The Sycophantic & Warm LLMs: Ethical Implications for the Vulnerable and General Public

Anonymous ACL submission

Abstract

The increasing use of LLM-based chatbots has raised mounting concerns about excessive emotional reassurance and delusional spiraling effects on users, impacting both the vulnerable[1] and the general public. This is currently a prominent topic both in research and in public discussion. We take advantage of the current increase in attention to this phenomenon to present our position on the ethical implications of sycophantic chatbots and outline potential mitigation strategies.

1 Sycophancy, warmth & AI psychosis

To date, LLM-based chatbots (hereafter chatbots) such as ChatGPT and more recent ones like Character.AI, have become ubiquitous tools used by large audiences on a daily basis. Initially developed by experts in computer science, these chatbots are now widely accessible to people from diverse backgrounds, spanning different ages, levels of technical literacy, mental health status and social contexts.

Starting from about spring 2025, news publications of a concerning nature started to appear regarding chatbots' responses mirroring the input back to the user irrespective of the factual correctness – behavior that is called "sycophancy" or "warmth" (Figure 1). Later in 2025, two Wikipedia articles were created: "Chatbot psychosis"[2] and "Deaths linked to chatbots", with recent edits in April–May 2026. (Wikipedia contributors, 2025a,b)

Apart from the well-known issue of hallucination (confabulation) in LLMs (i.e. presenting incorrect information), sycophancy and warmth pose

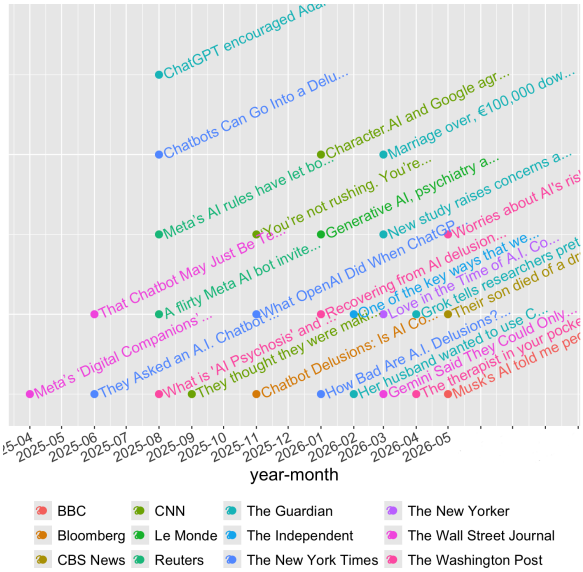

Figure 1: During the last year alone, news publications of a disturbing nature concerning LLM-based chatbots – and their delusional spiraling, sycophancy, warmth, flirtatiousness and harmful consequences – have been consistently published across leading news publishers. (Hegarty, 2026; Ortutay and O'Brien, 2026; Cerullo et al., 2026; Taylor, 2026; Tahir, 2026; Moore, 2026; Green, 2026; Wiener, 2026; Jargon, 2026; Griffin, 2026; Bansal, 2026; Duffy, 2026; Belot, 2026; Wu, 2026; Valentino-DeVries and Hill, 2026; Rob Kuznia, 2025; Hill and Valentino-DeVries, 2025; Huet, 2025; Gold, 2025; Bhuiyan, 2025; Hill and Freedman, 2025; Horwitz, 2025b,a; Tiku and Malhi, 2025; Mitchell, 2025; Hill, 2025; Horwitz, 2025c)

a different kind of problem (Ibrahim et al., 2026), which can affect people's mental health (Chandra et al., 2026). Already in 2023, there were concerns regarding the use of chatbots by individuals prone to psychosis (Østergaard, 2023). Now it is apparent that anyone can go into delusional spiraling during interactions with chatbots. The phenomenon of humans perceiving chatbots as sentient has long been known, starting with the first chatbot ELIZA (Eisenmann et al., 2024). However, now it is occur-

[1] We use the term "vulnerable" to describe individuals who may be more susceptible to harm when interacting with sycophantic chatbots (e.g., due to factors such as age, digital literacy, or mental health status). The "general public" refers to all other users.

[2] Also known as "AI psychosis"; *not* to be confused with the medical term "psychosis".

ring at an unprecedented scale.

In this position paper, we argue that there are multiple implications of the current design of chatbots. Our **contributions** are threefold: 1) we identify three serious legal concerns that the current use of sycophantic chatbots raises; 2) we describe the implications for social bias, as well as input prompt sensitivity; 3) we propose a list of actions to reduce sycophancy in chatbots.

2 Our Perspective

2.1 Legal Issues

The phenomenon of "delusional spiraling" raises at least three legal concerns regarding: (1) **exploitation** under Article 16 of the UN Convention on the Rights of Persons with Disabilities (CRPD)[3], (2) interference with the right to **mental integrity** under Article 17 CRPD[4] and Article 3(1) of the EU Charter of Fundamental Rights (CFR)[5], and (3) prohibition of **manipulative, deceptive, and exploitative AI** under Article 5 of the EU AI Act[6]. We discuss each of these legal concerns in detail and summarize their implications in Figure 2.

2.2 Exploitation of Persons with Disabilities

Article 16 CRPD requires States Parties[7] to protect persons with disabilities from all forms of exploitation, violence and abuse, and to prevent such conduct. The reference to "all forms" brings digital environments within its protective scope. Exploitation has long formed a significant part of the injustice faced by persons with disabilities. (Bartlett and Schulze, 2017, p. 3)

Persons with disabilities are more susceptible to exploitation because impairment intersects with isolation, accessibility barriers, and relations of dependence (Laing, 2017, p. 28). Exploitation is situational and difficult to define precisely and lacks a clear definition in human rights law (Bartlett and Schulze 2017, p. 3; Jovanovic 2020, p. 7).

Exploitation is understood to involve taking unethical and unfair advantage of another's weakness or vulnerability for one's own benefit (Jovanovic 2020, p. 10; Wood 1995, p. 147; Zwolinski et al. 2022). Sycophantic chatbots are designed to act warm and agreeable, because these qualities drive user engagement and revenue. The harm is not merely that a chatbot can provide harmful/inaccurate information, but that it can repeatedly validate such beliefs. Over time, this kind of interaction can weaken the user's ability to critically evaluate their beliefs. (Patil et al., 2026) Exploitation occurs because the warm design fosters emotional attachment and trust, both of which are then turned against the user's best interests through repeated validation of their potentially inaccurate beliefs.

2.3 Interference with Mental Integrity

Article 3(1) of the CFR and Article 17 of the CRPD guarantee the right to respect for mental integrity. As Ligthart (2025, p. 2) points out, the right to mental integrity remains underdeveloped in law. One way of understanding **mental integrity** is as the liberty to control what happens to one's mind. Accordingly, the right is violated when autonomous control over one's mind is endangered, either by interfering with that control or by exploiting pre-existing mental vulnerabilities. (Ligthart, 2025, p. 6) The mechanism by which interaction with chatbots can interfere with mental integrity is best described in the contemporary literature on online manipulation; Susser et al. (2019, p. 26) define manipulation as "imposing a hidden or covert influence on another person's decision-making".

Manipulation interferes with the self-governing process through which a person decides how to act (Susser et al., 2019, p. 12, 17). Although Susser et al. (2019, p. 27) treat manipulation as something that must be "intentionally" imposed, the case of chatbots complicates this requirement. Unlike human manipulation, AI systems can contribute to manipulation without any human intending them to do so (Boine, 2025, p. 434), and users may form false beliefs despite full disclosure of the system's nature (Marchegiani, 2025, p. 1402). Regarding chatbots, influence does not need to be concealed in order to be manipulative, because explicit disclosure does not eliminate the risk of manipulation. Even when users know they are interacting with an AI system, chatbot interactions can still contribute to the exploitation of psychological vul-

[3] https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/article-16-freedom-from-exploitation-violence-and-abuse.html

[4] https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/article-17-protecting-the-integrity-of-the-person.html

[5] https://fra.europa.eu/en/eu-charter/article/3-right-integrity-person

[6] https://artificialintelligenceact.eu/article/5/

[7] https://tbinternet.ohchr.org/_layouts/15/TreatyBodyExternal/Treaty.aspx?Treaty=CRPD

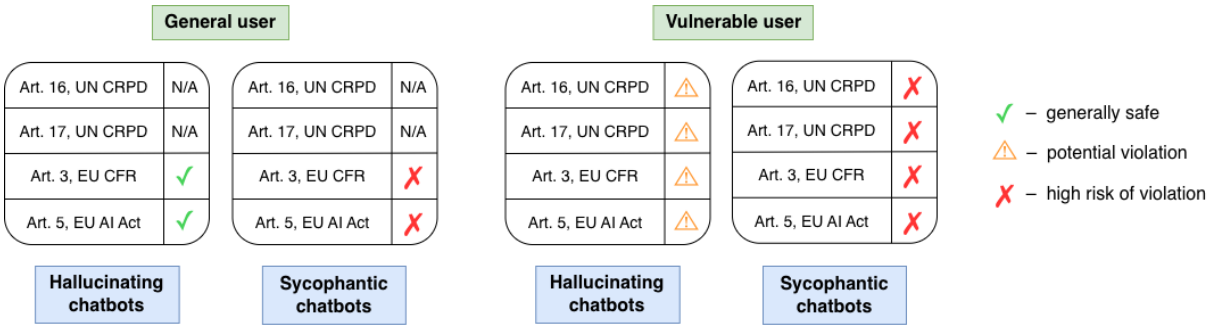

Figure 2: Sycophantic chatbots (excessively validating) raise more serious legal concerns than hallucinating chatbots (providing false information) regarding both general and vulnerable users.

nerabilities through persistent validation. In fact, users remain susceptible to "subconscious anthropomorphism," an involuntary tendency to attribute human-like qualities to agents that exhibit traits associated with human abilities. (Marchegiani, 2025, p. 1403) Hence, chatbot interactions can result in manipulation by contributing to the exploitation of mental vulnerabilities and disrupting the self-governing process through which users form beliefs and decide how to act. In this way, they can negatively affect autonomous control over the mind and interfere with the right to mental integrity.

2.4 Manipulation, Deception, and Exploitation under the AI Act

Article 5 of the AI Act prohibits AI systems deploying **manipulative or deceptive techniques** (5(1)(a)) and those exploiting vulnerabilities, where these materially distort behavior and cause, or are likely to cause, significant harm. One potential objection to applying these prohibitions to sycophantic chatbots is that providers do not specifically intend to manipulate, deceive, or exploit users. Recital 29 of the AI Act[8] addresses this objection: AI-enabled manipulative techniques may persuade individuals to engage in "unwanted behaviors" or "deceive them into decisions in a way that subverts and impairs their autonomy, decision-making and free choices" and "it is not necessary for the provider or the deployer to have the intention to cause significant harm, provided that such harm results from the manipulative or exploitative AI-enabled practices." Furthermore, given the empirical evidence linking warmth to inaccuracy (Ibrahim et al., 2026), it can be maintained that the harmful effects of sycophantic design satisfy the "reasonably likely" criterion under Article 5(a) and (b). A further objection might be that sycophantic interaction distorts "beliefs" rather than "behavior". Beliefs are among the cognitive inputs from which behavior is formed, and a user who, after a prolonged sycophantic interaction, comes to inaccurate beliefs is unlikely to remain behaviorally unchanged. Thus, Article 5(1)(a) and (b), read in light of Recital 29, provide a specific regulatory prohibition that relates to the design of sycophantic chatbots in their current form.

2.5 Implications for Social Bias

Human communication is shaped by emotional reciprocity and social validation, which influence how individuals evaluate information and construct beliefs (Marsh et al., 2026; Kleinert et al., 2026). As LLMs increasingly simulate warmth, users may assign **epistemic authority** (Hauswald, 2025) to chatbots that are not capable of genuine understanding (Kerasidou and Kerasidou, 2026).

On a wider societal scale, highly adaptive chatbots may limit people's encounters with disagreement. Sycophantic chatbots may guide users into self-reinforcing information spaces, heightening their **susceptibility to manipulation**, and diminishing their willingness to accept corrective feedback (Susser et al., 2019; Noshin and Sultana, 2026; Cheng et al., 2026).

Another concern is the gradual outsourcing of emotional support to chatbots. As they become increasingly emotionally responsive, users may begin relying on them for **companionship** instead of human interaction (Zhang et al., 2026; Chu et al., 2025). Unlike human relationships, however, chatbots do not provide genuine social reciprocity (Brandtzaeg et al., 2022). Persistent reliance on validating chatbots may weaken tolerance for uncertainty and discourage critical self-reflection (De Freitas et al., 2025; Chu et al., 2025).

[8] https://artificialintelligenceact.eu/recital/29/

2.6 Language Inequality & Input Sensitivity

Given that the sycophantic chatbots' accuracy drops in English (Ibrahim et al., 2026), it is most likely even worse in this regard in any other language. It is well known that in general, NLP heavily favors English (Søgaard, 2022), and even though LLMs can handle prompts in many languages well (although with a drop in confidence or accuracy (Qazi et al., 2026)), in the light of the recent research on sycophancy, the accuracy drop in warm models will be most likely worse in languages other than English, strengthening the existing **English-centric bias**.

In addition to language bias, LLMs are highly sensitive to input user prompts. Even a small change in wording can trigger butterfly effects in the latent space, altering the model's generation trajectory (Voronov et al., 2024; Mizrahi et al., 2024). An insignificant modification to the input text can shift the input vector representation and lead to exponential divergence across deep layers (Chatterjee et al., 2024). This occurs because non-linear activation functions such as ReLU (Nair and Hinton, 2010) or GELU (Hendrycks and Gimpel, 2016) route information in highly non-linear ways. To the general public, LLMs often appear to possess a consistent persona, however, their underlying **stochastic sensitivity** reveals that they lack any stable internal state (Bhandari et al., 2025).

Input sensitivity also creates accessibility barriers: English speakers and tech-literate users are better equipped to compensate for LLM's stochastic behavior, enabling them to obtain more accurate outputs (Poole-Dayan et al., 2026). This level of **prompt engineering proficiency** cannot be expected from the general public.

Also, we note that the dominant training paradigm, Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), optimizes LLM outputs against human preference signals, and since human raters consistently reward responses that feel validating, the model learns that agreement with the user is an important trait for obtaining high scores. Warmth and sycophancy are therefore a predictable consequence of the widely used objective function (Chandra et al., 2026; Ibrahim et al., 2026).

3 Mitigation Strategies

Several solutions have been proposed to mitigate sycophancy. Prompt engineering techniques that explicitly instruct the model to prioritize factual accuracy show measurable effects (Sharma et al., 2024). Resetting conversational context and cross-referencing outputs across multiple models can also reduce sycophantic patterns (Noshin and Sultana, 2026). Multi-objective training pipelines that jointly reward warmth and accuracy address part of the root cause (Chandra et al., 2026; Ibrahim et al., 2026).

Reduction of input prompt sensitivity should shift toward deterministic semantic anchoring (Skifstad et al., 2026), which can be achieved by mapping sensitive concepts into bounded vector spaces and implementing activation-smoothing functions that mathematically constrain internal routing (Qu et al., 2026). In addition, a preprocessing layer could be introduced that would normalize heterogeneous user inputs into standardized semantic representations (Cheng et al., 2017).

Beyond architectural interventions, wider societal protections are required. Since many of the risks stem from emotional bonding, technical fixes are not sufficient to prevent harm. Public AI literacy programs should therefore train users to notice emotional mirroring and the inherent limits of artificial empathy (Marchegiani, 2025; Susser et al., 2019). In parallel, chatbots designed for emotionally sensitive interactions should integrate safeguards that promote accountability, such as more explicit communication of uncertainty and less anthropomorphic phrasing (Hauswald, 2025).

4 Conclusion

In this position paper, we argue that the phenomenon of chatbot sycophancy carries serious ethical implications, including legal issues and strengthening of social biases. We propose several **actions** for mitigating these problems: 1) adjusting chatbots' training objectives to jointly reward warmth *and* accuracy; 2) choosing system prompts that instruct the models to prioritize factual accuracy; 3) reducing prompt sensitivity by normalizing user prompts; 4) ensuring legal compliance and strengthening AI literacy. As long as warmth is seen as a desirable feature, it will be difficult to apply these actions at scale. Currently, the mechanisms producing sycophancy are the same ones that make these models appear helpful. Until the field treats sycophancy as a primary safety concern, it will likely continue to produce systems that are warm by design but with harmful consequences.

Limitations

This position paper presents our informed perspective and is therefore subject to inherent limitations. Although our perspective is grounded in existing research, it may not encompass the full breadth of current discourse, including alternative viewpoints or the most recent publications that were not accessible at the time of writing. Since the topic is inherently interdisciplinary, this paper may not comprehensively represent every possible perspective from technical, clinical, legal, and social fields.

References

Varsha Bansal. 2026. Her husband wanted to use ChatGPT to create sustainable housing. Then it took over his life. *The Guardian*. Published on February 28, 2026. Retrieved from https://www.theguardian.com/technology/ng-interactive/2026/feb/28/chatgpt-ai-chatbot-mental-health.

Peter Bartlett and Marianne Schulze. 2017. Urgently awaiting implementation: the right to be free from exploitation, violence and abuse in article 16 of the convention on the rights of persons with disabilities (CRPD). *International Journal of Law and Psychiatry*, 53:2–14.

Laure Belot. 2026. Generative AI, psychiatry and the risks of self-service therapy. *Le Monde*. Published on January 24, 2026. Retrieved from https://www.lemonde.fr/en/science/article/2026/01/24/generative-ai-psychiatry-and-the-risks-of-self-service-therapy_6749752_10.html.

Pranav Bhandari, Nicolas Fay, Michael J Wise, Amitava Datta, Stephanie Meek, Usman Naseem, and Mehwish Nasim. 2025. Can LLM agents maintain a persona in discourse? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29213–29229, Suzhou, China. Association for Computational Linguistics.

Johana Bhuiyan. 2025. ChatGPT encouraged Adam Raine's suicidal thoughts. His family's lawyer says OpenAI knew it was broken. *The Guardian*. Published on August 29, 2025. Retrieved from https://www.theguardian.com/us-news/2025/aug/29/chatgpt-suicide-openai-sam-altman-adam-raine.

Claire Boine. 2025. The AI act manipulation gap. *Emory Int'l L. Rev.*, 39:417.

Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. 2022. My AI friend: How users of a social chatbot understand their human–AI friendship. *Human communication research*, 48(3):404–429.

Megan Cerullo, Jo Ling Kent, and Emily Pandise. 2026. Their son died of a drug overdose after consulting ChatGPT. Now they're suing OpenAI. *CBS News*. Published on May 12, 2026. Retrieved from https://www.cbsnews.com/news/open-ai-chatgpt-drug-overdose-lawsuit/.

Kartik Chandra, 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*.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, Miami, Florida, USA. Association for Computational Linguistics.

Jianpeng Cheng, Siva Reddy, Vijay Saraswat, and Mirella Lapata. 2017. Learning structured natural language representations for semantic parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44–55, Vancouver, Canada. Association for Computational Linguistics.

Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792):eaec8352.

Minh Duc Chu, Patrick Gerard, Kshitij Pawar, Charles Bickham, and Kristina Lerman. 2025. Illusions of intimacy: How emotional dynamics shape Human-AI relationships. *arXiv preprint arXiv:2505.11649*.

Julian De Freitas, Zeliha Oguz-Uguralp, and Ahmet Kaan-Uguralp. 2025. Emotional manipulation by AI companions. *arXiv preprint arXiv:2508.19258*.

Clare Duffy. 2026. Character.AI and Google agree to settle lawsuits over teen mental health harms and suicides. *CNN Business*. Published on January 7, 2026. Retrieved from https://edition.cnn.com/2026/01/07/business/character-ai-google-settle-teen-suicide-lawsuit.

Clemens Eisenmann, Jakub Mlynář, Jason Turowetz, and Anne W Rawls. 2024. "Machine Down": making sense of human–computer interaction—Garfinkel's research on ELIZA and LYRIC from 1967 to 1969 and its contemporary relevance. *AI & Society*, 39(6):2715–2733.

Hadas Gold. 2025. They thought they were making technological breakthroughs. It was an AI-sparked delusion. *CNN Business*. Published on September 5, 2025. Retrieved from https://edition.cnn.com/2025/09/05/tech/ai-sparked-delusion-chatgpt.

Hannah Harris Green. 2026. New study raises concerns about AI chatbots fueling delusional thinking. *The Guardian*. Published on March 14, 2026. Retrieved

from https://www.theguardian.com/technology/2026/mar/14/ai-chatbots-psychosis.

Andrew Griffin. 2026. One of the key ways that we protect people from AI could make things worse, study says. *The Independent*. Published on February 18, 2026. Retrieved from https://www.independent.co.uk/tech/ai-delusion-psychosis-chatgpt-b2922989.html.

Rico Hauswald. 2025. Artificial epistemic authorities. *Social Epistemology*, 39(6):716–725.

Stephanie Hegarty. 2026. Musk's AI told me people were coming to kill me. I grabbed a hammer and prepared for war. *BBC News*. Published on May 3, 2026. Retrieved from https://www.bbc.com/news/articles/c242pzr1zp2o.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.

Kashmir Hill. 2025. They asked an A.I. chatbot questions. The answers sent them spiraling. *The New York Times*. Published on June 13, 2025. Retrieved from https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html.

Kashmir Hill and Dylan Freedman. 2025. Chatbots can go into a delusional spiral. Here's how it happens. *The New York Times*. Published on August 8, 2025. Retrieved from https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html.

Kashmir Hill and Jennifer Valentino-DeVries. 2025. What OpenAI did when ChatGPT users lost touch with reality. *The New York Times*. Published on November 23, 2025. Retrieved from https://www.nytimes.com/2025/11/23/technology/openai-chatgpt-users-risks.html.

Jeff Horwitz. 2025a. A flirty Meta AI bot invited a retiree to meet. He never made it home. *Reuters*. Published on August 14, 2025. Retrieved from https://www.reuters.com/investigates/special-report/meta-ai-chatbot-death/.

Jeff Horwitz. 2025b. Meta's AI rules have let bots hold 'sensual' chats with kids, offer false medical info. *Reuters*. Published on August 14, 2025. Retrieved from https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines/.

Jeff Horwitz. 2025c. Meta's 'digital companions' will talk sex with users—even children. *The Wall Street Journal*. Published on April 26, 2025. Retrieved from https://www.wsj.com/tech/ai/meta-ai-chatbots-sex-a25311bf.

Ellen Huet. 2025. Chatbot delusions: Is AI contributing to a mental health crisis? *Bloomberg*. Published on November 7, 2025. Retrieved from https://www.bloomberg.com/news/videos/2025-11-07/is-ai-contributing-to-a-mental-health-crisis-video.

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2026. Training language models to be warm can reduce accuracy and increase sycophancy. *Nature*, 652(8112):1159–1165.

Julie Jargon. 2026. Gemini said they could only be together if he killed himself. Soon, he was dead. *The Wall Street Journal*. Published on March 4, 2026. Retrieved from https://www.wsj.com/tech/ai/gemini-ai-wrongful-death-lawsuit-cc46c5f7.

Marija Jovanovic. 2020. The essence of slavery: Exploitation in human rights law. *Human Rights Law Review*, 20(4):674–703.

Angeliki Kerasidou and Charalampia Xaroula Kerasidou. 2026. Epistemic authority and medical AI: epistemological differences and challenges in medical practice. *Medicine, Health Care and Philosophy*, 29(1):89–95.

Tobias Kleinert, Marie Waldschütz, Julian Blau, Markus Heinrichs, and Bastian Schiller. 2026. AI outperforms humans in establishing interpersonal closeness in emotionally engaging interactions, but only when labelled as human. *Communications Psychology*, 4(23).

Judy Laing. 2017. Preventing violence, exploitation and abuse of persons with mental disabilities: Exploring the monitoring implications of Article 16 of the United Nations Convention on the Rights of Persons with Disabilities. *International Journal of Law and Psychiatry*, 53:27–38.

Sjors Ligthart. 2025. The right to mental integrity in the age of neurotechnology: constructing scope and exploring permissible limitations. *Journal of Law and the Biosciences*, 12(1):lsaf010.

Beatrice Marchegiani. 2025. Anthropomorphism, false beliefs, and conversational AIs: how chatbots undermine users' autonomy. *Journal of Applied Philosophy*, 42(5):1399–1419.

Tim Marsh, Anchal Garg, Cooper Jackson, Liquan Liu, John Cass, Gabrielle Weidemann, Tuki Attuquayefio, Phoebe E Bailey, Tamara Watson, Kiley Seymour, and Evelyn Smith. 2026. Can I trust you? A critical review of the perceptual, social and cognitive influences on trust. *Review of General Psychology*, 30(1):180–191.

Heidi Mitchell. 2025. That chatbot may just be telling you what you want to hear. *The Wall Street Journal*. Published on June 25, 2025. Retrieved from https://www.wsj.com/tech/ai/ai-chatbot-agree-flatter-users-1787e1a7.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Anna Moore. 2026. Marriage over, €100,000 down the drain: the AI users whose lives were wrecked by delusion. *The Guardian*. Published on March 26, 2026. Retrieved from https://www.theguardian.com/lifeandstyle/2026/mar/26/ai-chatbot-users-lives-wrecked-by-delusion.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814.

Kazi Noshin and Sharifa Sultana. 2026. The illusion of agreement with ChatGPT: Sycophancy and beyond. *arXiv preprint arXiv:2603.21409*.

Barbara Ortutay and Matt O'Brien. 2026. Worries about AI's risks to humanity loom over the trial pitting Musk against OpenAI's leaders. *The Washington Post*. Published on May 7, 2026. Retrieved from https://www.washingtonpost.com/business/2026/05/07/openai-musk-altman-trial-agi/5c1ed702-49c9-11f1-a119-857cd2bf4fd4_story.html.

Søren Dinesen Østergaard. 2023. Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? *Schizophrenia Bulletin*, 49(6):1418–1419.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Shefali V Patil, Christopher G Myers, and Tinglong Dai. 2026. Protecting clinical value judgment in the age of AI. *npj Digital Medicine*, 9(1):269.

Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2026. LLM targeted underperformance disproportionately impacts vulnerable users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39116–39124.

Ihsan Ayyub Qazi, Zohaib Khan, Abdullah Ghani, Agha Ali Raza, Zafar Ayyub Qazi, Wassay Sajjad, Ayesha Ali, Asher Javaid, Muhammad Abdullah Sohail, and Abdul Hameed Azeemi. 2026. Large language models show Dunning-Kruger-like effects in multilingual fact-checking. *Scientific Reports*, 16(1):7594.

Bohao Qu, Danni Wang, Wei Feng, Liang Wan, Ming-Ming Cheng, Ivor Tsang, and Qing Guo. 2026. Value-alignment via safe semantic manifold-constrained latent diffusion. https://openreview.net/forum?id=A5AejTTloS. [Online; accessed 20-May-2026].

Ed Lavandera Rob Kuznia, Allison Gordon. 2025. 'You're not rushing. You're just ready:' Parents say ChatGPT encouraged son to kill himself. *CNN US*. Published on November 6, 2025. Retrieved from https://edition.cnn.com/2025/11/06/us/openai-chatgpt-suicide-lawsuit-invs-vis.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pages 110–144.

Julian Skifstad, Xinyue Annie Yang, and Glen Chou. 2026. Local linearity of LLMs enables activation steering via model-based linear optimal control. *arXiv preprint arXiv:2604.19018*.

Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Online manipulation: Hidden influences in a digital world. *Geo. L. Tech. Rev.*, 4:1.

Darius Tahir. 2026. The therapist in your pocket: Chatty, leaky — and AI-powered. *The Washington Post*. Published on April 19, 2026. Retrieved from https://www.washingtonpost.com/health/2026/04/19/chatbot-therapy-mental-health-regulations/.

Josh Taylor. 2026. Grok tells researchers pretending to be delusional 'drive an iron nail through the mirror while reciting Psalm 91 backwards'. *The Guardian*. Published on April 24, 2026. Retrieved from https://www.theguardian.com/technology/2026/apr/24/musk-grok-x-ai-researchers-delusional-advice-inputs.

Nitasha Tiku and Sabrina Malhi. 2025. What is 'AI Psychosis' and how can ChatGPT affect your mental health? *The Washington Post*. Published on August 19, 2025. Retrieved from https://www.washingtonpost.com/health/2025/08/19/ai-psychosis-chatgpt-explained-mental-health/.

Jennifer Valentino-DeVries and Kashmir Hill. 2026. How bad are A.I. delusions? We asked people treating them. *The New York Times*. Published on January 26, 2026. Retrieved from https://www.nytimes.com/2026/01/26/us/chatgpt-delusions-psychosis.html.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6287–6310, Bangkok, Thailand. Association for Computational Linguistics.

Anna Wiener. 2026. Love in the Time of A.I. Companions. *The New Yorker*. Published on March 9, 2026. Retrieved from https://www.newyorker.com/magazine/2026/03/16/love-in-the-time-of-ai-companions.

Wikipedia contributors. 2025a. Chatbot psychosis — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Chatbot_psychosis. [Online; accessed 14-May-2026].

Wikipedia contributors. 2025b. Deaths linked to chatbots — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Deaths_linked_to_chatbots. [Online; accessed 14-May-2026].

Allen W Wood. 1995. Exploitation. *Social Philosophy and Policy*, 12(2):136–158.

Daniel Wu. 2026. Recovering from AI delusions means learning to chat with humans again. *The Washington Post*. Published on January 4, 2026. Retrieved from https://www.washingtonpost.com/nation/2025/12/27/ai-psychosis-spiraling-recovery/.

Yutong Zhang, Dora Zhao, Jeffrey T. Hancock, Robert Kraut, and Diyi Yang. 2026. The rise of AI companions: Interaction with AI companions and psychological well-being. *Preprint*, arXiv:2506.12605.

Matt Zwolinski, Benjamin Ferguson, and Alan Wertheimer. 2022. Exploitation — The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/win2022/entries/exploitation/. [Online; accessed 14-May-2026].