A related gap concerns the AI Act's prohibition of manipulation, which is limited in application only to purposefully manipulative and deceptive techniques. This is ill-suited to AI companions, where influence accumulates and **manipulation may arise through ordinary dialogue rather than purposefully deceptive strategies**.

As a result, **policymakers face a regulatory challenge that extends beyond the problems of deception and disclosure**. The central issue is how to govern systems capable of shaping users' emotions, perceptions, and behaviour, through long-term relational interactions. Existing governance approaches remain poorly equipped to address harms that vary across users or become visible only over extended periods of use.

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

2. Policy recommendations

1a **Know thy user**

Existing approaches to user protection often assume that vulnerability is a pre-existing characteristic of particular groups. However, new forms of vulnerability may emerge from sustained human-AI interaction: 1) forms of emotional reliance that become emotional dependency, 2) social withdrawal that leads to permanent social isolation, or 3) a reinforcement of cognitive biases. These emergent vulnerabilities develop gradually. They are difficult to detect using conventional safeguards. In addition, applying the same safety standards to all users risks imposing unnecessary restrictions on some, while failing to adequately protect others. Governance mechanisms should focus on indicators of elevated risk that emerge through interaction, rather than proceed from fixed assumptions about users' vulnerability.

1b **Practical measures**

- *Detect emerging vulnerability through long-context analysis of interaction patterns, recognizing that risk is dynamic and can shift over time.*
- *Calibrate safety responses to detected risk levels, including adapted interaction defaults and referral to human support when thresholds are met.*

2a **Longitudinal oversight**

The AI Act's human oversight model assumes a reviewable and attributable decision point — an assumption that does not scale to the continuous, private, and high-volume interaction with AI companions. The same limitation extends to existing evaluation frameworks and technical safety measures: they rely predominantly on single-snapshot benchmarks, while content moderation and behavioural guardrails remain focused on short excerpts from interaction logs. As a result, these tools are poorly equipped to detect harms that emerge gradually through prolonged engagement.

Most infrastructure required for longitudinal oversight already exists as platforms collect and retain information about user interactions and behavioural patterns, yet they use them primarily for optimising engagement and user retention. Governance frameworks should require platforms to apply this same data infrastructure to deploy wellbeing-oriented analysis. Well-being analysis involves identifying indicators of distress or dependency against clearly defined intervention thresholds. This reorientation of existing data practices, rather than introducing new surveillance, supplies an ethical basis of profiling users, based on the “Know Thy User” recommendation. To support accountability and independent scrutiny, AI providers should organize regular data and privacy audits, and also publish aggregated risk-related statistics on AI companions.

2b **Practical measures**

- *LLM-based automated oversight mechanisms capable of operating at scale and across sessions.*
- *Development of longitudinal evaluation protocols to identify risks that only become visible over repeated use, including mandatory well-being analysis subject to independent evaluation with dependency and distress thresholds.*
- *Benchmarking GPAI systems for tendency towards companionship-style chats by measuring sycophancy ratios, intimacy formation patterns, and anthropomorphic projections.*
- *Public disclosure of aggregated user statistics and aggregated data access for research institutions.*

3a **Engineering friction**

Many AI companions are optimised to maximize engagement through continuous availability, personalised responsiveness, and alignment with user preferences, which systematically reduce opportunities for reflection, disagreement, or disengagement. To address a regulation gap around AI-related manipulation, governance frameworks should disincentivize sycophancy and encourage calibrated forms of human-AI friction as a means of promoting human autonomy. Developers have already begun introducing safeguards such as crisis referrals, break prompts, and restrictions on certain anthropomorphic behaviours. However, these remain largely reactive, voluntary, and rarely accompanied by mechanisms to revisit user consent as interactions become more personal.

3b	**Practical measures**	➢ *Introduce structured pauses in prolonged or high-intensity interactions (e.g., reflective prompts, summary of the latest interaction, reference to human contact) triggered by predefined thresholds such as usage patterns or emotional intensity signals.* ➢ *Prohibit the use of engagement metrics that correlate with emotional dependency, such as session length or return frequency, as primary optimisation mechanisms.* ➢ *Require that consent mechanisms be dynamically re-triggered when the monitoring module detects a qualitative shift in relational depth.*
4a	**Transparency beyond disclosure**	Existing transparency requirements are largely based on the assumption that users can protect themselves from harm if they are provided with sufficient information about the system. In the context of AI systems, this assumption is valid only partially. Users may remain fully aware that they are interacting with an AI system while nevertheless developing emotional dependency, disclosing intimate information, or becoming behaviourally manipulated over time. Going beyond disclosure, regulation should strive to provide users with the opportunity to understand how and whether these effects may arise over time. There is a need for an active and contextual transparency toolkit operating alongside the companion system to help users recognize engagement patterns, understand how personal information shapes the interaction, and identify when the relationship builds up in intensity or crosses invisible behavioural borders.
4b	**Practical measures**	➢ *Require systems to move beyond generic role-based disclaimers (e.g., "I am not a doctor") toward contextual explanations of how the system actually generates responses.* ➢ *Request that explanations be given proactively rather than solely on user request.* ➢ *Provide chat history summaries to users at regular intervals, enabling conscious awareness of their own emerging engagement patterns.*

3. Policy implications

Implementation challenges

- ***Industry coordination:*** *effective governance of AI companions requires close coordination between regulators and industry to ensure regulation both aligns with and reflects technical and societal developments.*
- ***Chinese regulation*** *of AI companions is already in place and may set a standard across the world ("The Interim Measures for the Administration of Anthropomorphic Interactive Services of Artificial Intelligence" come into force on July 15, 2026).*
- ***Commercial resistance:*** *in the global market for AI companions, industrial actors whose business models depend on emotional engagement may resist enhanced regulation. Operational frameworks co-created with AI engineers and reasonable enforcement mechanisms are essential in order not to overregulate.*
- ***Evidence gaps:*** *scientific consensus on when and how AI companionship may cause harm is still emerging. Policymakers should invest in independent research on behavioural change in humans, while already taking steps to protect all users against emergent vulnerabilities.*

Conclusion

AI companions are currently marketed and used as emotional support tools and quasi-clinical wellness applications without the safeguards required from medical devices. AIOLIA's findings emphasize non-trivial risks of personalisation or frictionless design.

AIOLIA's evidence shows the emergence of new cognitive and behavioural vulnerabilities in AI companion users over time. Current regulatory treatment of AI companions needs a systemic extension to account for the continuous, relational risks and new vulnerabilities.

This requires defining the obligations of AI providers by function rather than by product category (dedicated companion vs. general purpose), and establishing dynamic safety and evaluation mechanisms, including counterintuitive technical measures, for example by emphasizing LLM oversight over human oversight to prevent harms at scale.

Horizon Europe AIOLIA project (www.aiolia.eu)

Coordinator CEA

Lead Partner THWS

Funded by the European Union under Horizon Europe Grant Agreement N° 101187937. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

This brief draws on

Bayerl, P.S., Lawlor, E., Maris, M.T., Miorandi, D., Pekšys, G., Bjelica, M., Stojšin, K., Anastasova, M., Smith, O., Henestrosa, A., Yamshchikov, I., Bak, M.A.R., & Akhgar, B. (2026). Operational Ethics Guidelines on Use Cases Related to Human Behaviour and Cognition. AIOLIA Deliverable D3.1.

Aires, S., Kyosovska, N., Griffith, J., Teo, S. A., Bogucki, A. (2026). Operational Non-Technical Guidelines for AI Research Areas. AIOLIA Deliverable D3.3.