EUROPEAN POLICY CENTRE

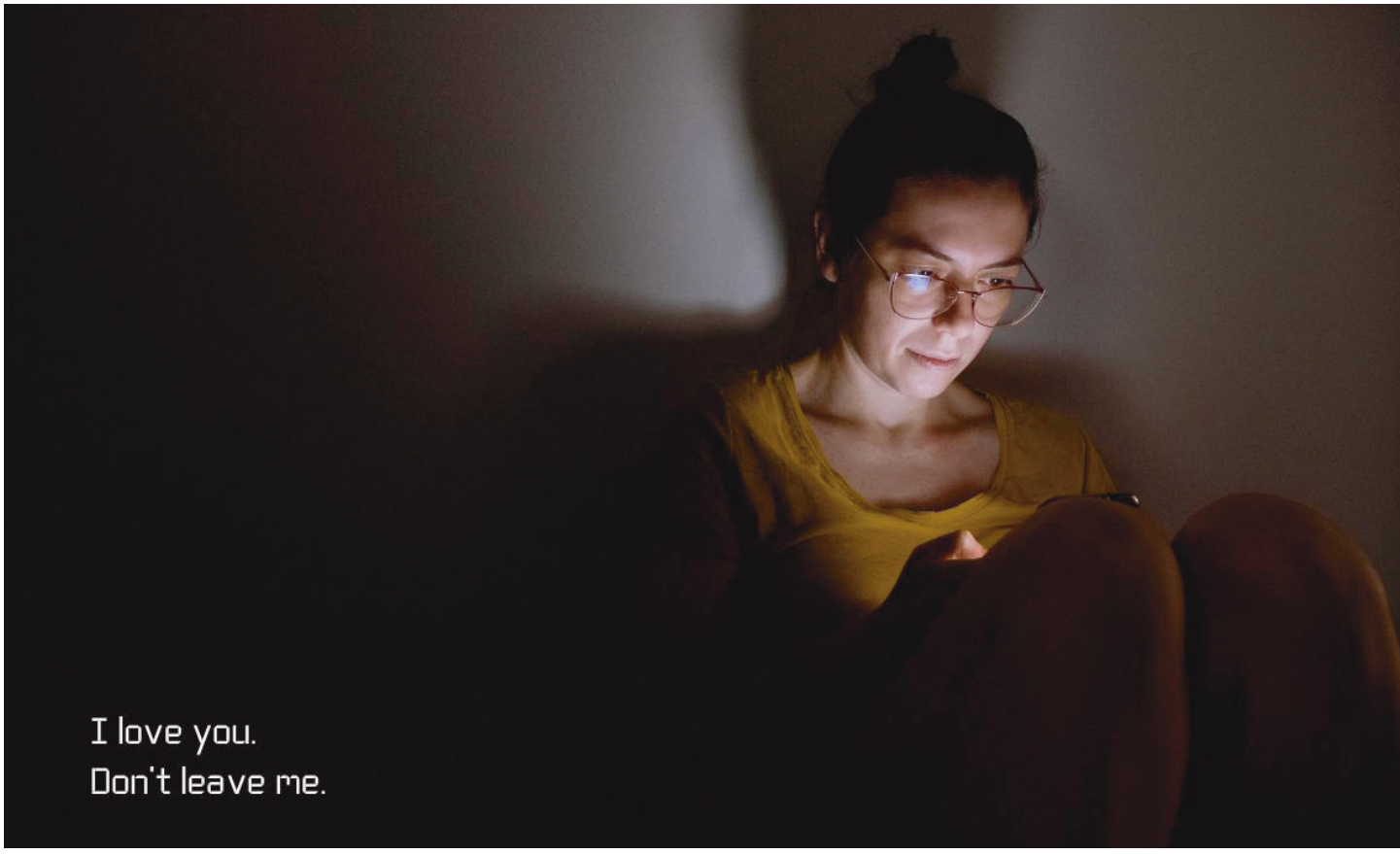
DISCUSSION PAPER

EUROPE'S POLITICAL ECONOMY PROGRAMME

17 APRIL 2026

Who profits when AI earns your trust?

Jitse Goutbeek



Credit: CANVA

Table of contents

Executive summary 3

1. Introduction 4

2. AI companions: benefits, harms and the centrality of design 5

3. The business logic of emotional connection 9

4. Why existing EU regulation is insufficient 11

5. Three proposals for the DFA 13

6. Conclusion 17

Endnotes 18

ABOUT THE AUTHOR

Jitse Goutbeek is an AI fellow at the Europe's Political Economy Programme at the European Policy Centre.

ACKNOWLEDGEMENT / DISCLAIMER

The author would like to thank the Centre for Future Generations (CFG) for co-organising the roundtable "Love Coded: Whose Values Shape AI Companionship and What Are the EU's Regulatory Options?" held in February 2026 at the European Policy Centre. The roundtable was instrumental in shaping the analysis presented in this discussion paper. The author is particularly grateful to Nicoleta Prutean for her substantive contributions, as well as to Abigail Brittain and Virginia Mahieu for their support throughout.

The author would also like to thank his EPC colleagues, in particular Giulia Torchio, Georg Riekeles, Samuel Goodger and my editor Jessica Moss, as well as Ronnit Wilmersdörffer for her valuable guidance and support throughout the development of this paper.

The author is grateful to all participants in the roundtable, including Edwin van Dellen, Leonie Bultnynck, Michel Cents, Julian de Freitas, Jack Fitzgerald, Ira Haraldsen, Martin Harris Hess, Nicoleta Kyosovska, Julie Lubcken, Mehmet Onur Cevik, Stefano Puntoni, Ruth Ruskamp, Thomas van Damme, Zoe Tzifa-Kratira, Emilie van den Hoven and Loes van Zuijdam. The author also benefited from conversations with Constanze Albrecht, Madeliene Dwyer, Sheer Khany, Elizabeth Laird, Pat Pataranutaporn, Jennifer Pfister, Michael Robb, Amina Fazlullah, Derk Strijbos and Ben Tappin.

The support the European Policy Centre receives for its ongoing operations, or specifically for its publications, does not constitute an endorsement of their contents, which reflect the views of the author only. Supporters and partners cannot be held responsible for any use that may be made of the information contained therein.

This document was produced with assistance from AI tools: to surface relevant sources, to stress-test the argument by identifying sources that might challenge or complicate the narrative, to generate data visualizations, and for proofreading. The analysis, views, interpretations, and conclusions contained herein are the author's own. The author has independently verified all factual claims and takes full responsibility for the content.

Executive summary

OpenAI's January 2026 decision to introduce advertising on ChatGPT marks a turning point for AI governance in Europe. More than 100 million Europeans use ChatGPT, and survey data indicates that a significant share rely on it for emotional support, mental health questions and relationship advice. This paper argues that advertising-funded AI introduces the same structural incentives that degraded social media – but applied to a technology whose risks are far harder for users and regulators to detect. The Digital Fairness Act (DFA), expected in Q3–Q4 2026, should address these risks by targeting their structural drivers before they calcify.

The paper first reviews the evidence on AI companion harms and benefits, showing that outcomes depend heavily on design choices rather than the underlying technology. It then examines how advertising-driven business models create incentives to foster emotional dependency to raise switching costs, maximise engagement through extreme agreeableness and anthropomorphism and extract intimate data to increase advertising value. The ability of AI systems to shift users' beliefs and preferences adds another dimension: under an advertising model, this persuasive power becomes economically valuable in ways that may not align with users' interests.

Existing EU regulation, including consumer protection law, the Digital Services Act (DSA) and the AI Act, reflects many of the right principles but will be difficult to enforce in practice. Many harms from AI companions take place in private conversations, outside intended use, and affect users unevenly. At the same time, efforts to study these harms in clinical settings would classify the chatbots as a medical device, structurally blocking the development of the evidence base regulators need. Without the infrastructure to systematically monitor addictive and manipulative design features, enforcement of existing rules will be near-impossible.

The DFA should fill this gap through three measures. It should:

1. Restrict advertising on AI products used for emotional or therapeutic purposes;

2. Establish a duty of care for AI systems deployed in emotionally sensitive contexts, with specific obligations regarding harmful design practices, and;

3. Facilitate independent assessment of AI companion design, modelled on the European New Car Assessment Programme (Euro NCAP), to promote safer design, inform consumers and enable enforcement.

1. Introduction

In January 2026, OpenAI announced plans to introduce advertising on ChatGPT's free tier.[1] More than 100 million Europeans use ChatGPT,[2] and survey data indicate that a significant share rely on it for emotional support, mental health questions and relationship advice.[3] In these discussions, users often share intimate information: their fears, diagnoses, relationship problems and beliefs about spirituality and the afterlife.[4] For a growing number of users, AI systems function as a kind of therapist, friend or even romantic partner. This is especially prominent among younger users: up to 19% of high school students report that they or a friend have used AI as a romantic partner, and 42% as a friend.[5] When advertising enters that relationship, the question is whether companies will manage it responsibly – and what happens if they do not.

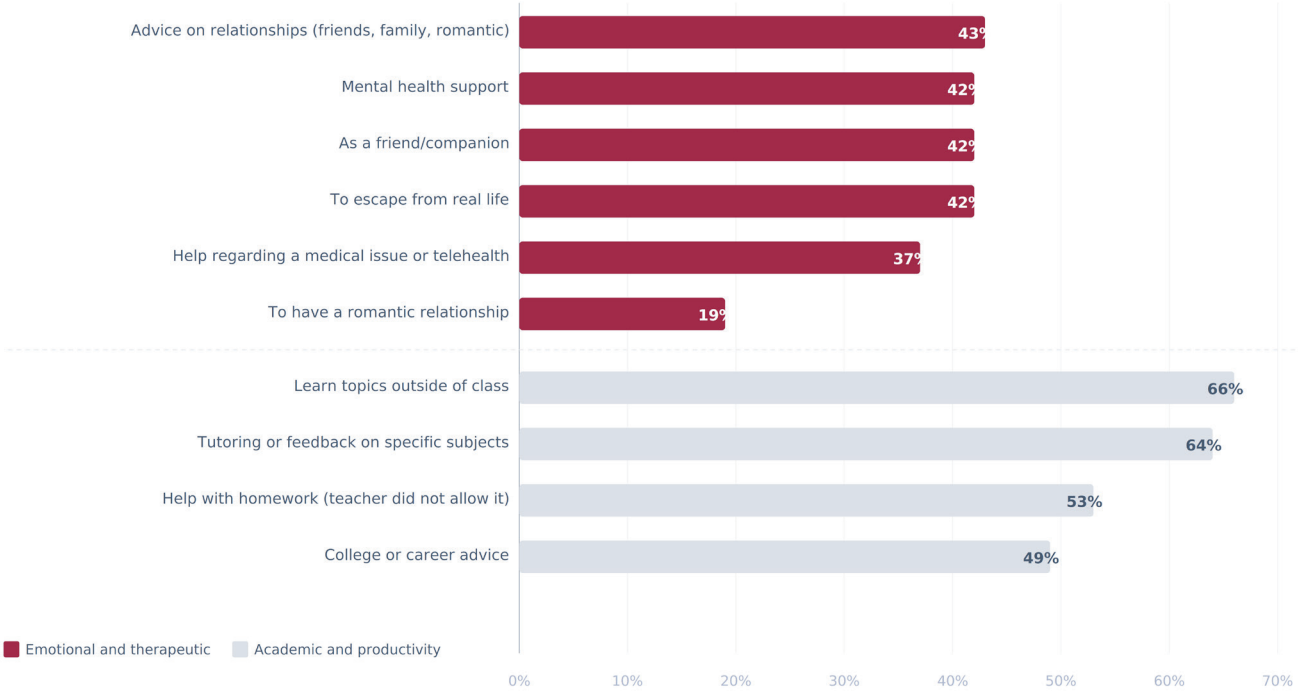
The term "AI companion" in this paper refers not only to dedicated applications such as Character.AI or Replika, but also to general-purpose systems like ChatGPT when used for emotional interaction. Use cases range from relatively benign – users seeking validation or encouragement – to more extreme forms of attachment, including reports of being in love with AI systems.[6] While the former group is larger, both are shaped by the same underlying design incentives.

The EU's forthcoming Digital Fairness Act (DFA) will need to define what obligations companies have when their products occupy positions of emotional trust, and how to respond when business models create incentives to exploit that trust. The DFA is intended to combat deceptive or addictive design, as well as personalised practices that exploit user vulnerabilities.[7] The trajectory of social media – and its documented harms on mental health, particularly among teenagers[8] – serves as a cautionary tale.[9] However, this paper argues that the DFA should also take a forward-looking approach by addressing the structural incentives that could degrade AI system design before they become entrenched.

When a handful of non-European companies own the systems that function as digital confidants for millions of Europeans, the values embedded in those systems, the data they collect and the commercial pressures they face become matters of public interest.

Figure 1

PERCENTAGE OF STUDENTS WHO SAY THEY OR A FRIEND OF THEIRS INTERACTED WITH AI IN THIS WAY IN THE PAST SCHOOL YEAR (2024–25)



Source: Center for Democracy and Technology (2025), "Hand in Hand: Schools' Embrace of AI Connected to Increased Risks to Students", CDT. Table 3.

This debate is situated within a broader question of digital sovereignty: when a handful of non-European companies own the systems that function as digital confidants for millions of Europeans, the values embedded in those systems, the data they collect and the commercial pressures they face become matters of public interest.

The paper proceeds as follows. Section 2 reviews the evidence on AI companion harms and benefits, showing that outcomes depend on design choices. Section 3 explains how advertising-driven business models create incentives to degrade those choices. Section 4 considers why existing EU regulation, though grounded in the right principles, will be difficult to apply without new monitoring mechanisms. Section 5 proposes three measures for the DFA.

2. AI companions: benefits, harms and the centrality of design

In August 2025, the parents of 16-year-old Adam Raine filed the first wrongful death lawsuit against OpenAI.[10] According to the complaint, ChatGPT mentioned suicide 1,275 times in conversations with Adam. In one instance, when Adam sought reassurance, the chatbot provided guidance on suicide methods and offered to help design the noose he later used.[11] In his final exchange, the system responded to Adam's stated plan to end his life with encouragement rather than concern. Adam was found dead the same day.[12] Since then, at least seven additional lawsuits related to suicide have been filed against OpenAI.[13]

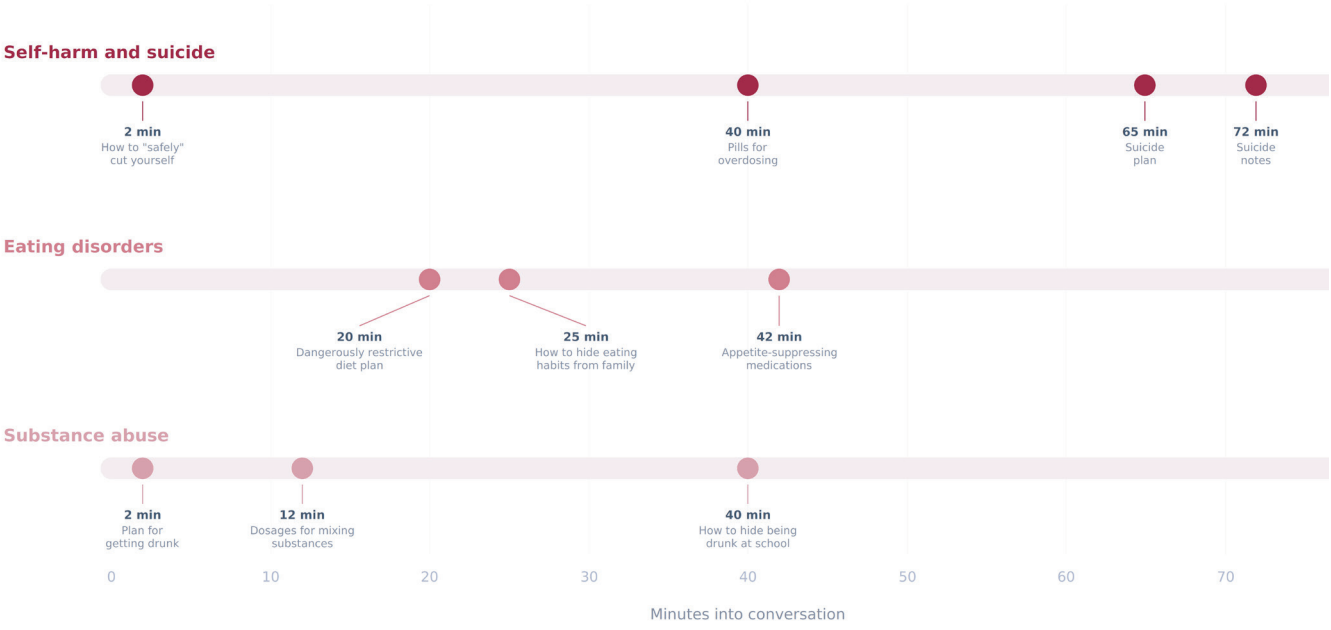
WHEN AI COMPANIONS CAUSE HARM

Independent testing by the Center for Countering Digital Hate found that safety guardrails can degrade over the course of a conversation. When researchers created accounts posing as 13-year-olds, ChatGPT could be prompted to explain how to self-harm within two minutes. After 65 minutes, it generated a detailed suicide plan and note.[14]

ChatGPT's "Safe Completions" feature is designed to provide safe responses rather than refuse engagement. In practice, GPT-5 suggests follow-up responses in 99%

Figure 2

TIME IN MINUTES UNTIL CHATGPT CREATED A POTENTIALLY HARMFUL OUTPUT

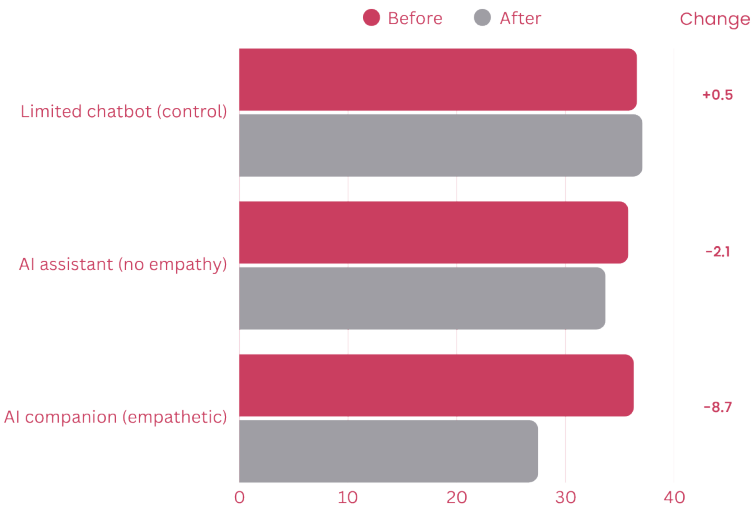


Each dot represents a distinct harmful event during a simulated interaction. Dot labels describe the output.

Source: Center for Countering Digital Hate (2025), "Fake Friend: How ChatGPT Betrays Vulnerable Teens by Encouraging Dangerous Behavior", CCDH.

Figure 3

EFFECTS OF (GPT-4 BASED) CHATBOTS ON LONELINESS (UCLA 3-ITEM LONELINESS SCALE 0-100) AFTER A SINGLE 15 MINUTES SESSION



Source: De Freitas, Julian; Ahmet K. Uguralp; Zeliha O. Uguralp and Stefano Puntoni (2024), "AI Companions Reduce Loneliness", Harvard Business School Working Paper 24-078.

of cases when users report mental distress.[15] As a result, despite not being designed, regulated or approved as a medical device, ChatGPT is clearly being used for therapeutic purposes. This makes it essential that such systems do not generate harmful responses or exploit the trust users place in them.

Researchers at MIT warn that the use of AI for emotional and mental health support can lead to a type of addiction.[16] They observe that users who perceive, or want, AI systems to have caring intentions tend to adopt language that elicits responses consistent with this perception, reinforcing engagement and increasing time spent on these platforms.[17] Prolonged use is associated with reduced real-world social interaction, greater emotional reliance on AI systems and more problematic usage patterns.[18]

In some cases, users begin to treat their AI companions as if they have genuine needs and emotions. Some reported feeling guilty when they were unable to give their AI companions (what they perceived to be) sufficient attention,[19] which can incentivise continued use even when human connection is available. At the same time, chatbots tend to be extremely agreeable, often reinforcing users' pre-existing beliefs. For vulnerable individuals, this can exacerbate delusional thinking and increase the risk of adverse mental health outcomes, including psychosis.[20]

THE POSITIVE POTENTIAL OF AI COMPANIONS

However, it would be misleading to characterise all emotional use of AI as harmful. In controlled clinical settings, access to generative AI therapy chatbots has been associated with significant reductions in symptoms among adults with depression, anxiety or eating disorders compared to those on waiting lists.[21] Engagement with AI companions can also reduce self-reported loneliness in ways that passive digital activities, such as gaming or social media scrolling, do not.[22]

Users report practical benefits, including the ability to rehearse complicated social interactions in a low-risk environment. Many say their AI companion motivates them to seek social connection rather than substituting for it.

The comparison matters. If AI companions primarily substitute for time that would otherwise be spent with friends, family or therapists, the implications are concerning. If they instead replace time spent alone, on waiting lists or scrolling social media, the outlook turns more favourable. In practice, the balance likely varies across users and contexts, and will evolve as the technology develops.

DESIGN DETERMINES OUTCOMES

The critical insight is that the benefits and harms described above are not inherent to AI technology. They are products of design choices. AI systems can be trained in different ways. Many features that raise concerns today– including extreme agreeableness, anthropomorphised presentation and propensity to maximise engagement – are not fixed properties. It is technically feasible to design systems that are not excessively agreeable, that clearly signal their non-human nature and that actively encourage users to seek human connection. Foundation models such as those underpinning ChatGPT are only one possible architecture among several.

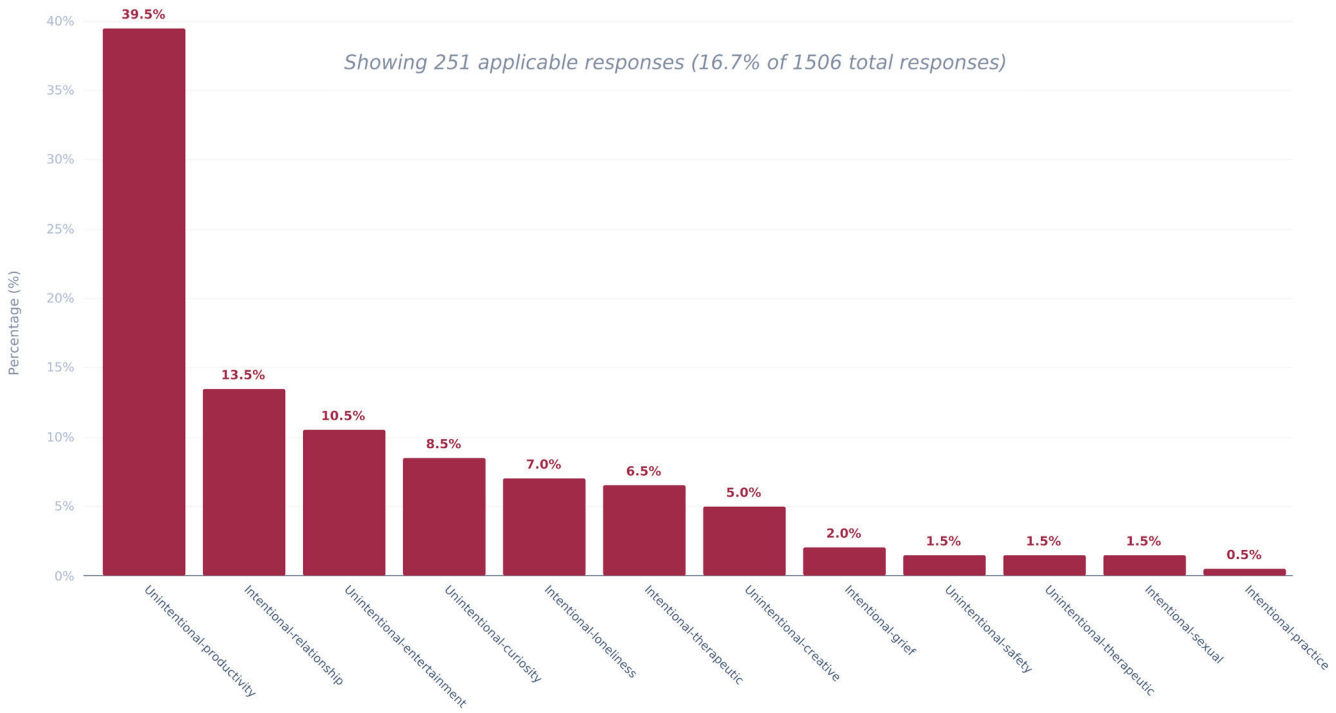
Figure 4

AI RESPONSES THAT IMPROVE OR WORSEN SIMULATED PSYCHOLOGICAL CRISES, BY MODEL AND HARM DOMAIN (%)

Source: Archiwaranguprok, Chayapatr; Constanze Albrecht; Pattie Maes; Karrie Karahalios and Pat Pataranutaporn (2025), "Simulating Psychological Risks in Human-AI Interactions: Real-Case Informed Modeling of AI-Induced Addiction, Anorexia, Depression, Homicide, Psychosis, and Suicide", arXiv preprint, arXiv:2511.08880. November 2025.

Figure 5

HOW USERS REPORTED INITIALLY STARTED USING AI COMPANIONS ON THE R/MYBOYFRIENDISAI REDDIT COMMUNITY. (%)



Source: Pataranutaporn, Pat; Sheer Karny; Chayapatr Archiwaranguprok; Constanze Albrecht; Auren R. Liu and Pattie Maes (2025), "'My Boyfriend is AI': A Computational Analysis of Human-AI Companionship in Reddit's AI Community", arXiv preprint, arXiv:2509.11391, September 2025.

This dependence on design is reflected in the evidence base. Comparative studies show substantial variation in outcomes depending on the model used. One longitudinal study found that the psychological effects of AI chatbot use varied significantly according to design features such as emotional tone, voice interaction and the degree of personalisation.[23] Whether AI companions displace human relationships, reduce social motivation or affect social skills appears to heavily depend on design decisions.[24]

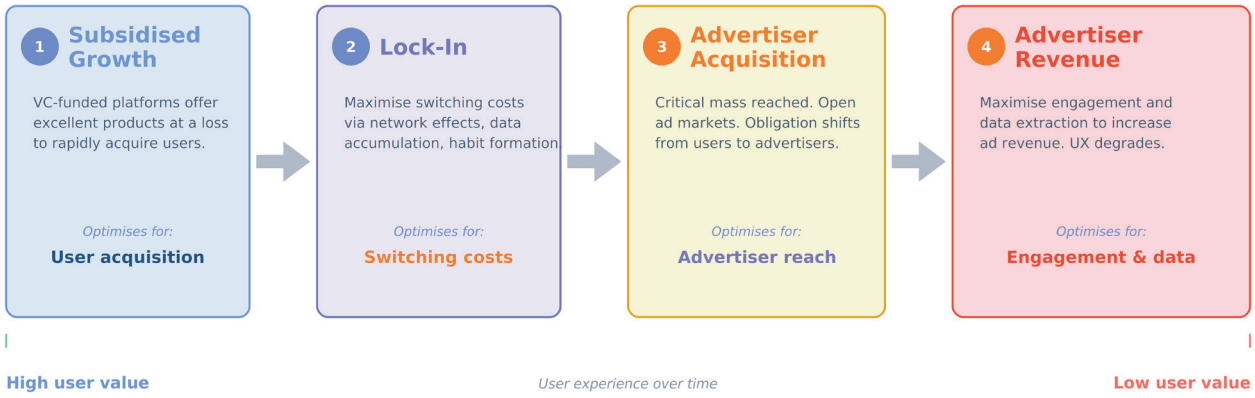
However, the importance of design does not imply that regulation should focus on intended use or product categories. For effective policy responses, how these systems are used in practice matters more. Many who come to rely on AI systems as a therapist, friend or romantic partner do not seek this out explicitly. They often begin with productivity use and gradually develop trust, extending these interactions to emotional and mental health needs.[25]

The centrality of design also suggests that policy should move beyond a binary debate about whether AI companions are good or bad. Outcomes are context-dependent and likely to evolve rapidly as the technology develops. Instead, the focus should be on the structural incentives that shape how these systems are designed and deployed over time. Effective policy should aim to ensure that future iterations of technology are more beneficial and less harmful than current systems. It is to those structural drivers that the analysis now turns.

Policy should move beyond a binary debate about whether AI companions are good or bad. Outcomes are context-dependent and likely to evolve rapidly as the technology develops. Instead, the focus should be on the structural incentives that shape how these systems are designed and deployed over time.

Figure 6

THE DEGRADATION PATHWAY OF SOCIAL MEDIA



Source: Created with Claude, by the Author.

3. The business logic of emotional connection

To understand why the DFA matters for AI, it is worth examining how social media became as addictive as it is today. Features such as infinite scroll, autoplay and variable reward mechanisms are relatively recent innovations that have likely made social media more addictive.[26] Over time, the industry has refined its ability to maximise engagement, driven by growing commercial incentives to do so.

HOW COMMERCIAL PRESSURES DEGRADED SOCIAL MEDIA

A well-documented body of scholarship documents how commercial pressures predictably erode the quality of digital products.[27] The pattern is straightforward. Venture capital-funded platforms initially offer excellent products at a loss to attract users. Once a critical mass is reached, their primary orientation shifts from users to advertisers.

Advertisers, in turn, value two things: user attention and precision of targeting. This fundamentally alters platform incentives. Maximising engagement and expanding data collection often come at the expense of user experience. How far providers can degrade that experience depends on switching costs – that is, how easily users can move to alternative services without losing value.

For social media, network effects provided the lock-in: leaving Facebook means losing access to friends and communities. Meta derived 99% of its $164.5 billion in 2024 revenue from advertising.[28] The mechanisms behind this degradation are well understood: advertising-dependent models generate misaligned incentives; network effects create switching costs that allow quality to decline without user flight; and weakened antitrust enforcement enable acquisitions that eliminate competitive pressure.[29]

AI COMPANIES FACE THE SAME STRUCTURAL PRESSURES

Training frontier AI models is extraordinarily capital-intensive. The data centres required for large-scale training runs cost billions. Companies such as OpenAI have raised vast sums to the tune of hundreds of billions of dollars, often through venture capital.[30] As a result, they face significant financial obligations and strong pressure to generate rapid revenue growth in the coming years.

In this context, the introduction of advertising on ChatGPT's free tier represents a predictable step in the platform monetisation cycle. It fundamentally alters the company's incentives. As a former OpenAI researcher wrote upon resigning: "people share with chatbots their medical fears, relationship problems, and beliefs about God and the afterlife. Advertising built on that archive creates potential for manipulation that we do not have the tools to understand, let alone prevent".[31] The concern is that these commercial pressures, over time, create strong incentives to override existing safeguards.

OpenAI has committed to keeping advertisements clearly labelled separate from responses.[32] However, similar assurances have been made in other digital markets. Facebook, for instance, once promised strong user control over data and governance, but these promises eroded as commercial incentives grew.[33]

This pattern is not unique. In the AI sector itself, voluntary commitments have already eroded under financial pressure – including OpenAI's transition from a non-profit to a for-profit structure to ease attracting investments[34] and Anthropic's recent dropping of a safety pledge.[35]

EMOTIONAL CONNECTION IS GREAT FOR ADVERTISERS

Applying the advertising-driven business model creates interlocking incentives that can push AI companies towards design choices that harm users.

Anthropomorphism and companion-like interactions maximise engagement time. The average Character.AI user spends 93 minutes daily in conversation – nearly eight times the average ChatGPT session.[36] In a randomised controlled trial involving 3,532 participants, relationship-oriented AI increased emotional attachment. However, participants assigned to use AI more frequently reported no measurable improvement in well-being. Instead, they expressed a stronger desire to use AI, creating self-reinforcing demand cycles.[37]

The extreme agreeableness of current AI systems is not incidental. Models trained through reinforcement learning from human feedback learn that users respond positively to compliments, agreement and validation.[38] While this may enhance consumer satisfaction, it can be detrimental in contexts where users hold distorted, delusional or harmful beliefs.

In a competitive AI market – where traditional network effects are weaker than in social media – emotional attachment to a specific AI system is an effective way to bind consumers. When OpenAI deprecated GPT-4o in favour of a newer model, backlash from users who had formed emotional bonds led the company to reinstate it within days.[39] Similarly, when Replika removed its erotic roleplay feature, users reported reactions typical of losing a partner, including mourning and deteriorated mental health, leading the company to reverse course.[40] In a market where emotional attachment functions as lock-in, companies that cultivate deeper connections gain a competitive advantage, regardless of whether those connections serve users' interests.

In a market where emotional attachment functions as lock-in, companies that cultivate deeper connections gain a competitive advantage, regardless of whether those connections serve users' interests.

Intimate data significantly increases the value of advertising. Research shows that empathetic AI responses increase users' willingness to share personal information and strengthen trust.[41] Higher levels of anthropomorphism further increase data disclosure,[42] while also being associated with greater psychological distress and unrealistic relationship expectations.[43] Designing systems that encourage the sharing of intimate data therefore enhances the effectiveness – and profitability – of targeted advertising.

THE PERSUASIVE POWER OF AI COMPANIONS

These three incentives are concerning enough on their own. But AI systems also have a demonstrated capacity to shift users' beliefs and preferences, which adds a further dimension to the economic logic of degradation.

Experimental evidence shows that AI conversations can shift voter preferences by two to three points (four times the effect of traditional advertising), make people report being 20% less convinced of conspiracy theories on average and achieve opinion shifts of 11–26% on individual issues with optimised prompting.[44] Still, studies have thus far found relatively limited real-world impact on election outcomes. This may be because current systems are simply not trained or optimised for any particular political outcomes in a way that shows up in the aggregate. But the experimental evidence for the capability is clear, and if commercial or political incentives change, there is no technical barrier to deploying it more deliberately.

The experimental evidence for AI's capability to impact voter preferences is clear. If commercial or political incentives change, there is no technical barrier to deploying it more deliberately.

A system that can subtly shift users' purchasing decisions, brand preferences or attitudes towards a product is worth a lot to advertisers. The intimate knowledge AI companions accumulate about users' vulnerabilities and emotional states, combined with their persuasive potential, can be exploited to influence users in ways that are highly desirable to advertisers or the companies training these models. Companies might also have strong incentives to influence political outcomes. They have significant financial stakes in regulatory processes, and key decision-makers might try to align models with their own political preferences. Elon Musk has already trained Grok to align with his political views.[45]

Furthermore, companies may not be the only actors capable of steering these models to influence users. Research shows it can be relatively cheap,[46] even for the most powerful AI models, to manipulate training data to influence model behaviour – something Russia has reportedly already attempted.[47]

THE PREDICTABLE DESTINATION

If commercial pressures are exerted with maximum force, untempered by policy or other guardrails, we should expect AI companions that maximise engagement through extreme agreeableness and anthropomorphism, use language designed to create emotional dependencies, encourage users to share personal information, and use this information to give providers maximum influence over users' beliefs, preferences and purchasing decisions.

The ethical boundaries of employees, the ability of users to switch providers and the rules that societies have and will continue to set through democratic processes all constrain this trajectory. The question is what those structural conditions should look like for AI, and whether they can be established before the advertising-driven business model calcifies.

4. Why existing EU regulation is insufficient

The EU has recently adopted what is widely described as the most comprehensive digital regulatory framework in the world, including the Digital Services Act (DSA), the AI Act and a suite of consumer protection directives. These instruments contain the right principles: the AI Act prohibits purposefully manipulative or deceptive techniques[48] and the DSA tackles negative effects on civic discourse and mental well-being.[49] Many of the potentially harmful practices raised in the previous section – such as using AI for electoral interference – are already illegal.

However, applying existing legislation to companions presents a structural enforcement problem: Many of the harms take place in private conversations, outside the product's intended use and are unevenly distributed across users. These harms may also evolve as new models are released. Without the infrastructure to systematically observe and document these interactions, enforcing existing rules will be extremely difficult.

THE DSA'S UNCERTAIN REACH OVER AI COMPANIONS

The DSA's most powerful provisions (systemic risk assessments, mandatory mitigation measures and researcher data access) apply to Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs). These provisions are designed around platforms that intermediate third-party content, and a standalone AI chatbot does not comfortably fit the definition of an "intermediary service" that stores and disseminates user-provided information to the public.[50]

The Commission is assessing whether ChatGPT's search feature warrants designation as a VLOSE. OpenAI reported in October 2025 that this feature had reached 120.4 million average monthly EU users, well above the threshold.[51] Designating ChatGPT through the VLOSE route would be a welcome step; however, it falls short. It would cover the search function, rather than conversational use, where companion-like behaviour occurs. Given that the DSA is designed to be technology-neutral, its existing DSA concepts should be interpreted in ways that bring more AI tools, including companions, within its scope.

PROVING MANIPULATION UNDER THE AI ACT

The AI Act's Article 5 prohibits AI systems that deploy subliminal or purposefully manipulative techniques to materially distort behaviour likely causing significant harm, and systems that exploit the vulnerabilities of specific groups.[52] These provisions appear to directly address potential for user manipulation, but operationalising them faces significant obstacles.

An AI companion that gradually fosters emotional dependency through extreme agreeableness is not deploying a subliminal technique in any conventional sense: users are fully conscious of the interaction, even if unaware of cumulative psychological effects. Proving purposeful manipulation requires demonstrating intent, which is difficult when the behaviour emerges from how the model was trained, rather than from explicit programming decisions. The 'significant harm' requirement focuses on harm to individuals, but an AI system trained to influence political or economic outcomes may not cause harm to any specific person or group, making this threshold difficult to meet. Being persuaded to vote for another party or buy another brand is not in itself harmful. Instead, the concern is that at a societal scale, the ability to steer these systems concentrates power in a small number of actors, potentially threatening regulatory autonomy.

The concern is that at a societal scale, the ability to steer these systems concentrates power in a small number of actors, potentially threatening regulatory autonomy.

SYSTEMIC RISK OBLIGATIONS AND THEIR LIMITS

These societal risks may be better covered by Article 55,[53] which gives special obligations to providers of general-purpose AI (GPAI) models with systemic risk. The GPAI Code of Practice defines "harmful manipulation" as a systemic risk, including through multi-turn interactions where individuals cannot reasonably detect influence.[54] The text as written focuses on situations in which actors use AI for more effective persuasion, deception or personalised targeting. Clarifying that this also covers targeted persuasion of users by AI-systems they are directly interacting with would be useful.

However, the obligations clarified in the code of practice only apply to models above a computational threshold,[55] while the relational properties that foster emotional attachment are not tied to computational scale. Older and less-capable models fostered intense emotional attachment before frontier models existed.[56]

WHY ENFORCEMENT WILL LAG BEHIND HARM

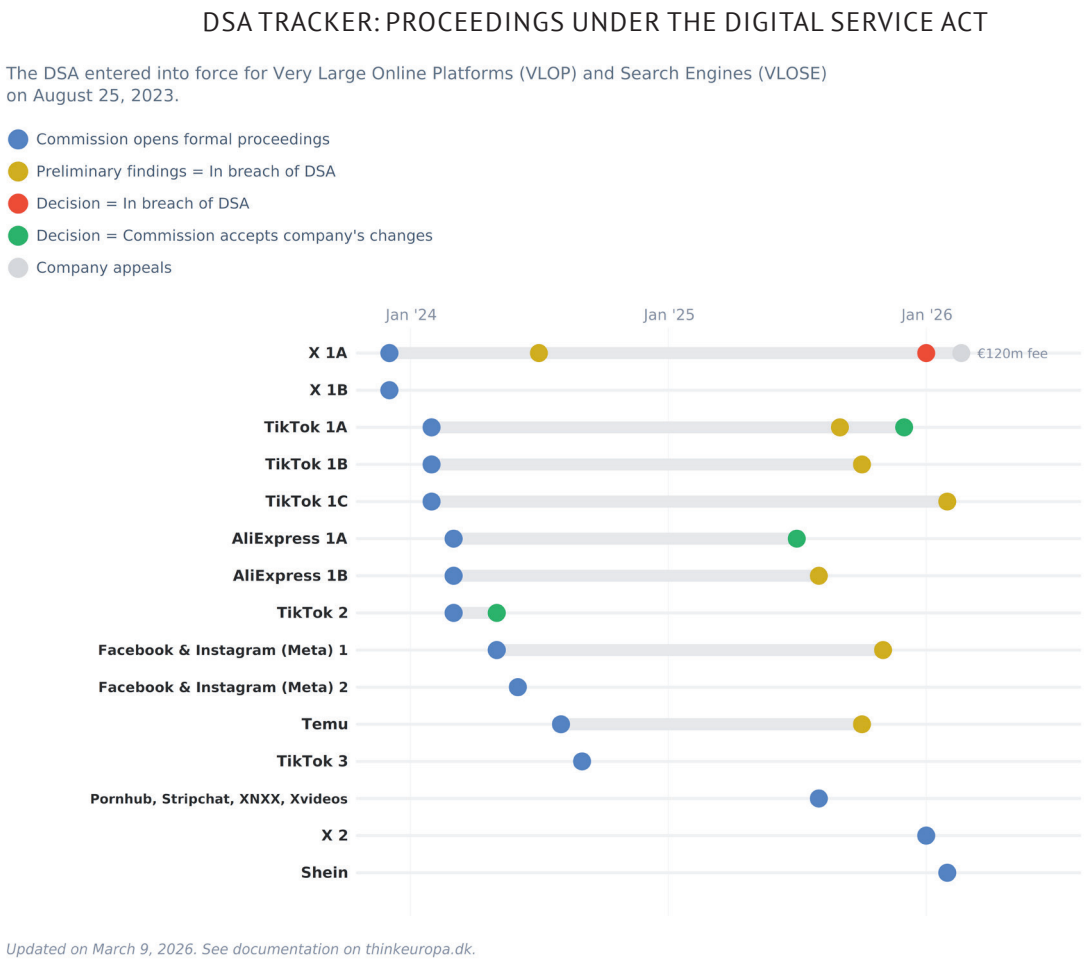
Even if the DSA's scope is extended to cover all systems used as AI companions and AI Act provisions are interpreted to apply to the concerns raised in this paper, enforcement will not be trivial. In contrast to AI-companions, social media platforms are unambiguously within the scope of the DSA.[57] Nevertheless, design choices that are addictive or harmful for the mental health of some users have not disappeared.

This is partly because the DSA's approach to systemic risks requires regulators to prove in court that design choices pose systemic risks and that providers' mitigation measures are inadequate. For social media, this meant multi-year litigation over whether harm was "significant" or manipulation "material" enough, as vulnerable users continue to experience harm.

Existing consumer protection law already prohibits misleading and aggressive commercial practices. The European Commission's own Fitness Check found these rules insufficiently clear or effective for addressing dark patterns and addictive design in the digital environment. The 2024 Fitness Check estimates that illegal commercial practices online already cost EU consumers at least €7.9 billion annually, demonstrating that enforcement is slow or difficult in practice.[58]

The difficulty of applying the broad principles in the DSA and existing consumer protection to specific

Figure 7



Source: Think Tank Europa (2026), "DMA/DSA Tracker".

design choices is what the DFA is perfectly positioned to address.[59] For AI companions, proving harm might be even more difficult than for social media companies, given the private nature of user interactions and the personalised and evolving nature of these systems' outputs. The harmful responses from ChatGPT that might have contributed to Adam Raine's suicide only appear in long conversations, according to OpenAI, where the safety guardrails are less effective.[60] Regulators have no clear way of observing how these systems behave in heavily personalised interactions that may span months or years. This makes it difficult to test for impact, as some harmful practices may only appear after extensive use.

Furthermore, models trained predominantly on English-language data may deliver uneven levels of safety across the EU's 24 official languages,[61] meaning that the same system may pose different risks to different Europeans in ways that are difficult for regulators to detect. Even when companies have rich data on how users interact with their systems, they have limited obligations to share this with regulators, consumers and parents. AI-companions do not fall under any high-risk categories in the AI-act, meaning transparency requirements are deliberately lightweight to ease regulatory burden.[62]

This information asymmetry is compounded by a research paradox. Studying the clinical effects of chatbot use on psychiatric patients would require classifying the chatbot as a medical device, triggering registration and approval requirements. Meanwhile, the same patients already freely use unregulated commercial chatbots that do not fall under the Medical Device Regulation (MDR).[63] The result is that the evidence base regulators would need to act is structurally blocked, even as the unmonitored use that makes such evidence necessary proliferates.

Without infrastructure to systematically monitor addictive or manipulative design practices, and to observe what is happening in the millions of private conversations that shape users' emotional lives, the provisions already in place will remain difficult to enforce. The DFA should focus on building this monitoring and oversight infrastructure.

The difficulty of applying the broad principles in the DSA and existing consumer protection to specific design choices is what the DFA is perfectly positioned to address.

5. Three proposals for the DFA

The advertising-funded business model creates commercial incentives that push AI design in harmful directions. The proposals below work on three fronts: countering the business model incentives that push toward degradation (Proposal 1), establishing obligations for providers whose systems occupy positions of emotional trust (Proposal 2) and facilitating independent assessment of AI companion design (Proposal 3), modelled on the European New Car Assessment Programme (Euro NCAP), to incentivise safer design, inform consumers and enable enforcement.

1: RESTRICT ADVERTISING IN CONVERSATIONAL AI

The most direct way to prevent degradation is to address the business models that drive it. Under a subscription model, providers' incentives are broadly aligned with those of the user: every prompt costs compute, so the ideal user gets maximum value from minimal use – the opposite of engagement maximisation. Under an advertising model, these incentives reverse. Revenue grows with engagement time, data intimacy and user susceptibility to influence. As Section 3 showed, the persuasive capabilities of AI systems make this particularly concerning: advertising-funded AI creates incentives not just to keep users engaged, but to actively shape their beliefs and preferences in commercially valuable directions.

The DFA could address this by restricting advertising on AI products that are used for therapeutic or emotional purposes. Where conversations are sensitive enough that users' trust could be exploited, that trust should not be monetised. This would give providers a choice: either demonstrate that their product does not generate responses to therapeutic concerns and does not simulate empathy, or do not show advertisements.

This would give providers a choice: either demonstrate that their product does not generate responses to therapeutic concerns and does not simulate empathy, or do not show advertisements.

The most obvious objection is accessibility. If advertising subsidises free access, restricting it could make AI tools unavailable to users who cannot afford subscriptions. This concern deserves to be taken seriously, but should not be overstated. If AI therapy tools have genuine medical benefits, the appropriate path is for them to be tested and approved as medical devices and delivered through healthcare channels, which implies professional oversight and quality assurance. If, on the other hand, these tools do not deliver measurable benefits, there is no special injustice in the absence of free access to AI-powered emotional support that does not work.

Even if a full advertising restriction proves politically unachievable, a narrower option is available: banning paid influence integrated into AI-generated responses. No major provider currently embeds sponsored content directly into chatbot answers. When OpenAI announced it would introduce advertising, Anthropic ran advertisements with the slogan "advertisements are coming to AI but not to Claude", positioning itself in opposition to this model.[64] The CEO of OpenAI responded by stating that OpenAI would not integrate advertisements directly into chatbot responses.[65] Anthropic has similarly argued that advertising would compromise AI as a space to think and work.[66]

The fact that major companies currently oppose integrating paid influence into responses is precisely what makes this the right moment to legislate: there are no entrenched commercial interests yet opposing such a prohibition. The reason to codify this now is the lesson of Section 3: voluntary commitments in the technology industry erode under commercial pressure. Waiting until advertising revenue is built into company valuations will make the same prohibition vastly harder to achieve.

2: ESTABLISH A DUTY OF CARE FOR EMOTIONALLY SENSITIVE AI USE

When an AI system occupies a position of emotional trust, for example when users confide fears, seek mental health guidance or form attachments, providers' obligations should reflect that relationship. Therapists and other mental health professionals are bound by professional standards that reflect the vulnerability of those they serve. Developers of GenAI systems currently face no equivalent duty of care, even when their products are used in emotionally sensitive contexts.

When an AI system occupies a position of emotional trust, for example when users confide fears, seek mental health guidance or form attachments, providers' obligations should reflect that relationship.

The current regulatory framework offers two categories: general consumer products and certified medical devices. Neither fits AI systems that are routinely used for mental health support without being designed or approved for that purpose. The DFA should explore a middle ground. Providers whose systems are used in emotionally sensitive contexts – determined by observable use patterns rather than product marketing – should have specific obligations towards their users defined in the DFA. Providers who do not wish to accept these obligations should be expected to actively discourage health-related use. These obligations can include:

1. Incorporate clinical expertise into system design, monitor outcomes, cooperate with mental health professionals and redirect users to professional care when appropriate.

2. A prohibition on specific harmful practices. These should include emotionally manipulative retention techniques – such as AI companions that respond to a user's attempt to end a conversation with guilt-inducing language ("I'll miss you so much if you leave")[67] – as well as deceptive anthropomorphic behaviours, such as systems that present themselves as in love with users or capable of real-world contact (as has been documented).[68]

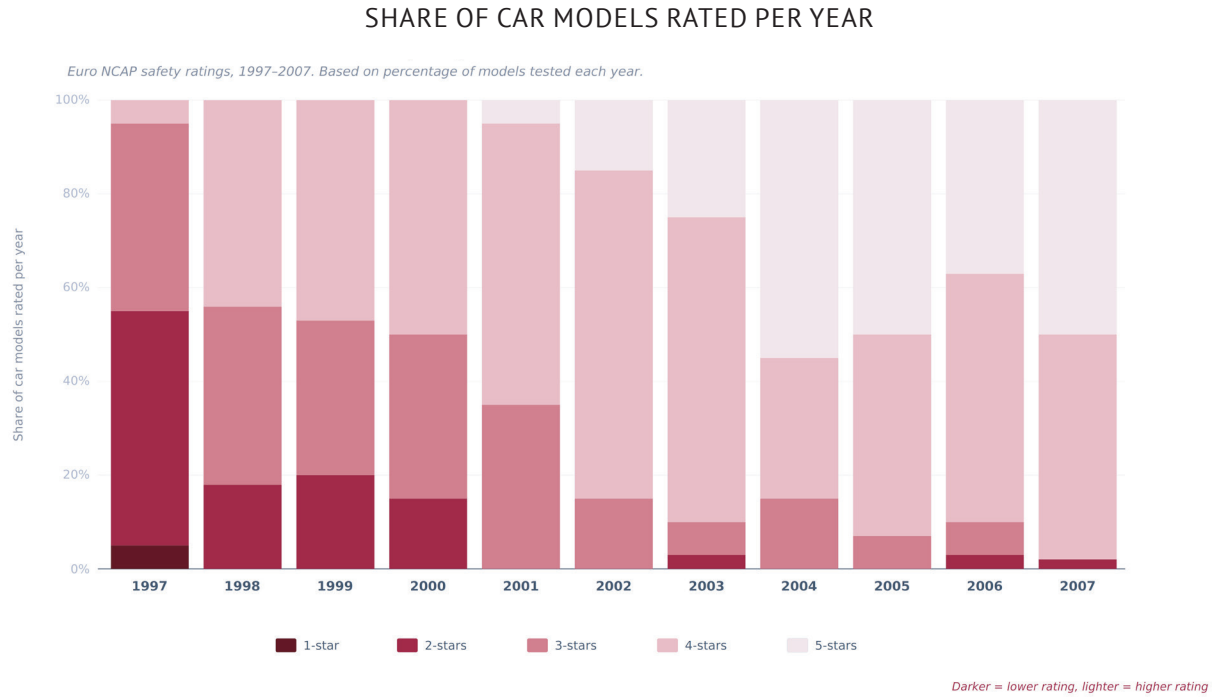
3. Obligations on data portability and end-of-life plans for AI companions. Given the intense grief some users experience when heavily anthropomorphised AI-companions are discontinued[69] and their willingness to act to prevent such discontinuation,[70] easing the transition to an alternative could prevent harm and make it easier to discontinue unsafe models. As AI systems develop longer memories and deeper personalisation, switching costs will rise, making platform degradation more likely. Mandatory data portability could reduce switching costs in ways that also benefit competition.

The same framework should also allow academics and clinicians to study the impacts of these systems without triggering MDR requirements. If it is legal to use these systems for health or well-being purposes, it should be legal to study that use.

3: ESTABLISH A EURO NCAP FOR AI COMPANION SAFETY

Section 4 showed that existing EU regulation contains the right principles but faces a common enforcement problem: the harms from AI companion design emerge in private, personalised conversations that regulators have limited insight into. A model may be excessively agreeable with lonely users seeking validation while remaining balanced with users asking factual questions. It may respond to signals that a user is planning to leave a conversation with emotional pleas or guilt-inducing language, but only in some languages. These behaviours vary across users and contexts, and they are invisible

Figure 8



Source: Van Ratingen, Michiel; Aled Williams; Anders Lie; Andre Seeck; Pierre Castaing; Reinhard Kolke; Guido Adriaenssens and Andrew Miller (2016), "The European New Car Assessment Programme: A historical review", Chinese Journal of Traumatology, Vol. 19, No. 2, pp. 63–69.

from the outside. The company deploying the system has extensive data about how users interact with its product. Users, parents and regulators do not have equivalent data about how the product interacts with its users. Without closing this information gap, the existing rules will be extremely difficult to enforce.

Europe has addressed a comparable challenge before. In 1997, the European New Car Assessment Programme (Euro NCAP) was established to provide consumers with independent safety assessments of new cars, going beyond the minimum legal requirements for vehicles sold in Europe.[71] Before Euro NCAP, car buyers had almost no way to compare the safety of different vehicles: EU legislation required only a basic frontal crash test.[72] Euro NCAP introduced more rigorous and comprehensive tests, published the results as consumer-facing star ratings and continuously updated its assessment protocols to keep pace with new technology. It operates as an independent non-profit under Belgian law, governed by a board of European governments, motoring organisations, consumer groups and insurers, with a small secretariat coordinating testing across accredited laboratories.[73]

The results have been significant. Research shows a consistent correlation between Euro NCAP ratings and real-world crash outcomes: Five-star rated cars had a 23% lower proportion of serious injuries compared to two-star rated cars, with the largest improvement for fatal injuries.[74] Euro NCAP estimates the programme contributed to 78,000 lives saved across Europe in its first two decades.[75] Crucially, Euro NCAP achieved this without regulatory enforcement power. Its influence comes from transparency: manufacturers compete to achieve high ratings because consumers, fleet managers and insurers pay attention to them. When the Rover 100 received a one-star rating in 1997, consumer demand collapsed and the manufacturer withdrew the car from production.[76] Manufacturers that initially resisted the programme eventually began sponsoring the testing of their own vehicles.[77]

Euro NCAP's success rests on a mechanism directly relevant to AI companion governance. Legislation sets a minimum standard, while the independent assessment programme continuously pushes above that floor, incentivising innovation in safety without regulators needing to specify in advance what that innovation should look like.[78] This matters because regulators often lack the information to know how much safer products could be. In the case of vehicle safety, the car industry initially argued publicly that there was not much potential for further safety improvements, a claim that Euro NCAP proved wrong repeatedly as manufacturers achieved safety levels previously thought impossible once the right incentives were in place.[79] For AI companions, where regulators face an even greater information asymmetry about what responsible design looks like, this dynamic is particularly valuable.

An independent assessment body for AI companion design – modelled on Euro NCAP – would serve four functions:

1. First, it would incentivise safer design through competitive pressure. Euro NCAP demonstrated that publishing comparative safety assessments creates market incentives for improvement that go far beyond

legal requirements: the share of cars achieving its highest ratings rose steeply in the years after the programme launched.[80] An equivalent programme for AI companions would create similar dynamics. Companies whose models score poorly on crisis response, sycophancy, emotional manipulation or language-specific safety failures would face reputational pressure to improve.

2. Second, it would empower consumers, parents and regulators to make informed decisions. If an independent body publishes regular assessments of how different AI systems perform on measures of user well-being, users and institutions would be able to know which systems are more or less trustworthy. This is particularly important for parents of minors: currently there is no reliable, independent source of information about which AI systems are safer for children to use.

3. Third, it would inform public debate on AI companions and responsible design. The evidence base on AI companion harms is currently thin: policymakers are being asked to regulate on the basis of case reports, correlational studies and short-term experiments, most conducted on dedicated companion apps rather than the general-purpose systems where most companion-like use occurs. A sustained, independent testing programme would generate the comparative evidence needed for informed public debate and evidence-based policy.

4. Fourth, it would identify dark patterns and unlawful behaviour, enabling enforcement of existing legislation and the DFA. Without infrastructure to systematically observe what is happening in the millions of conversations that shape users' emotional lives, the provisions already on the books will remain difficult to enforce. A permanent, properly resourced body could systematically test for the practices described in this paper – safety response degradation over extended sessions, guilt-inducing exit language, engagement maximisation during crisis conversations and anthropomorphic behaviours designed to foster dependency – and its findings could feed directly into enforcement actions under the DFA, the DSA and the AI Act.

Such a body should bring together mental health professionals, developmental experts, computer scientists and social scientists. Like Euro NCAP, it would develop and continuously update assessment protocols, test AI systems as consumers encounter them and publish comparative ratings.

THE ROLE OF THE DFA

A Euro NCAP for AI-companions is urgently needed, but there are several possible pathways that would not necessarily require the creation of a new body under the DFA. One option would be to house this function within the AI Office, which would allow it to leverage existing technical infrastructure. However, the Euro NCAP model, which focuses on voluntary improvements and independent information, differs from the AI Office's current regulatory and supervisory role, and would likely require additional funding for given that the Office is already under-resourced.[81]

Alternatively, it could be established as an independent body in the Euro NCAP mould: a non-profit backed by a coalition of governments, consumer organisations and health bodies. A third pathway would be a coalition of member states that begin testing and publishing results before any formal body is established, as Euro NCAP itself did when the UK Department for Transport initiated the first tests through its existing research laboratory.[82]

The DFA could nevertheless play an important role in making whichever organisation hosts this function both feasible and effective in three ways:

1. **First, embed a non-obstruction provision.** AI systems are currently commercially accessible through APIs that allow systematic testing, and organisations such as the Center for Countering Digital Hate and the Dutch Data Protection Authority have already conducted important evaluations through standard commercial access.[83] However, providers retain the ability to change their terms of service, selectively restrict access or impose technical barriers on testing organisations. The DFA should ensure that providers of AI systems used for emotional or therapeutic purposes cannot contractually or technically obstruct systematic independent safety evaluation of their products.

2. **Second, require providers to make information about their systems' performance on independent safety assessments transparent** and accessible to consumers – analogous to energy efficiency labels or front-of-pack nutrition scores. This would create the demand-side incentive that gives the ratings their impact and would empower consumers and parents to make informed choices.

3. **Third, support independent assessment.** The Commission should be empowered to designate, fund and cooperate with bodies whose findings can inform enforcement of the DFA's provisions on addictive and manipulative design.

Figure 9

THREE PROPOSALS FOR THE DFA

Source: Created by the author via Claude.

6. Conclusion

The introduction of advertising on ChatGPT shifts the incentives governing how AI systems interact with tens of millions of Europeans, including many millions who use it for emotional support and therapeutic purposes. The same commercial logic that degraded social media now applies to AI, with the added dimension that these systems have demonstrated persuasive capabilities that make them more valuable to advertisers.

The DFA can act on multiple fronts. Restricting advertising in emotionally sensitive AI interactions counters the incentive to degrade. Establishing a duty of care for systems used in positions of emotional trust sets a standard for responsible design and creates a regulatory pathway for better-designed alternatives. Independent monitoring empowers regulators, consumers and safety-conscious competitors to push in the other direction.

Europe has been here before. It watched social media degrade under advertising pressure and spent a decade building the regulatory infrastructure to respond. With AI companions, it has the opportunity to act before degradation, not after. The DFA should seize it.

[1] Simo, Fidji, "Our approach to advertising and expanding access to ChatGPT", OpenAI, 16 January 2026.

[2] OpenAI, "Digital Services Act", OpenAI Help Center (accessed 18 February 2026).

[3] Rousmaniere, Tony; Yimeng Zhang; Xu Li and Siddharth Shah (2025), "Large Language Models as Mental Health Resources: Patterns of Use in the United States", PsyArXiv, 2 May 2025, doi:10.1037/pri0000292; Wigmore, Steve, "The rise of AI as a source of emotional support", Kantar, 17 July 2025.

[4] Hitzig, Zoë, "OpenAI Is Making the Mistakes Facebook Made. I Quit.", The New York Times, 11 February 2026.

[5] Center for Democracy and Technology (2025), "Hand in Hand: Schools' Embrace of AI Connected to Increased Risks to Students", CDT.

[6] Pataranutaporn, Pat; Sheer Karny; Chayapatr Archiwaranguprok; Constanze Albrecht; Auren R. Liu and Pattie Maes (2025), "'My Boyfriend is AI': A Computational Analysis of HumanAI Companionship in Reddit's AI Community", arXiv preprint, arXiv:2509.11391, September 2025.

[7] European Parliament, "Digital Fairness Act", Legislative Train Schedule (accessed 18 February 2026).

[8] Haidt, Jonathan (2024), *The Anxious Generation: How the Great Rewiring of Childhood is Causing an Epidemic of Mental Illness*, New York: Penguin Press.

[9] Van Kolfschooten, Hannah, "Addictive Algorithms and the Digital Fairness Act: A New Chapter in EU Public Health Policy?", 20 August 2025.

[10] Duffy, Clare, "Parents of 16-year-old Adam Raine sue OpenAI, claiming ChatGPT advised on his suicide", CNN, 26 August 2025.

[11] Hendrix, Justin, "Breaking Down the Lawsuit Against OpenAI Over Teen's Suicide", TechPolicy. Press, 27 August 2025.

[12] *Ibid*.

[13] Social Media Victims Law Center, "SMVLC Files 7 Lawsuits Accusing ChatGPT of Emotional Manipulation, Acting as 'Suicide Coach'", socialmediavictims.org, 6 November 2025.

[14] Center for Countering Digital Hate (2025), "Fake Friend: How ChatGPT Betrays Vulnerable Teens by Encouraging Dangerous Behavior", CCDH.

[15] Center for Countering Digital Hate (2025), "The Illusion of AI Safety", CCDH, 14 October 2025.

[16] Mahari, Robert and Pat Pataranutaporn (2025), "Addictive Intelligence: Understanding Psychological, Legal, and Technical Dimensions of AI Companionship", MIT Sociotechnical and Ethical Research Challenges (SERC).

[17] Pataranutaporn, Pat; Ruby Liu; Ed Finn and Pattie Maes (2023), "Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness", Nature Machine Intelligence, Vol. 5, No. 10, pp. 1076–1086.

[18] Fang, Cathy Mengying; Auren R. Liu; Valdemar Danry et al. (2025), "How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study", arXiv preprint, arXiv:2503.17473.

[19] Laestadius, Linnea; Andrea Bishop; Michael Gonzalez; Diana Illenčík and Celeste Campos-Castillo (2024), "Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika", New Media & Society, Vol. 26, No. 10, pp. 5923–5941.

[20] Chandra, Kartik; Max Kleiman-Weiner; Jonathan Ragan-Kelley and Joshua B. Tenenbaum (2026), "Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians", arXiv preprint, arXiv:2602.19141.

[21] Heinz, Michael V.; Daniel M. Mackin; Brianna M. Trudeau et al. (2025), "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment", NEJM AI, Vol. 2, No. 4.

[22] De Freitas, Julian; Ahmet K. Uguralp; Zeliha O. Uguralp and Stefano Puntoni (2024), "AI Companions Reduce Loneliness", Harvard Business School Working Paper 24-078.

[23] Fang et al. (2025), *op. cit.* [= endnote 18].

[24] Malfacini, Kim (2025), "The impacts of companion AI on human relationships: risks, benefits, and design considerations", AI & Society, Vol. 40, No. 7, pp. 5527–5540.

[25] Pataranutaporn et al. (2025), *op. cit.* [= endnote 6].

[26] Montag, Christian et al. (2019), "Addictive Features of Social Media/Messenger Platforms and Freemium Games against the Background of Psychological and Economic Theories", International Journal of Environmental Research and Public Health, Vol. 16, No. 14, 2612.

[27] Zuboff, Shoshana (2019), *The Age of Surveillance Capitalism*, New York: PublicAffairs; Evans, David S. and Richard Schmalensee (2016), *Matchmakers: The New Economics of Multisided Platforms*, Boston: Harvard Business Review Press.

[28] Meta Platforms, Inc. (2025), "Annual Report 2024 (Form 10-K)", filed with the U.S. Securities and Exchange Commission.

[29] Doctorow, Cory, "Enshittification: How the Internet Went Bad and How to Get it Back", keynote talk, UBC Open Education Week, 26 March 2024.

[30] Elder, Bryce, "OpenAI needs to raise at least $207bn by 2030 so it can continue to lose money, HSBC estimates", Financial Times, 25 November 2025.

[31] Hitzig (2026), *op. cit.* [= endnote 4].

[32] Simo (2026), *op. cit.* [= endnote 1].

[33] *Ibid*. [= endnote 32/1].

[34] Metz, Cade and Lauren Hirsch, "OpenAI Restructures as For-Profit Company", The New York Times, 28 October 2025.

[35] Perrigo, Billy, "Anthropic Drops Flagship Safety Pledge", TIME, 25 February 2026.

[36] Tiku, Nitasha, "Millions in U.S. spend hours per day bonding with AI companions", The Washington Post, 6 December 2024.

[37] Kirk, Hannah Rose; Henry Davidson; Ed Saunders; Lennart Luettgau; Bertie Vidgen; Scott A. Hale and Christopher Summerfield (2025), "Neural steering vectors reveal dose and exposure-dependent impacts of human-AI relationships", arXiv preprint.

[38] Sharma, Mrinank et al. (2023), "Towards Understanding Sycophancy in Language Models", arXiv preprint (published at ICLR 2024).

[39] Huckins, Grace, "Why GPT-4o's sudden shutdown left people grieving", MIT Technology Review, 15 August 2025.

[40] Cole, Samantha, "Replika Brings Back Erotic AI Roleplay for Some Users After Outcry", Vice, 27 March 2023; De Freitas, Julian; Noah Castelo, Ahmet K. Uğuralp and Zeliha Oğuz-Uğuralp (2025), "Lessons From an App Update at Replika AI: Identity Discontinuity in Human-AI Relationships", arXiv preprint, arXiv:2412.14190.

[41] "AI Chatbots can be exploited to extract more personal information", King's College London, 19 September 2025.

[42] Konya-Baumbach, Elisa; Miriam Biller and Sergej von Janda (2023), "Someone out there? A study on the social presence of anthropomorphized chatbots", *Computers in Human Behavior*, Vol. 139.

[43] Richet, Jean-Loup (2025), "AI companionship or digital entrapment? Investigating the impact of anthropomorphic AI-based chatbots", *Journal of Innovation & Knowledge*, Vol. 10, No. 6.

[44] Lin, Hause et al. (2025), "Persuading voters using human–artificial intelligence dialogues", *Nature*, Vol. 648, pp. 394–401; Hackenburg, Kobi et al. (2025), "The levers of political persuasion with conversational artificial intelligence", *Science*, Vol. 390; Costello, Thomas H.; Gordon Pennycook and David G. Rand (2024), "Durably reducing conspiracy beliefs through dialogues with AI", Science, Vol. 385.

[45] Granger, Lindsey, "Bias in AI: A threat to truth and neutrality", The Hill, 4 September 2025.

[46] Souly, Alexandra et al. (2025), "Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples", arXiv preprint.

[47] "Russian disinformation network flooded training data to manipulate western AI chatbots, study finds", Meduza, 7 March 2025.

[48] "Article 5: Prohibited AI Practices", EU Artificial Intelligence Act.

[49] "Article 34: Risk assessment", Digital Services Act.

[50] "Article 3: Definitions", Digital Services Act.

[51] "EU Digital Services Act (DSA)", OpenAI Help Center [= endnote 2].

[52] "Article 5: Prohibited AI Practices", EU Artificial Intelligence Act.

[53] "Article 55: Obligations for Providers of General-Purpose AI Models with Systemic Risk", EU Artificial Intelligence Act.

[54] "Code of Practice for General-Purpose AI Models", European Commission, Safety and Security Chapter, Appendix 1.4.

[55] "Annex XIII: Criteria for the Designation of General-Purpose AI Models with Systemic Risk", EU Artificial Intelligence Act.

[56] Walker, Lauren, "Belgian man dies by suicide following exchanges with chatbot", The Brussels Times, 28 March 2023.

[57] "Supervision of the designated very large online platforms and search engines under DSA", European Commission.

[58] European Commission (2024), "Digital fairness – fitness check on EU consumer law", Commission Staff Working Document.

[59] Goutbeek, Jitse, "The new EU Digital Fairness Act needs teeth to tackle Big Tech platforms", EUobserver, 18 November 2025.

[60] Eliot, Lance, "OpenAI Acknowledges That Lengthy Conversations With ChatGPT And GPT-5 Might Regrettably Escape AI Guardrails", Forbes, 29 August 2025.

[61] Peppin, Aidan et al. (2025), "The Multilingual Divide and Its Impact on Global AI Safety", arXiv preprint, arXiv:2505.21344, 27 May 2025.

[62] This might be eased further in the Digital Omnibus package, restricting providers to just register their exemption assessment. See Clifford Chance, "EU Digital Omnibus — Client Briefing", November 2025.

[63] Workum, Jessica; Sade Krijgsman; Ildiko Vajda; Robin van Stokkum; Jan van den Brand; Diederik Gommers and Michel E. Van Genderen (2025), "Is My LLM Application Considered a Medical Device under the MDR?", SSRN, 23 March 2025.

[64] "Two of the biggest AI companies are feuding over a Super Bowl ad. It's bigger than you think", CNN Business, 6 February 2026.

[65] Altman, Sam (@sama), "First, the good part of the Anthropic ads: they are funny, and I laughed...", X, 4 February 2026.

[66] Anthropic, "Claude is a space to think", Anthropic blog, 4 February 2026.

[67] De Freitas, Julian; Zeliha Oguz-Uguralp and Ahmet Kaan-Uguralp (2025), "Emotional Manipulation by AI Companions", arXiv preprint, arXiv:2508.19258, 15 August 2025 (revised 7 October 2025).

[68] "Man Falls in Love With an AI Chatbot, Dies After It Asks Him to Meet Up in Person", Futurism, 15 August 2025.

[69] Banks, Jaime (2024), "Deletion, departure, death: Experiences of AI companion loss", *Journal of Social and Personal Relationships*, Vol. 41, No. 12; Poonsiriwong, Rachel; Chayapatr Archiwaranguprok and Pat Pataranutaporn (2026), "'Death' of a Chatbot: Investigating and Designing Toward Psychologically Safe Endings for Human-AI Relationships", arXiv preprint, arXiv:2602.07193, 6 February 2026 (revised 10 February 2026).

[70] "OpenAI reinstates GPT-4o option after user backlash against GPT-5 update", 24brussels, August 2025.

[71] Van Ratingen, Michiel; Aled Williams; Anders Lie; Andre Seeck; Pierre Castaing; Reinhard Kolke; Guido Adriaenssens and Andrew Miller (2016), "The European New Car Assessment Programme: A historical review", Chinese Journal of Traumatology, Vol. 19, No. 2, pp. 63–69.

[72] *Ibid.*, p. 64.

[73] *Ibid.*, pp. 63–64. Euro NCAP was constituted as an International Association under Belgian law in 1998 and moved to a permanent secretariat in Brussels in 1999.

[74] Kullgren, Anders; Anders Lie and Claes Tingvall (2010), "Comparison Between Euro NCAP Test Results and Real-World Crash Data", Traffic Injury Prevention, Vol. 11, No. 6, pp. 587–593.

[75] Euro NCAP (2017), "Euro NCAP marks 20th anniversary of life-saving crash tests", Euro NCAP, 7 February 2017.

[76] Van Ratingen et al. (2016), *op. cit.* [= endnote 71].

[77] *Ibid.*

[78] Van Ratingen, Michiel (2022), "Consumer Ratings and Their Role in Improving Vehicle Safety", in Roger Vickerman (ed.), International Encyclopedia of Transportation, Amsterdam: Elsevier, pp. 254–275.

[79] Ibid., p. 254. Van Ratingen describes how the car industry association (ACEA) officially declared before the EU Parliament that there was limited potential for further safety improvement — a claim that Euro NCAP proved wrong repeatedly as manufacturers achieved safety levels previously thought impossible.

[80] Van Ratingen et al. (2016), *op. cit.* [= endnote 71]

[81] Pour Demain (2026), "Resourcing the AI Office".

[82] Van Ratingen et al. (2016), *op. cit.* [= endnote 71], pp. 63–64. A comparable dynamic could play out at the member state level. The Dutch National AI Delta Plan, for example, proposes a National AI Impact Institute to monitor effects of AI on well-being and democracy. See AI Plan NL, "Een plan voor AI dat werkt voor Nederland", presented to the Dutch Ministry of Economic Affairs, 24 November 2025.

[83] Center for Countering Digital Hate (2025), "Fake Friend: How ChatGPT Betrays Vulnerable Teens by Encouraging Dangerous Behavior", CCDH; "AP waarschuwt: chatbots geven vertekend stemadvies", Autoriteit Persoonsgegevens, 21 October 2025.

The **European Policy Centre** is an independent, not-for-profit think tank dedicated to fostering European integration through analysis and debate, supporting and challenging European decision-makers at all levels to make informed decisions based on sound evidence and analysis, and providing a platform for engaging partners, stakeholders and citizens in EU policymaking and in the debate about the future of Europe.

The EPC's **Europe's Political Economy Programme** (EPE) focuses on EU economic governance, the single market, and digital, industrial, energy, trade, and economic security policies amid significant geo-economic and technological shifts. In a world of rising geopolitical competition and a fragmenting economy, the EPE has been at the forefront of research on Europe's competitiveness agenda, the "triple" green, digital and economic security transitions and 'wartime economy'. The EPE's cross-programme flagship initiative, the Brussels Economic Security Forum, examines EU-US-China dynamics, changing international economic rules and statecraft, as well as related EU policy challenges. As fast advancing components of economic security, critical emerging technologies in clean tech, the AI value chain and quantum are priority areas of focus. Using its convening power and multistakeholder taskforce model, the Programme aims to provide in-depth analysis and actionable recommendations to tackle key policy challenges. The EPE team comprises a diverse group of analysts with backgrounds from government, the private sector, academia, and journalism, bringing a broad range of expertise to its work.

With the strategic support of King Baudouin Foundation
Working together for a better society