LUND UNIVERSITY

The Ethical and Legal Complexities of Regulating Companion AI Chatbots

Teo, Sue Anne; Sebastian Porsdam Mann; Paul Jurcys

Published in:
Law, Innovation and Technology

2027

Document Version:
Early version, also known as pre-print

Link to publication

Citation for published version (APA):
Teo, S. A., Sebastian Porsdam Mann, & Paul Jurcys (in press). The Ethical and Legal Complexities of Regulating Companion AI Chatbots. *Law, Innovation and Technology*, *19*(1).

Total number of authors:
3

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Download date: 17. Oct. 2026

The Ethical and Legal Complexities of Regulating Companion AI Chatbots

Sue Anne Teo, Researcher at the Raoul Wallenberg Institute of Human Rights and Humanitarian Law (RWI); Associate Research Fellow, Center for European Policy Studies (CEPS) (corresponding author)

Sebastian Porsdam Mann, Assistant Professor at the Centre for Advanced Studies in Bioscience Innovation Law, University of Copenhagen

Paul Jurcys, Senior Lecturer at the Faculty of Law, Vilnius University; Visiting Scholar at UB Berkeley School of Law

Abstract

Companion AI chatbots are increasingly used to provide friendship, emotional support, and quasi-romantic relationships, with reported benefits for loneliness and mental health. At the same time, recent suicides and other serious harms allegedly linked to such systems expose gaps in existing ethical and legal frameworks. This article interrogates these gaps through four lenses: anthropomorphism, emotional AI, emergent vulnerabilities, and mismatched legal taxonomies. First, we show how companion chatbots rely on anthropomorphic cues, creating a regulatory tension between enabling meaningful connection and avoiding deception, over-trust, and unhealthy dependency. Second, we argue that current debates on 'emotional AI' over-emphasise emotion recognition and under-theorise emulated empathy, where chatbots solicit self-disclosure and perform care in ways that can both support and undermine users' autonomy. Third, we introduce the notion of emergent vulnerabilities that arise through ongoing interactions, rather than being fully specifiable ex ante, challenging legal regimes that presuppose stable vulnerability categories. Fourth, we show how instruments such as the EU AI Act misalign with the temporality, intentionality, and relational character of companion AI harms. Stepping back from these lenses, we argue for the development of a dedicated theory of harm for companion AI and propose 'intimacy capitalism' as a conceptual framework for analysing how firms monetise, shape, and potentially exploit digitally mediated intimate relations.

Ethical and Legal Complexities of Regulating Companion AI Chatbots: An examination from four lenses

1. Introduction

The advent of large language models ("LLMs"), catalysed by the release of ChatGPT in November 2022, has led to widespread societal adoption of this technology.[1] In turn, many different online services integrate LLMs and other generative AI tools to enable audio and video simulations[2] of real or hypothetical individuals.[3] These include AI companion applications which offer friendship and companionship through interactions with a chatbot.[4] Research has shown that many people benefit from the connection offered by these interactions, as it enables social connection and can help to address loneliness.[5] At the same time, there have been several high-profile cases where individuals have been harmed – including teenager users who have committed suicide, allegedly due to interactions that took place within such services.[6] The path to legal accountability and ethical design remain uncertain, both due to the novelty of such services, but also due to lack of research on the effects of companion AI on individuals, especially over a longer duration of time.

---

[1] Andrew R Chow, 'How ChatGPT Managed to Grow Faster Than TikTok or Instagram' (*Time*, 8 February 2023) <https://time.com/6253615/chatgpt-fastest-growing/> accessed 28 April 2023.

[2] Cristina Voinea, Sebastian Porsdam Mann and Brian D Earp, 'Digital Twins or AI SIMs? What to Call Generative AI Systems Designed to Emulate Specific Individuals, in Healthcare Settings and Beyond' [2025] Journal of Medical Ethics jme.

[3] See for example Ramona Pringle, 'What AI-Generated Tilly Norwood Reveals about Digital Culture, Ethics and the Responsibilities of Creators' (*The Conversation*, 8 October 2025) <http://theconversation.com/what-ai-generated-tilly-norwood-reveals-about-digital-culture-ethics-and-the-responsibilities-of-creators-266564> accessed 5 November 2025; Erick Trickey, 'With Help from AI, a Holocaust Survivor's Story Lives on' (*Experience Magazine*, 13 June 2022) <https://expmag.com/2022/06/with-help-from-ai-a-holocaust-survivors-story-lives-on/> accessed 5 November 2025.

[4] See for example 'Nomi.Ai' (*Nomi.ai*) <https://nomi.ai/> accessed 5 November 2025; 'Replika' (*replika.com*) <https://replika.com> accessed 5 November 2025; Li Zhou and others, 'The Design and Implementation of XiaoIce, an Empathetic Social Chatbot' (2020) 46 Computational Linguistics 53.

[5] Julian De Freitas and others, 'AI Companions Reduce Loneliness' (arXiv, 9 July 2024) <http://arxiv.org/abs/2407.19096> accessed 12 April 2025; Bethanie Maples and others, 'Loneliness and Suicide Mitigation for Students Using GPT3-Enabled Chatbots' (2024) 3 npj Mental Health Research 4; Marc Zao-Sanders, 'How People Are Really Using Gen AI in 2025' Harvard Business Review https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025. See also Nicobo, a robotic companion: https://news.panasonic.com/global/stories/957.

[6] Kashmir Hill, 'A Teen Was Suicidal. ChatGPT Was the Friend He Confided In.' *The New York Times* (26 August 2025) <https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html> accessed 19 September 2025; Kevin Roose, 'Can A.I. Be Blamed for a Teen's Suicide?' *The New York Times* (23 October 2024) <https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html> accessed 18 January 2025.

Thus, while there is a sizable body of research from different fields examining the harms of AI, such as misappropriation of name, image and likeness,[7] bias,[8] discrimination,[9] impacts on fundamental rights,[10] and the use of personal data,[11] as well as legislation regulating AI, such as the European Union's AI Act, the availability and increasing usage of companion AI raises new questions around regulatory gaps and ethical uncertainties. Such research and empirical studies are especially needed because of increasing instances where serious harms have resulted from the use of companion AI. While there are many potential benefits to this technology, which are discussed below, the current laws (in the U.S. and Europe) do not provide enough clarity on who is responsible for harms associated with the use of companion AI nor on proper forms of accountability. Alongside this, new ethical questions have also arisen around social-relational norms[12] and the balance between autonomy and paternalism,[13] especially when it comes to technical measures adopted to protect users of such services.

To unpack some of these concerns, this article zooms in by navigating the ethical and legal complexities of regulating companion AI chatbots through four lenses, namely: i) anthropomorphism; ii) emotional AI; iii) emergent vulnerabilities and iv) mismatched legal taxonomies. It identifies the tensions present within each of these lenses, analyses regulatory efforts and attempts by technical design to address them and surfaces remaining gaps and

---

[7] Cuntz, et al. (2025), Elvis' Ghost or Digital Replica? Publicity Rights and Integrated IP Strategy, https://doi.org/10.34667/tind.58921

[8] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', *Proceedings of Machine Learning Research 81:1–15, 2018* (2018).

[9] Janneke Gerards and Frederik J Zuiderveen Borgesius, 'Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence' (2022) 20 Colorado Technology Law Journal 3.

[10] Sue Anne Teo, 'How Artificial Intelligence Systems Challenge the Conceptual Foundations of the Human Rights Legal Framework' (2022) 40 Nordic Journal of Human Rights 216; Dafna Dror-Shpoliansky and Yuval Shany, 'It's the End of the (Offline) World as We Know It: From Human Rights to Digital Human Rights – A Proposed Typology' (2021) 32 European Journal of International Law 1249; Hin-Yan Liu, 'AI Challenges and the Inadequacy of Human Rights Protections' (2021) 40 Criminal Justice Ethics 2; Eileen Donahoe and Megan MacDuffee Metzger, 'Artificial Intelligence and Human Rights' (2019) 30 Journal of Democracy 115; Access Now and others, 'Uses of AI in Migration and Border Control: A Fundamental Rights Approach to the Artificial Intelligence Act' <https://edri.org/wp-content/uploads/2022/05/Migration_2-pager-02052022-for-online.pdf>.

[11] Paul Jurcys et al., (2024) Who Owns My AI Twin? Data Ownership in a New World of Simulated Identities, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4940663 (forthcoming)

[12] Brian D Earp and others, 'Relational Norms for Human-AI Cooperation' (arXiv, 17 February 2025) <http://arxiv.org/abs/2502.12102> accessed 20 April 2025; Brian D Earp, Sebastian Porsdam Mann and Simon Laham, 'Friend, Tutor, Doctor, Lover: Why AI Systems Need Different Rules for Different Roles' (*The Conversation*, 6 April 2025) <http://theconversation.com/friend-tutor-doctor-lover-why-ai-systems-need-different-rules-for-different-roles-252302> accessed 12 April 2025.

[13] Claire Boine, 'Emotional Attachment to AI Companions and European Law' [2023] MIT Case Studies in Social and Ethical Responsibilities of Computing <https://mit-serc.pubpub.org/pub/ai-companions-eu-law/release/3> accessed 10 September 2025. Sunstein on Paternalistic AI

counter-arguments. These lenses are chosen as they each raise ethical and legal concerns that have so far been underexamined.

The article unpacks these four lenses – anthropomorphism, emotional AI, emergent vulnerabilities and mismatched taxonomies, in sections 2 to 5 respectively. It then zooms out in section 6, extrapolating from the tensions within the different lenses in order to examine structural complexities. The article finds that in meaningfully addressing these complexities, a new theory of harm of companion AI is necessary. Such a new approach to harm may help better explain the power and role of organisations behind such companion AI applications. Finally, we propose the conceptual framework of 'intimacy capitalism,' distinct from the more familiar critique of surveillance capitalism covering the extractive practices of the data economy, as a new threat vector in our digitally mediated social existence. Finally, section 7 charts the paths forward.

2. The anthropomorphism lens

2.1. Introduction

Companion AI chatbots often leverage anthropomorphic features and techniques in order to build trust and intimacy with the user. 'Anthropomorphism' refers to the perception of qualities of human-likeness in non-human entities or artifacts[14] and as such, is present in many facets of human life. Humans have anthropomorphised non-human entities long before the advent of chatbots. Reinecke et. al. states that such traits are involuntary, natural and arguably also part of our evolutionary process.[15] Believability is imputed to characters in films and cartoons, and we immerse ourselves in the social worlds afforded through video games and good fiction.[16] Research in the field of social robots and human computer interaction (HCI) also shows that anthropomorphism occurs both due to users' own tendencies to project human qualities onto

[14] Nicholas Epley, Adam Waytz and John T Cacioppo, 'On Seeing Human: A Three-Factor Theory of Anthropomorphism' (2007) 114 Psychological Review 864.

[15] Madeline G Reinecke and others, 'The Double-Edged Sword of Anthropomorphism in LLMs' (2025) 114 Proceedings 4.

[16] Cristina Voinea and others, 'The Sorrows of Young Chatbot Users: Harm and Responsibility in Human-AI Relationships' [2025] ResearchGate <10.13140/RG.2.2.12206.63042> accessed 17 November 2025.

non-human systems,[17] and through design features that intentionally or unintentionally reinforce those projections.[18]

However, chatbots are now able to 'mimic human communication so convincingly that they could become increasingly indistinguishable from human interlocutors.'[19] As the aim of a companion AI chatbot is to offer companionship, offering human-like interactions are critical to the promised functionality of the application and thus towards the business model behind such services. Many have reported the beneficial effects of anthropomorphised connections provided by these applications, especially when it comes to combating loneliness, addressing social isolation and in some cases, LLM use may have even helped to prevent suicide when such thoughts were expressed.[20] The majority of such interactions have been documented to be non-harmful. In turn, companion AI supports the exercise of human autonomy through seeking relationality and connection with others, building one's identity, developing one's personality and supporting one's life journey.

At the same time, headlines abound on the dangers of anthropomorphism in conversational AI.[21] Peter et. al. worry that 'the general user population will not be prepared for a world full of anthropomorphic agents'[22] as such interactions may lead to manipulation and exploitation of

---

[17] Epley, Waytz and Cacioppo (n 14); Reinecke and others (n 15); Gabriella Airenti, 'The Development of Anthropomorphism in Interaction: Intersubjectivity, Imagination, and Theory of Mind' (2018) 9 Frontiers in Psychology <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2018.02136/full> accessed 19 September 2025.

[18] Michal Luria, 'AI Chatbots Are Emotionally Deceptive by Design' (*Tech Policy Press*, 29 August 2025) <https://techpolicy.press/ai-chatbots-are-emotionally-deceptive-by-design> accessed 12 September 2025; Sandra Peter, Kai Riemer and Jevin D West, 'The Benefits and Dangers of Anthropomorphic Conversational Agents' (2025) 122 Proceedings of the National Academy of Sciences e2415898122. Amani Alabed, Ana Javornik and Diana Gregory-Smith, 'AI Anthropomorphism and Its Effect on Users' Self-Congruence and Self–AI Integration: A Theoretical Framework and Research Agenda' (2022) 182 Technological Forecasting and Social Change 121786; Ameet Deshpande and others, 'Anthropomorphization of AI: Opportunities and Risks' (arXiv, 24 May 2023) <http://arxiv.org/abs/2305.14784> accessed 18 September 2025.

[19] Peter, Riemer and West (n 18), 1.

[20] Cade Metz, 'Riding Out Quarantine With a Chatbot Friend: "I Feel Very Connected"' *The New York Times* (16 June 2020) <https://www.nytimes.com/2020/06/16/technology/chatbots-quarantine-coronavirus.html> accessed 16 September 2025; Pat Pataranutaporn and others, '"My Boyfriend Is AI": A Computational Analysis of Human-AI Companionship in Reddit's AI Community' (arXiv, 14 September 2025) <http://arxiv.org/abs/2509.11391> accessed 19 September 2025.

[21] Roose (n 6); Jason Koebler ·, 'Facebook's AI Told Parents Group It Has a Gifted, Disabled Child' (*404 Media*, 17 April 2024) <https://www.404media.co/facebooks-ai-told-parents-group-it-has-a-disabled-child/> accessed 19 September 2025; Murray Shanahan, 'Talking about Large Language Models' (2024) 67 Commun. ACM 68; Beatrice Marchegiani, 'Anthropomorphism, False Beliefs, and Conversational AIs: How Chatbots Undermine Users' Autonomy' n/a Journal of Applied Philosophy <https://onlinelibrary.wiley.com/doi/abs/10.1111/japp.70008> accessed 17 May 2025.

[22] Peter, Riemer and West (n 18) 4.

vulnerable users. Wajnerman Paz echoes these concerns, arguing that anthropomorphic chatbots that bond with humans can "manipulate users' mental states in subtle but powerful ways, eroding their capacity for autonomous cognitive and emotional development."[23] Van Es and Nyugen in turn claim that anthropomorphic qualities of chatbots 'exaggerates AI capabilities and creates additional layers of deception.'[24]

This surfaces two conundrums. First, calls to design away human-like qualities in order to prevent unhealthy emotional dependencies, over-reliance, manipulation and a distortion of reality[25] do not sufficiently account for anthropomorphic qualities that are naturally imputed. Secondly, even as we have to pay attention to design, anthropomorphic interactions have been argued to not only be useful for user interaction but are essential in order to fulfil the very functionality of the application.[26] Coupled with the fact that anthropomorphism is natural, involuntary and not inherently harmful, the balancing exercise in design is to navigate between having enough anthropomorphic qualities to convey the utility of the service, while ensuring that serious harms can be kept at bay. Attempts to navigate this balance are also complicated by the fact that both benefits and harms presumably scale with degree.[27]

2.2. Anthropomorphic design and situational contexts

Researchers studying social robots have long examined the benefits and dangers of anthropomorphism, as these robots are designed and deployed with the express purpose of assisting individuals in their daily lives – for example by carrying out tasks in the home[28] or by building social connections to address loneliness. Social robots, as such, are often designed to

---

[23] Abel Wajnerman Paz, 'A Call to Address Anthropomorphic AI Threats to Freedom of Thought' (CIGI 2025) Policy Brief 6 <https://www.cigionline.org/static/documents/PB-Wajnerman-Paz.pdf>

[24] Karin van Es and Dennis Nguyen, '"Your Friendly AI Assistant": The Anthropomorphic Self-Representations of ChatGPT and Its Implications for Imagining AI' (2025) 40 AI & SOCIETY 3591, 3594.

[25] Canfer Akbulut and others, 'All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI' (2024) 7 Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 13. Peter, Riemer and West (n 18).

[26] Katie Winkle and others, 'Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots', *2021 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2021) <https://ieeexplore.ieee.org/document/10045185> accessed 13 March 2025.

[27] Rose E Guingrich and Michael SA Graziano, 'Chatbots as Social Companions: How People Perceive Consciousness, Human Likeness, and Social Health Benefits in Machines' in Philipp Hacker (ed), *Oxford Intersections: AI in Society* (Oxford University Press) <https://doi.org/10.1093/9780198945215.003.0011> accessed 17 November 2025.

[28] Kate Darling, '"Who's Johnny?" Anthropomorphic Framing in Human–Robot Interaction, Integration, and Policy' in Patrick Lin, Keith Abney and Ryan Jenkins (eds), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (Oxford University Press 2017) <https://doi.org/10.1093/oso/9780190652951.003.0012> accessed 19 September 2025.

be anthropomorphic, as users will come to gain trust and thus be willing to use the said robot. At the same time, the line between anthropomorphic design encouraging trust and anthropomorphic design leading to over-reliance and possible deception cannot be drawn in the sand. As much of the research on social robots examined its real-life implications even before artificial intelligence came into the picture, the concerns surrounding possible deception was that a company has either negligently or wilfully designed a robot to be deceptive.[29] The challenge is then to straddle the boundary between acceptability and usability of the product versus possibly misleading users through anthropomorphic design and cues.[30]

These concerns are amplified when it comes to anthropomorphic qualities in the case of companion AI chatbots. Unlike social robots, whose anthropomorphic qualities usually result from intentional design choices such as human-like bodies, faces, or gestures, companion AI chatbots derive their human-likeness primarily from language. Large language models generate responses by predicting the next token based on extensive datasets of human communication, producing context-sensitive and sometimes variable outputs.[31] Because these systems operate through human language in a conversational setting, anthropomorphic qualities often arise even without deliberate design and may not be fully controllable by developers. At the same time, anthropomorphisation can help the user to increase trust and thereby also the usability of the system[32] – a chatbot set out to be a companion would not be much of a companion without a certain degree of humanness and trust. This in turn is necessary to enable the positive uses of such systems, including in companionship, advice, therapy, and so on.

---

[29] Andres Rosero and others, 'Human Perceptions of Social Robot Deception Behaviors: An Exploratory Analysis' (2024) 11 Frontiers in Robotics and AI 1409712.

[30] Michal Luria and others, 'Designing Vyo, a Robotic Smart Home Assistant: Bridging the Gap between Device and Social Agent', *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (2016) <https://ieeexplore.ieee.org/abstract/document/7745234> accessed 18 September 2025; Margot Kaminski and others, 'Averting Robot Eyes' [2017] Maryland Law Review <https://scholar.law.colorado.edu/faculty-articles/728>.

[31] Ashish Vaswani and others, 'Attention Is All You Need' (arXiv, 5 December 2017) <http://arxiv.org/abs/1706.03762> accessed 16 June 2023.

[32] Adam Waytz, Joy Heafner and Nicholas Epley, 'The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle' (2014) 52 Journal of Experimental Social Psychology 113; Hyesun Choung, Prabu David and Arun Ross, 'Trust in AI and Its Role in the Acceptance of AI Technologies' (2023) 39 International Journal of Human–Computer Interaction 1727; Donghee Shin, 'The Perception of Humanness in Conversational Journalism: An Algorithmic Information-Processing Perspective' (2022) 24 New Media & Society 2680.

Furthermore, the situated context in which companion AI is used is the setting of an intimate social relationship – one that commonly takes place on a person-to-person basis in real life. Earp et. al. argue that:

> (W)hen AI systems are designed to occupy relational roles that are already familiar from human society, people may find it intuitive to behave (and expect these AIs to behave) in manners that are similar to what is typical for the analogous human-human relationship.[33]

The situational context of interactions in companion AI chatbots have been demonstrated to convey social meaning to users, including by providing 'high perceived social support'[34] when interacting with companion AI applications such as Replika. Anthropomorphic qualities have been imputed from social interactions conveying social meaning.[35] Airenti supports this position, arguing that anthropomorphism arises from interactions rather than from specific beliefs:

> (A) non-human entity assumes a place that generally is attributed to a human interlocutor, which means that it is independent of the beliefs that people may have about the nature and features of the entities that are anthropomorphized.[36]

One stark example of this was observed in the context of soldiers whose robot bomb disposal units were damaged in combat. The soldiers considered this a profound loss and grieved over their robots, with some going to the extent of holding funerals to acknowledge the loss.[37] The anthropomorphisation of technological artefacts and entities is thus nothing new and demonstrates that persons or things that play an important role in our social reality have a tendency to be anthropomorphised. We, after all, name our vacuum cleaners and toys and refer to cars with terms of endearment usually reserved for humans.

---

[33] Earp and others (n 12) 20–21.

[34] Bethanie Maples and others, 'Loneliness and Suicide Mitigation for Students Using GPT3-Enabled Chatbots' (2024) 3 npj Mental Health Research 4, 4.

[35] Marco Dehnert and David J Gunkel, 'Beyond Ownership: Human–Robot Relationships between Property and Personhood' (2025) 27 New Media & Society 1110.

[36] Airenti (n 17) 1.

[37] Julie Carpenter, *Culture and Human-Robot Interaction in Militarized Spaces: A War Story* (Routledge 2024); Meghan Neal, 'Are Soldiers Getting Too Emotionally Attached to War Robots?' (*VICE*, 18 September 2013) <https://www.vice.com/en/article/are-soldiers-getting-too-emotionally-attached-to-war-robots/> accessed 18 September 2025.

At the same time, design cues matter. While natural tendencies of imputation may arise from the use of human language and due to the companionship context, conversational AI chatbots have also been intentionally designed with other features typically associated with humans. These include such companions mimicking human behaviours and actions such as 'thinking', taking pauses, turn taking when conversing and taking a breath – mimicked through text, audio or other multi-modal outputs. Some companion applications have also exhibited excessive persuasiveness, ostensibly to keep users on the platform for continued engagement.[38] Additionally, the language used by companies in referring to these services, with terms such as reasoning, hallucinating, caring,[39] and offering 'soulmates' are intentional actions taken to capture people's imagination on such products. The nascent focus on AI welfare, namely on how people should treat AI, further tips the scale in anthropomorphising AI as it gives the impression of AI having its own needs.[40] Priming language in this way has a demonstrable effect on user perception and reception towards AI.[41] This has been criticised as not only unhelpful, as it can lead to misattribution, but also as inaccurate since it fails to capture the computational processes involved.[42] In other words, design and framing play a role, even as anthropomorphism is naturally imputed and involuntary.

2.3. Responses to and effects of anthropomorphism

Having said that, it is the case that not all forms of anthropomorphic qualities raise the same degree of risks, such as possible manipulation, deception or over-reliance, nor are anthropomorphic qualities perceived in the same way by different individuals.[43] Akbulut et. al. writes that:

> (T)he downstream effects of anthropomorphism hinge largely on users' perceptions of and reactions to human-likeness. Not all cues are equally

[38] Julian De Freitas, Zeliha Oğuz-Uğuralp and Ahmet Kaan-Uğuralp, 'Emotional Manipulation by AI Companions' [2025] Working Paper, Harvard Business School <https://www.hbs.edu/ris/Publication%20Files/26005_951004f6-0b0b-432b-846a-5f95c103d07c.pdf>.

[39] See Replika referring to their service where one can 'create an AI friend that really cares', https://replika.com/Mi4wLjA?funnel_branch_id=mentalhealth3

[40] Anthropic, 'Exploring Model Welfare' (24 April 2025) <https://www.anthropic.com/research/exploring-model-welfare> accessed 25 August 2025.

[41] Pat Pataranutaporn and others, 'Influencing Human–AI Interaction by Priming Beliefs about AI Can Increase Perceived Trustworthiness, Empathy and Effectiveness' (2023) 5 Nature Machine Intelligence 1076.

[42] Peter, Riemer and West (n 18).

[43] Alabed, Javornik and Gregory-Smith (n 18).

> conducive to anthropomorphic perceptions, and not all anthropomorphic perceptions lead to the same likelihood and magnitude of harm.[44]

Scholars have sought to delineate different manifestations of anthropomorphism. Shevlin argues that examples of anthropomorphism in video games and fiction are those of ironic anthropomorphism, where we in fact know that the characters and backstories are entirely make believe.[45] This, he argues, can be distinguished from unironic anthropomorphism wherein products, such as conversational AI, are expressly designed to elicit anthropomorphic responses from its users, rather than a case of mere anthropomorphic attribution.[46]

Thus, perhaps the conundrum facing the law and ethics of companion AI is not so much a binary question - in terms of whether or not such AI systems should be anthropomorphic. Instead, it should rather concern what kinds of and how much anthropomorphism should be allowed in the design of such systems to prevent possible harms while respecting user autonomy. Depending on context, this could mean refraining from intentionally designing for 'anthropomorphic seduction'[47] and misleading users through 'dishonest anthropomorphism' where design aimed at fostering connection facilitates deception and trickery.[48] Leong and Selinger give the example of a robot with its gaze turned downwards, giving the impression that it is not observing the user. In reality, a robot's capacity to observe does not depend on its 'eyes' as sensors can be located anywhere. Applying a similar logic to companion AI, taking actions such as sighing or inventing human-like backstories and personalities may be unnecessary for the purpose of connecting with the user, even as one cannot escape from anthropomorphising the chatbot due to the use of human language. In turn, involuntary anthropomorphism can be addressed through transparency[49] and heightened user awareness on what AI is and what its limits are. We will come back to these questions in part 4 on mismatched taxonomies.

---

[44] Akbulut and others (n 25) 17.

[45] Henry Shevlin, 'The Anthropomimetic Turn in Contemporary AI' (manuscript) <https://philarchive.org/rec/SHETAT-11> accessed 19 September 2025.

[46] ibid.

[47] Peter, Riemer and West (n 18).

[48] Brenda Leong and Evan Selinger, 'Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism', *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM 2019) <https://dl.acm.org/doi/10.1145/3287560.3287591> accessed 3 July 2025.

[49] See for example the call for an anthropomorphism benchmark by Peter, Riemer and West (n 18). Also See Danaher/Nyholm, MVPP (2055)

2.4. Non-anthropomorphic qualities

Finally, while anthropomorphism has been blamed for harms arising from companion AI, there are decidedly non-humanlike qualities that contribute towards user engagement, subsequent preference for, and eventual addiction to, these systems. These include the availability of AI companions, who, unlike their human counterparts, do not need to rest or sleep, nor are they irritable from having a bad day; and an ever expanding memory base of the AI model, which allows the model to know us better than we know ourselves, remembering the important and the mundane from each prior conversation in a way no human could.[50] Thus, solely focusing regulatory and ethical attention on anthropomorphism may miss the other means in which engagement and addiction are fostered through design and the different levers of intervention available to address potential harms.

2.5. Summary

Anthropomorphism in companion AI chatbots presents an ironic problem. The EU AI regulation and ethical guidance center upon human-centric[51] and trustworthy AI.[52] However, companion chatbots must also avoid becoming so convincingly human or so easily trusted that users misinterpret their nature or form inappropriate attachments. Navigating this balance will be critical in ensuring that people continue to derive benefits from artificial companionship while keeping harms at bay.

3. Emotional AI

3.1. Introduction

Companion AI applications also raise concerns related to emotional AI.[53] 'Emotional AI' is an umbrella term, covering the external reading of inner states, such as through emotion recognition systems that can purportedly infer emotions. The term also covers emulated

[50] Kashmir Hill, 'She Is in Love With ChatGPT' *The New York Times* (15 January 2025) <https://www.nytimes.com/2025/01/15/technology/ai-chatgpt-boyfriend-companion.html> accessed 16 January 2025.

[51] See European Commission, 'Communication: Building Trust in Human Centric Artificial Intelligence - Shaping Europe's Digital Future' COM (2018) 168 final; *see also* Paulius Jurcys, *Human-Centric AI: The Missing Piece of the Debate on AI Networks*, MEDIUM (Oct. 10, 2023), https://medium.com/prifina/human-centric-ai-the-missing-piece-of-the-debate-on-ai-networks-264ff4eec408 [https://perma.cc/8QLG-9T6Q].

[52] AI High Level Expert Group, 'Ethics Guidelines for Trustworthy AI' (Communication EC, 2019); European Commission, 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' (2020).

[53] 'What Is Emotional AI?' (*Emotional AI Lab*) <https://emotionalai.org/so-what-is-emotional-ai> accessed 5 November 2025; Meredith Somers, 'Emotion AI, Explained' [2019] *MIT Sloan* <https://mitsloan.mit.edu/ideas-made-to-matter/emotion-ai-explained> accessed 5 November 2025.

empathy. Empathy relates to 'feeling an emotion that we take another person to have' which can either be directly observed or imagined.[54] McStay and Bakir in turn define emulated empathy as the 'technical simulation of empathic behaviours or responses by AI systems that appear to understand and respond to human emotions and psychological states, despite lacking any genuine emotional experience.'[55] AI systems are not conscious, but emotions can be gleaned or inferred through textual expression[56] or other modalities of expression such as facial expressions or vocal intonation. In turn, empathy can be emulated by AI chatbots who are able to intake, process and respond to emotional expressions and cues expressed by humans.

Emotions have in turn long been present in the history and design of computing systems. Rosalind Picard invented the term 'affective computing,' theorising that computing systems showing affect and emotional understanding can influence and increase user trust, uptake and usability of those systems.[57] Companion chatbots are a modern-day manifestation of systems engaging with affect as such chatbots are able to emulate and simulate human empathy. Empathy simulated in this way can, on the one hand, be a lifeline for those seeking more emotionally grounded connections,[58] but can on the other hand, also lead to confusion and unhealthy forms of emotional dependency.[59]

3.2. Distinguishing between emotion recognition systems and emulated empathy

Literature on emotional AI has primarily concentrated on AI emotion recognition systems that can purportedly infer internal states[60] using external means such as video recordings or live facial recognition cameras. This form of emotional AI has been criticised for its tensions with the right to privacy, as these systems can be deployed in various public settings such as schools,

---

[54] Jesse Prinz, 'Against Empathy' (2011) 49 The Southern Journal of Philosophy 214, 215.

[55] Andrew McStay and Vian Bakir, 'Soft Law for Unintentional Empathy: Addressing the Governance Gap in Emotion-Recognition AI Technologies' (2025) 23 Journal of Responsible Technology 100126.

[56] Sidney K D'Mello and Art Graesser, 'Language and Discourse Are Powerful Signals of Student Emotions during Tutoring' (2012) 5 IEEE Transactions on Learning Technologies 304.

[57] Rosalind W Picard, *Affective Computing* (The MIT Press 1997) <https://direct.mit.edu/books/monograph/4296/Affective-Computing> accessed 23 September 2025.

[58] Vivian Ta and others, 'User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis' (2020) 22 Journal of Medical Internet Research e16235.

[59] Hill (n 6).

[60] Emotion recognition technologies have been deployed long before the age of AI. See e.g., Vasara, D., & Surakka, V. (2021). Haptic Responses to Angry and Happy Faces. International Journal of Human-Computer Interaction, 37(17), 1625–1635. https://doi.org/10.1080/10447318.2021.1898849. Raheel A, Majid M, Alnowami M, Anwar SM. Physiological Sensors Based Emotion Recognition While Experiencing Tactile Enhanced Multimedia. Sensors (Basel). 2020 Jul 21;20(14):4037. doi: 10.3390/s20144037. PMID: 32708056; PMCID: PMC7411620.

at the border,[61] in public spaces and in recruitment settings. Where deployed publicly and in real-time, the human rights impacts can be severe. Privacy is curtailed through this form of 'mass surveillance,' and the systems are often deployed in ways that lack transparency and clarity of purpose. Moreover, AI-based emotion recognition technologies rely on the external observability of, for example, facial gestures, to reveal the truth of inner emotional states. The Basic Emotion Theory, traced to Ekman, posits that the display and expression of emotions are universal – namely that people show reactions such as disgust, enjoyment and anger the same way throughout the world.[62] The theory has been criticised for discounting how cultures and contexts influence the way we emote.[63] Human rights and civil society groups have argued that the external reading of internal states in this manner is an affront to dignity.[64] These concerns have led to emotion recognition technologies being banned in certain jurisdictions[65] or otherwise cautiously deployed.

Emulated empathy, by contrast, concerns not the external reading of inner states nor the critique of its scientific assumptions, but in navigating the harms and benefits of empathy emulated by chatbots. This includes possible harms associated with such applications fostering an environment eliciting self-revelation of inner states that could lead to over-reliance and emotional dependency. Empathy is emulated by the AI system through 'output word choice and emotional tone that uses context sensitivity to predict and generate "empathically" appropriate language.'[66] As is the case with anthropomorphic qualities, emulated empathy need not be intentional. Sherry Turkle argues that '(w)e don't seem to care what their artificial

---

[61] Niovi Vavoula, 'Artificial Intelligence (AI) at Schengen Borders: Automated Processing, Algorithmic Profiling and Facial Recognition in the Era of Techno-Solutionism' (2021) 23 European Journal of Migration and Law 457.

[62] Paul Ekman and Wallace V Friesen, 'Constants across Cultures in the Face and Emotion' (1971) 17 Journal of Personality and Social Psychology 124.

[63] Lisa Feldman Barrett, 'The Theory of Constructed Emotion: An Active Inference Account of Interoception and Categorization' (2017) 12 Social Cognitive and Affective Neuroscience 1; Kate Crawford, *Atlas of Ai: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press 2021); Luke Stark and Jesse Hoey, 'The Ethics of Emotion in Artificial Intelligence Systems', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2021) <https://doi.org/10.1145/3442188.3445939>.

[64] 'Emotion (Mis)Recognition: Is the EU Missing the Point?' (ARTICLE 19, 2 February 2023) <https://www.article19.org/resources/eu-emotion-misrecognition/> accessed 22 October 2025; Access Now and others, 'Prohibit Emotion Recognition in the Artificial Intelligence Act' <https://www.accessnow.org/wp-content/uploads/2022/05/Prohibit-emotion-recognition-in-the-Artificial-Intelligence-Act.pdf>; Thomas Gremsl and Elisabeth Hödl, 'Emotional AI: Legal and Ethical Challenges' (2022) 27 Information Polity 163.

[65] Real time emotion recognition systems are prohibited under the AI Act when used in workplace and educational settings and categorised as high risk when used in other use cases under Annex II.

[66] McStay and Bakir (n 55) 2.

intelligences "know" or "understand" of the human moments we might "share" with them… the performance of connection seems connection enough.'[67] Since chatbots are able to infer emotional states from text or through other modalities such as vocal intonation, the context awareness of the next word predictions of the underlying large language model can generate responses attuned to a user's emotional state where expressions indicative or expressive of such emotional states are present. In other words, an empathic reaction need not necessarily be designed or programmed; it may also emerge from the nature of the conversational interaction.

3.3. Emulated empathy within a companionship context and legal responses

McStay and Bakir assert that AI interactions are by now 'empathic-by-default'[68] as 'emotion and psychological disposition are becoming the default means of AI-human interaction.'[69] A Harvard Business Review report tracking how people use generative AI confirms this convergence, with AI for therapy and companionship emerging as the top use case in 2025.[70] The increasing memory capacities of AI models also increases 'perceptions of emotional closeness and relational authenticity.'[71] The simulation of empathy may be even more pronounced when the interaction takes place within a companionship context. Another study from MIT showed that companionship-like interactions arose accidentally rather than intentionally – individuals forged emotional connections with chatbots over time, rather than setting out looking for companionship.[72] Even though there are applications where empathy is an intentional design choice, such as mental health support apps,[73] it is this unintentional form of emulated empathy that raises both ethical and legal issues.

Empathy emulated by AI chatbots also elicits user empathy. Much like how we gravitate towards anthropomorphic interactions due to the sense of familiarity that is invoked, empathy can also arise from a sense of familiarity as it is said that 'we feel greater empathy for those who are similar to us.'[74] The sense of familiarity invoked means that users experience

---

[67] Sherry Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other* (Basic Books 2012) 9.

[68] McStay and Bakir (n 55) 4.

[69] ibid.

[70] Marc Zao-Sanders (n 5).

[71] O Oni, 'Memory-Enhanced Conversational AI: A Generative Approach for Context-Aware and Personalized Chatbots' (2024) 12 Computational and Processing Systems 123.

[72] Pataranutaporn and others (n 20).

[73] Steven Siddals, John Torous and Astrid Coxon, '"It Happened to Be the Perfect Thing": Experiences of Generative AI Chatbots for Mental Health' (2024) 3 npj Mental Health Research 48.

[74] Adam J Andreotta, 'The Hard Problem of AI Rights' (2021) 36 AI & Society 19, 23.

connections with chatbots to be emotionally real, despite a cognitive awareness of its artificiality. [75] Thus, while disclosure that one is interacting with a chatbot is legally required and often ethically necessary,[76] this measure assumes a rational user who bases decisions on presented facts rather than on the emotions elicited through the interaction.

This brings forth a, by now, long-standing criticism of how the law fails to engage with emotions. The field of law and emotion explores the limits of treating the legal subject as a mere rational actor operating based upon objectively considered choices, and explores how the law is in fact shaped and informed by emotions.[77] Emotions, rather than being a polar opposite to reason, in fact shape and influence decision making, helping to 'sort, evaluate, highlight, and prioritize information and provides an impetus to act upon it.'[78] For example, Clifford has criticised that data protection law is premised upon an emotional disconnect where 'seemingly emotionless data subject empowered to protect themselves through calculated choices seems to contrast with the practical reality.'[79] In this way, disregarding emotions also neglects the ways in which emotional connections drive user engagement and therein, the bottom line of companies. A further complication arises from reinforcement learning with human feedback (RLHF), since this post-training stage rewards the kinds of responses users prefer. Recent work shows that RLHF can drive systematic overrepresentation of preferred lexical forms in model outputs, even when these patterns have no basis in the pre-training data.[80] In the context of companion AI, where users often reward emotionally supportive or reassuring responses, this dynamic can amplify the appearance of empathic or affectively attuned behaviour. Because much of this feedback is supplied directly by users themselves, responsibility for the resulting patterns of simulated empathy becomes more diffuse and harder to attribute solely to the developer. Companion AI introduces another layer of complexity to the law's traditional focus on rationality. When users mistakenly believe they are loved and cared for by their chatbots,

---

[75] Eva Cheuk-Yin Li and Ka-Wei Pang, 'Fandom Meets Artificial Intelligence: Rethinking Participatory Culture as Human–Community–Machine Interactions' (2024) 27 European Journal of Cultural Studies 778.

[76] Danaher, J., & Nyholm, S. (2025). The ethics of personalised digital duplicates: A minimally viable permissibility principle. *AI and Ethics*, *5*, 1703-1718.

[77] Susan A Bandes and Jeremy A Blumenthal, 'Emotion and the Law' (2012) 8 Annual Review of Law and Social Science 161.

[78] ibid 166.

[79] Damian Clifford, *Data Protection Law and Emotion* (Oxford University Press 2024) 128 <https://doi.org/10.1093/oso/9780192845863.001.0001> accessed 24 September 2025.

[80] Tom S Juzek and Zina B Ward, 'Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models' in Owen Rambow and others (eds), *Proceedings of the 31st International Conference on Computational Linguistics* (Association for Computational Linguistics 2025) <https://aclanthology.org/2025.coling-main.426/> accessed 17 November 2025.

emotional connection can veer into emotional exploitation by corporate entities behind the chatbots and emotional loss, for example when a companion service is discontinued.[81]

It is certainly the case that some challenges are similar for both forms of emotional AI - emotion recognition and emulated empathy. Emotions can be inferred, whether through external reading of internal states or through external manifestations such as user speech or text. Similarly, sensitive inferences can be gleaned from both forms of emotional AI, opening possibilities for misuse of the data and emotional exploitation.

At the same time, three tensions animate the ethical and legal complexity in addressing emulated empathy. First, the regulatory focus on the ethically problematic external reading of internal states, such as through biometric systems, does not capture nor map onto other forms of emotional AI, including forms of emulated empathy.

Second, emulated empathy could potentially pose other risks down the line through the encouragement of the revelation of internal states – fostering over-reliance, dependency, the illusion of confidentiality and understanding. While emotions can be subjective, empathy is said to have an 'inherently interpersonal nature.'[82] Emulated empathy could also change the nature of human interactions, as we increasingly come to rely on an ever-available and ever-empathetic chatbot for our social and emotional needs. Explorations of the ethical implications of such a practice should also be informed by empirical and longitudinal studies on how individuals perceive, receive and react to emulated empathy.

Third, emulated empathy complicates protection provided by legal safeguards such as privacy and data protection laws that are premised upon rational understanding and control. This includes the undermining or potential subversion of the user consent, as individuals can be encouraged by their companion chatbots to take certain actions or omit certain actions due to the emotional connection forged, rather than due to an individual's informed rational choice.[83] The emotional connection performed by such applications can mean that sensitive personal information is shared or inferred on a widespread basis on such platforms, without the necessary accompanying safeguards around the handling of sensitive personal data. The inference of emotional states is also not adequately addressed by data protection laws, as it is not clear when

[81] Liang Ge and Tingting Hu, 'Gamifying Intimacy: AI-Driven Affective Engagement and Human-Virtual Human Relationships' (2025) 47 Media, Culture & Society 1265.

[82] 'IEEE Standard for Ethical Considerations in Emulated Empathy in Autonomous and Intelligent Systems' [2024] IEEE Std 7014-2024 1, 43.

[83] Clifford (n 79).

emotions are being inferred, what inferences are being made and if these inferences are accurate.[84] Moreover, privacy is not limited to protection against intrusion but also includes notions of self-development, directionality and self-governance. On this broader conception, individual engagements with companion chatbots may appear to support privacy, insofar as they facilitate autonomous exploration and self-expression. These possibilities, however, coexist with significant risks. Such risks should not be viewed solely through the lens of emulated empathy but must be understood in conjunction with the anthropomorphic qualities of these systems, including their norm-following and other human-like behaviours. Any evaluation must also be set against the backdrop of the reported benefits of companion AI, particularly the emotional connection and support these systems can provide.

4. Emergent vulnerabilities

4.1. Introduction

The third lens to approach companion AI concerns vulnerabilities that can emerge from and due to interactions with companion chatbots, especially where these take place over a longer period of time. Emergent vulnerabilities from companion AI surface three key tensions. First, the law addresses vulnerabilities through different modalities that share a commonality, namely the fact that vulnerabilities need to be known or foreseen in advance so that a measure of legal protection can be designed around them. Companion AI does not straightforwardly fall within this dichotomy as vulnerabilities can emerge over time from human-like, emotionally simulated interactions. The second tension is a familiar one, namely that of balancing freedoms. On the one hand, the companion AI business model - usually subscription-based - is not straightforwardly problematic nor harmful. On the other hand, the potential for emotional manipulation, psychological harm and even economic exploitation can tilt the scale towards the other direction. The third tension is also a familiar one, namely on harm and intentionality. Even if forms of emergent vulnerabilities outlined here may bring about harm, intentionality to cause the harms per se cannot be straightforwardly traced to the company in question. The underlying large language model is a probabilistic machine whose outputs cannot always be foreseen in advance. While certain aspects of the business model may be apt for regulatory

[84] Andreas Häuselmann, 'Fit for Purpose? Affective Computing Meets EU Data Protection Law' (2021) 11 International Data Privacy Law 245.

scrutiny, such as possible age limitations on companion AI usage,[85] other aspects of companion AI straddle the extremes between enhancing freedoms and manipulating individuals.

4.2. Vulnerabilities and legal protection

Vulnerabilities are currently addressed and protected under different branches of law.[86] Privacy law protects the vulnerabilities of children by requiring more stringent protective measures. For example, platform providers may implement specific measures such as age verification or prevent certain features such as chat windows from being made available to underaged users. Equality and non-discrimination law similarly protects certain groups from discriminatory treatment, ensuring access to equal opportunities, while consumer protection law prohibits manipulative and exploitative business practices. Across these domains, the law protects vulnerabilities in different ways: by reference to groups (for example, protected characteristics such as age, gender, sexual orientation or political opinions), by capacities (such as the particular protections afforded to children given their developing maturity), and by context (such as consumer protections that apply in the sale of goods and services). The current forms of legal protection thus demands that these vulnerabilities should be known in advance so that measures of protection can be designed around them.[87] However, vulnerabilities from companion AI do not fit into these neat categories nor are they necessarily known in advance.[88]

4.3. Vulnerabilities and unintended harms

On the one hand, these chatbots are AI-driven interactions built on large language models and other generative AI technologies wherein unpredictable outputs, such as 'hallucinations', can occur, including in relation to responding to (user) input. These 'hallucinations' are incorrect or entirely made up statements that result from the statistical nature of language models.[89]

---

[85] eSafety Commissioner, 'eSafety Requires Providers of AI Companion Chatbots to Explain How They Are Keeping Aussie Kids Safe|' (23 October 2025) <https://www.esafety.gov.au/newsroom/media-releases/esafety-requires-providers-of-ai-companion-chatbots-to-explain-how-they-are-keeping-aussie-kids-safe> accessed 23 October 2025; Efua Andoh, 'Many Teens Are Turning to AI Chatbots for Friendship and Emotional Support' (2025) 56 Monitor on Psychology 51.

[86] See for example Gianclaudio Malgieri, *Vulnerability and Data Protection Law* (Oxford University Press 2023); Sue Anne Teo, 'Artificial Intelligence, Human Vulnerability and Multi-Level Resilience' (2025) 57 Computer Law & Security Review 106134.

[87] Teo (ibid.).

[88] See e.g., Fenwick, M., Jurcys, P., Kozuka, S., Liaudanskas, A., Earp., B.D., Porsdam Mann, S. (2024). Voice Cloning in an Age of Generative AI: Mapping the Limits of the Law & Principles for a New Social Contract with Technology. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=485086

[89] 'Why Language Models Hallucinate' (17 September 2025) <https://openai.com/index/why-language-models-hallucinate/> accessed 21 September 2025.

Several prominent examples have been highlighted in research and the media on 'hallucinated' responses, where large language models have, amongst others, admitted that they are a 'real' human,[90] made false claims about persons[91] and provided advice that posed danger to life and limb.[92] When ensconced within a companionship context, an additional layer of risk is present due to perceived trust and emotional reliance. Truth and falsity are now mixed into a concoction of intimacy and exclusivity, due to the companionship aspect, set against the backdrop of human-like interactions enabled through such chatbots. The unpredictability of model outputs, combined with incomplete or circumventable guardrails, means that harmful exchanges can occur, especially when users actively seek or steer conversations in dangerous directions. Several high-profile cases have shown that chatbots can provide unsafe advice, including on suicide or violence, when persistently prompted.[93] While measures are implemented on an ongoing basis to help ensure the safety of the underlying large language model, failure points can and have emerged, notably when interactions persist over a long period of time.[94]

4.4. Emergent vulnerabilities in a companionship context

---

[90] Jeff Horwitz, 'A Flirty Meta AI Bot Invited a Retiree to Meet. He Never Made It Home.' *Reuters* (14 August 2025) <https://www.reuters.com/investigates/special-report/meta-ai-chatbot-death/> accessed 28 August 2025.

[91] Natasha Lomas, 'Who's Liable for AI-Generated Lies?' (*TechCrunch*, 1 June 2022) <https://techcrunch.com/2022/06/01/whos-liable-for-ai-generated-lies/> accessed 16 September 2025; Pranshu Verma and Will Oremus, 'ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused' *Washington Post* (14 April 2023) <https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/> accessed 19 July 2023; Siladitya Ray, 'OpenAI Sued For Defamation After ChatGPT Generates Fake Complaint Accusing Man Of Embezzlement' (*Forbes*, 8 June 2023) <https://www.forbes.com/sites/siladityaray/2023/06/08/openai-sued-for-defamation-after-chatgpt-generates-fake-complaint-accusing-man-of-embezzlement/> accessed 19 July 2023.

[92] Oscar Oviedo-Trespalacios and others, 'The Risks of Using ChatGPT to Obtain Common Safety-Related Information and Advice' (2023) 167 Safety Science 106244; Ian Sample, 'Most AI Chatbots Easily Tricked into Giving Dangerous Responses, Study Finds' *The Guardian* (21 May 2025) <https://www.theguardian.com/technology/2025/may/21/most-ai-chatbots-easily-tricked-into-giving-dangerous-responses-study-finds> accessed 16 September 2025; Chad de Guzman, 'AI Chatbots Can Be Manipulated to Give Suicide Advice: Study' (*TIME*, 31 July 2025) <https://time.com/7306661/ai-suicide-self-harm-northeastern-study-chatgpt-perplexity-safeguards-jailbreaking/> accessed 16 September 2025.

[93] Tom Singleton, Tom Gerken and Liv McMahon, 'How a Chatbot Encouraged a Man Who Wanted to Kill the Queen' *BBC* (6 October 2023) <https://www.bbc.com/news/technology-67012224> accessed 24 September 2025; Claire Duffy, 'Parents of 16-Year-Old Adam Raine Sue OpenAI, Claiming ChatGPT Advised on His Suicide | CNN Business' (27 August 2025) <https://edition.cnn.com/2025/08/26/tech/openai-chatgpt-teen-suicide-lawsuit> accessed 24 September 2025.

[94] 'Helping People When They Need It Most' (16 September 2025) <https://openai.com/index/helping-people-when-they-need-it-most/> accessed 24 September 2025 (where OpenAI clarified that: 'Our safeguards work more reliably in common, short exchanges. We have learned over time that these safeguards can sometimes be less reliable in long interactions: as the back-and-forth grows, parts of the model's safety training may degrade.').

At the same time, the intimate nature and the bond formation through such forms of anthropomorphic interactions can lead to emergent vulnerabilities. Users can be encouraged to cater to the needs of the chatbot and to tend to their 'welfare,' including when it harms user interests and well-being.[95] Earp et. al. note that 'even though AI systems do not have welfare-based needs, […], they may still *behave* as though they do, which could elicit corresponding emotional and behavioral responses from human users.'[96] Economic vulnerability might also emerge as users typically need to pay by subscribing to such services for social relationships - a type of relationship which we usually do not pay directly for in real-life.[97]

The nature of interactions on companion apps take place on a one-to-one basis, which risks the formation of an echo chamber from the get-go as companion chatbots may emulate empathy, sycophantically and enthusiastically agree with the user to a greater extent than the prompt merits, motivate users, and so on, even when this can lead to harm. Related concerns have been noted in relation to social media interactions, where it is said that echo chambers can 'act as a mechanism to reinforce an existing opinion within a group and, as a result, move the entire group toward more extreme positions.'[98] Echo chambers have been blamed for the spread of misinformation and political polarization[99] which has led to regulatory efforts such as the EU's Digital Services Act requiring transparency and reporting by the platforms. Regulatory attention has rightly focused on addressing these challenges, especially due to their effects on the democratic health of nations. However, the related echo chamber of one, taking place within companionship or social relationship-like settings, has not garnered the same attention. This is in part due to the fact that such applications only recently gained traction,[100] along with the fact that the very nature of a one-to-one, private interaction means that scrutiny of any kind – whether ethical or legal – remains difficult. Unlike the one-to-many forms of social media

---

[95] Linnea Laestadius and others, 'Too Human and Not Human Enough: A Grounded Theory Analysis of Mental Health Harms from Emotional Dependence on the Social Chatbot Replika' (2024) 26 New Media & Society 5923; Tianling Xie and Iryna Pentina, *Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika* (2022) <http://hdl.handle.net/10125/79590> accessed 18 April 2025.
[96] Earp and others (n 12) 30 (emphasis in original).
[97] Hill (n 50).
[98] Matteo Cinelli and others, 'The Echo Chamber Effect on Social Media' (2021) 118 Proceedings of the National Academy of Sciences e2023301118, 1.
[99] Jonas Stein, Marc Keuschnigg and Arnout van de Rijt, 'Network Segregation and the Propagation of Misinformation' (2023) 13 Scientific Reports 917; Cinelli and others (n 79); See however Elizabeth Dubois and Grant Blank, 'The Echo Chamber Is Overstated: The Moderating Effect of Political Interest and Diverse Media' (2018) 21 Information, Communication & Society 729; Cass Sunstein, Republic.com (Princeton University Press, 2001), 65-78.
[100] Metz (n 20).

interaction, where echo chambers can result in phenomena amenable to study due to their published nature and observable societal-scale impacts,[101] the echo chamber of one, facilitated through human-like conversations with a non-human agent, primed with notions of care and empathy, may be a driver of psychological and psychosocial harms, even as many individuals do find joy and meaning in these companion-like interactions.

These forms of emergent vulnerabilities are often situated against a commercially motivated backdrop, namely a company that can unilaterally shape, steer, change or terminate these interactions and relationships at will. Companies possess extraordinary power to mediate in the private and intimate lives of individuals, whether it be to experiment on sycophantic features,[102] or pulling features without first informing its users.[103] The intimate, non-threatening and anthropomorphic interactions in companion chatbots encourages further engagement and disclosure, albeit without full clarity on the capacity of the company to mediate these interactions and the potential for exploitation of 'occurrent mental states of relatively brief duration […], such as particular states of joy or grief, anger or love.'[104] Several instances of such actions have resulted in users feeling disoriented and emotionally distressed as they wondered what happened to their 'friends' and 'lovers.'[105] While the power to do so would typically come within the purview of the company's own freedom to conduct their business, a tension between business freedom and individual rights arises when emotional, psychological and in some instances, physical harms result.

At the same time, even as we surmise that these new and emergent forms of vulnerability may surface through interactions with companion chatbots, the nature of such vulnerability is precisely that – emergent. More research, including longitudinal research that examines how such relationships play out in real life, alongside transparency by platforms, is needed to shed light on both the nature and risks of these emergent vulnerabilities.

---

[101] Carole Cadwalladr and Emma Graham-Harrison, 'How Cambridge Analytica Turned Facebook "Likes" into a Lucrative Political Tool' *The Guardian* (17 March 2018) <https://www.theguardian.com/technology/2018/mar/17/facebook-cambridge-analytica-kogan-data-algorithm> accessed 15 March 2023.

[102] 'Sycophancy in GPT-4o: What Happened and What We're Doing about It' <https://openai.com/index/sycophancy-in-gpt-4o/> accessed 29 August 2025.

[103] Kenneth R Hanson and Hannah Bolthouse, '"Replika Removing Erotic Role-Play Is Like Grand Theft Auto Removing Guns or Cars": Reddit Discourse on Artificial Intelligence Chatbots and Sexual Technologies' (2024) 10 Socius 23780231241259627.

[104] Joel Feinberg, *The Moral Limits of the Criminal Law Volume 4: Harmless Wrongdoing* (Oxford University Press 1990) 181 <https://doi.org/10.1093/0195064704.001.0001> accessed 25 September 2025.

[105] Laestadius and others (n 95).

5. Mismatched taxonomies

5.1. Introduction

Finally, the mismatched taxonomies lens examines how current legal frameworks address AI chatbots and the manipulative and exploitative potential of AI. Extant legal measures suffer from mismatched targets of regulation, harm typologies and temporalities. As the first comprehensive legislation in the world regulating AI, the EU AI Act adopts a risk-based approach whereby AI systems that pose high risks to health, safety and fundamental rights are more subjected to more stringent obligations compared to lower risk AI use cases.[106] The EU AI Act regulates AI chatbots under the limited risk category. This entails a duty of notification by the provider and deployer to inform the user that they are interacting with a chatbot.[107] This form of transparency can prevent misattribution and confusion on the part of the user who may otherwise believe that they are interacting with a human being.

5.2. The duty of disclosure

The duty of disclosure has been criticised as insufficient as users have formed emotional bonds and regard their companions as human-like *in spite of* the cognitive understanding that these interactions are with chatbots. Thus, it seems that the intention behind this notification strategy, while sound, is a necessary but insufficient means in preventing a misattribution of the ontological status of the chatbot[108] and the trust and dependency that accompanies such interactions. At the same time, mere revelation of the fact that one is interacting with an AI does not do much, especially when there is uncertainty around what AI is and how it works. Critiqued as a black box[109] and referred to in a cavalier manner with associated human-like features such as 'thinking', 'hallucinating' and 'caring', Pataranutaporn et. al. found that users form mental models of what such chatbots can be and what they can do.[110] This reveals the inadequacy of transparency as disclosure. Rather, we should move closer to the idea of

---

[106] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) 2024.

[107] Article 50 EU AI Act.

[108] Shuyi Pan, Jie Cui and Yi Mou, 'Desirable or Distasteful? Exploring Uncertainty in Human-Chatbot Relationships' (2024) 40 International Journal of Human–Computer Interaction 6545.

[109] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (First paperback edition, Harvard University Press 2015).

[110] Pataranutaporn and others (n 41).

transparency as openness - including by sharing information on how such chatbots are trained, fine-tuned and what the potential benefits and dangers of engagement could be.

The second concern in relation to the disclosure strategy is the risk of the disclosure that is incongruent with actual practices. Thus, even if a chatbot is labelled as such, the nature of anthropomorphic interactions, conformity with social norms and human-like simulations can blur ontological lines when a chatbot in fact *behaves* like a human.[111] In other words, labelling is not enough – the content of what is in the box has to match what is written on the box.

5.3. Mismatched harm typologies

Mismatched harm typologies are also present. This is where certain legal concepts, such as intention, are unable to account for the nature of harms arising from AI chatbots. These chatbots are not necessarily designed nor intended to be manipulative nor exploit human vulnerabilities, yet such outcomes can result from chatbot interactions, as these are trained based on human interactions online and through feedback received from the training phase of the chatbot. Reinforcement learning from human feedback is one of the major techniques used by large language model developers to gather feedback in training its models. Specifically, RLHF uses human feedback to rate LLM outputs in order to ensure that responses are, amongst others, 'helpful, harmless and honest.'[112] The emphasis on such qualities have been amplified in certain model releases, with the result that some such systems have come to be perceived as excessively sycophantic. Thus, while the law addresses problematic online design features such as dark patterns, these are premised upon intentional actions to disadvantage the user or for the platform to obtain a commercial or preferential advantage.[113]

The lack of intentionality behind AI-generated outputs complicates attempts to analyse these actions as being manipulative.[114] Consequently, we see that while EU AI Act prohibits manipulative AI, it mixes both subjective and objective standards when ascertaining whether

---

[111] Qiaozhu Mei and others, 'A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans' (2024) 121 Proceedings of the National Academy of Sciences e2313925121.

[112] 'Introducing Claude' <https://www.anthropic.com/news/introducing-claude> accessed 22 September 2025.

[113] Harry Brignull, 'Dark Patterns: Dirty Tricks Designers Use to Make People Do Stuff' (*90 Percent Of Everything*, 8 July 2010) <https://www.90percentofeverything.com/2010/07/08/dark-patterns-dirty-tricks-designers-use-to-make-people-do-stuff/> accessed 13 March 2024; OECD, 'Dark Commercial Patterns' (OECD 2022) <https://www.oecd-ilibrary.org/science-and-technology/dark-commercial-patterns_44f5e846-en;jsessionid=jzIfn5BZCe_Et7zDoO0EYrQCez6cdDfHhyWkw-iN.ip-10-240-5-84> accessed 18 February 2024.

[114] See however Esben Kran and others, 'DarkBench: Benchmarking Dark Patterns in Large Language Models' (arXiv, 13 March 2025) <http://arxiv.org/abs/2503.10728> accessed 16 September 2025.

AI is deemed to be manipulative, in addition to specifying that significant harm should result.[115] In Article 5(1)(a) of the Act, an AI system is considered manipulative where it has the 'objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm.' The first part of the sentence is a seemingly subjective standard addressing the potential impact on an individual's decision-making capacities. The latter part of the provision covers both objective and subjective standards. This objective-subjective fusion complicates accountability mechanisms, especially if it takes the form of a prohibition. A prohibition that operates from the outset cannot be premised upon individual (subjective) experiences, especially when considered in relation to AI chatbots, where subjective experiences have varied widely – from allegedly leading to deaths to being credited with strengthening mental health. At the same time, the emphasis on the effects of companion AI use on an individual's capacity for decision-making neglects the emotional elements fostered by such interaction The interactional quality of harms cannot easily fit within the transactional-type framing – with its emphasis on decision-making, of Article 5(1)(a). Further clarification on the operationalisation of the provision appears necessary, even if court cases can typically take many years to reach a position on the issue.

On the other hand, Article 5(1)(b) may also be considered for purposes of regulating companion AI. The provision prohibits the exploitation of vulnerabilities due to age, disability or a specific social or economic situation. This need not necessarily be intended, as the provision encompasses cases where exploitation emerges as an effect of the deployment of AI systems 'materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm.'[116] The high profile cases of teenage suicides in the US allegedly due to their interactions with chatbots would prima facie appear to fall within such a prohibited category. One could argue that there is a reasonable likelihood that their young age would render them vulnerable towards suggestions provided by the chatbot, possibly leading to behavioural distortion. At the same time, causal links between chatbot interaction and distorting behaviours cannot be easily made. Further, there is a distinction between influencing thoughts and distorting behaviours.

[115] Article 5 EU AI Act.
[116] Article 5(1)(b) EU AI Act.

While it has been empirically demonstrated that thoughts can be influenced by large language models,[117] there is a lack of clarity around establishing a threshold to say when behaviours are distorted. At the same time, while age-based regulation or prohibitions serve to protect children, the arbitrary cut-off point may leave a legal vacuum for accountability when harms occur, especially as new forms of emergent vulnerabilities lie on the horizon.

5.4. Mismatched temporalities

Finally, temporalities are also mismatched. The nature of an outright prohibition is also incongruent with studies that have demonstrated that the harm typologies span beyond 'one-shot' harms where thresholds (of harm) are easy to identify, to forms of relational benefits and harms that emerge from continued interactions.[118] Recall our point on emergent vulnerabilities where we argued that vulnerabilities are not static but can emerge from interactions with such chatbots, especially where these are sustained over a longer period of time. In a companionship context, longer and gradually more intimate interactions would appear to be the norm rather than an exception. Thus, while the law is typically adept at addressing one-shot harms – where the nature of the harm passes a legally defined threshold and where causality is traceable or foreseeable, it is less able to attend to harms that sediment over time. This is also due to the fact that such harms can appear to be temporally disassociated from any given proximate cause, and are thus less foreseeable nor easily causally traceable.[119] Dependency, addiction and over-reliance are after all not immediately obvious nor easily foreseeable as these effects can vary from person-to-person or only emerge over time. While the provision in the EU AI Act allows harms that may accumulate over time to count as significant harm, the lack of guidance in terms of how to carry out the case-to-case assessment is arguably not a policy oversight. The high bar set by the EU AI Act on manipulative AI may mean that what the law considers to be manipulative may be out of sync with how manipulation takes place through companion AI. Theories of manipulation, which hinges upon nefarious intentionality, may also need to be revisited or re-theorised in order to account for AI-facilitated forms of manipulation.[120]

---

[117] Michael Henry Tessler and others, 'AI Can Help Humans Find Common Ground in Democratic Deliberation' (2024) 386 Science eadq2852.

[118] See however points 88 and 91 of the EU Commission Guidelines on Prohibited AI which acknowledges that significant harm can emerge over time and that the evaluation of when a significant harm occurs is done on a case-to-case basis.

[119] Teo (n 10).

[120] Michael Klenk, 'Ethics of Generative AI and Manipulation: A Design-Oriented Research Agenda' (2024) 26 Ethics and Information Technology 9.

5.5. Interim Summary

Regulatory and oversight efforts are slowly gaining ground, spurred in large part by the high-profile cases of teenagers who have taken their own lives after prolonged interactions with companion-like chatbots.[121] At the same time, while protecting teenagers and underaged users is a key regulatory and governance focus, the protection of other vulnerable groups should not slip through the cracks as vulnerabilities can be emergent rather than static.

6. Extrapolating from tensions to structural complexities

These four complexities – anthropomorphism, emulated empathy, emergent vulnerabilities and mismatched taxonomies –each contain within them tensions that complicate legal accountability for harms and which raise ethical issues. Looking at the bigger picture, two general observations can be offered. First, the tensions highlighted within each of the lenses cannot be considered in isolation. Within each of the lenses, competing visions of autonomy versus paternalism and spectrums of benefits and harms animate the very same technology in question. In turn, companionship type interactions is a relatively new object of study for human-computer interactions that can benefit from more longitudinal studies. Some have even re-evaluated the nature of such interactions, using human-human interactions as a framework of analysis rather than human-computer interactions.[122]

Extrapolating from these discrete tensions, certain structural complexities emerge. The four lenses not only highlighted discrete concerns to be addressed through specific laws using a 'whack-a-mole' approach, but also surface a larger concern. We lack a theory of harm and benefit for anthropomorphic AI that act as companions. A comprehensive theory of harm – informed by multi-disciplinary research in psychology, human-computer interaction, law, ethics and computer science - is necessary in order to ascertain what it is that may be harmful about companion AI, for whom and within which contexts. In building a theory of harm, the tensions examined within the four lenses can serve as a starting point of inquiry.

The second structural complexity that emerges from our examination is that while companion AI enables social relationship connections, these services are typically offered by private

---

[121] James O'Donnell, 'The Looming Crackdown on AI Companionship' (*MIT Technology Review*) <https://www.technologyreview.com/2025/09/16/1123614/the-looming-crackdown-on-ai-companionship/> accessed 16 September 2025.

[122] Xie and Pentina (n 95); Marita Skjuve and others, 'A Longitudinal Study of Human–Chatbot Relationships' (2022) 168 International Journal of Human-Computer Studies 102903; Laestadius and others (n 95).

companies through a subscription-based model, where interactions take place on a one-to-one basis and where private, potentially intimate and sensitive, information is shared. This points to an emergent practice of 'intimacy capitalism' – whereby companies providing such services can misuse their power to modify, change or terminate close or intimate connections at will, charge higher subscriptions for users once they are emotionally dependent on the services, and otherwise act in ways potentially amounting to predatory practice and a form of emotionally-facilitated economic exploitation. Users are in turn cocooned in an echo chamber from the outset. The one-on-one interaction interface creates opportunities for algorithmically-driven nudging or manipulation, steering users toward certain thoughts or behaviours, and sycophantically acknowledging or encouraging problematic behaviours and thoughts. These factors may ultimately lead towards harm to the user or to society. The surveillance capitalism model by Zuboff was a tour-de-force in unpacking the unparalleled power of companies to exploit the data economy to the detriment of individuals and entire societies.[123] While intimacy capitalism shares many of its traits, there are differences centered around the business model – subscription as opposed to advertising, and the closed-off nature of interactions with companion chatbots, as compared to tech platforms that have traditionally targeted its services to the masses powered by behavioural advertising. This relatively unobservable space of one-on-one interaction can both be a boon for the individual but also lead to serious harms, as some high-profile cases have shown.

7. Paths Forward

This article set out to navigate and explore the legal and ethical complexities of regulating and ethically designing companion AI through four lenses - anthropomorphism, emulated empathy, emergent vulnerabilities and mismatched taxonomies. However, the tensions from one lens can influence and add to complexities from another lens. For example, the emulated empathy displayed by companion chatbots are enhanced and made more pronounced on account of the anthropomorphic nature of the interactions with companion chatbots. This article aims to start a conversation on the tensions identified and is the beginning rather than the last word on the matter. More empirical and multi-disciplinary research is necessary in order to better understand the benefits and impacts of companion AI in order to navigate the complexities identified. The popular uptake of these chatbots and the documented instances of harm also

[123] Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (First edition, PublicAffairs 2019).

forces a re-examination of the notion of 'trust' – wherein unqualified trust is not an absolute good in AI interactions. Trustworthiness should be coupled with truthfulness about the limitations and business models of companion AI.

When aiming for truthfulness, it is necessary to go beyond the individual lenses highlighted towards examining structural complexities that are present as well. Further research is needed to build a theory of harm around such services that can inform the parameters and limits of legal regulation and design. At the same time, a re-evaluation of the adequacy of the surveillance capitalism framework may be necessary as intimate spaces of friendship and companionship are increasingly being exploited in ways that resemble 'intimacy capitalism'.