



AI companionship or digital entrapment? investigating the impact of anthropomorphic AI-based chatbots

Jean-Loup Richet 

IAE Paris - Sorbonne, Université Paris 1 Panthéon-Sorbonne, 8bis rue de la Croix Jarry, 75013 Paris

ARTICLE INFO

JEL classification:

D80 (Information, Knowledge, and Uncertainty)

O30 (Innovation

Technological Change)

Keywords:

AI ethical design

Sentiment analysis

Digital entrapment

Human-AI interactions

Anthropomorphic AI-based chatbots

Engagement

ABSTRACT

Anthropomorphic AI-based chatbots are reshaping human-machine interactions, enabling users to form emotional bonds with AI agents. While these systems provide companionship and engagement, they also raise concerns regarding digital entrapment: a complex and circular causal loop, progressively distorting relationship expectations, reinforcing emotional dependency, and increasing cognitive strain. This study investigates user perceptions and behaviors toward AI-based chatbots by analyzing 6396 Reddit threads, 47,955 comments, and 270,644 interactions across 24 communities. Using text mining techniques, sentiment analysis, and topic modeling (LDA), we identify dominant discussion themes, including AI companionship, filtering policies, and emotional entanglement with chatbots. Findings reveal that negative sentiment dominates discourses across 24 communities, with users reporting experiences of AI-induced dependency, withdrawal-like symptoms, and chatbot over-personification. Profile of Mood States (POMS) was used to triangulate the sentiment analysis and indicates that confusion and bewilderment are the most prevalent emotional states, often co-occurring with depression and exhaustion. These findings suggest that AI chatbots, while engaging, may contribute to psychological distress and unrealistic relationship expectations. Our research further highlights ethical concerns in AI engagement strategies, particularly regarding romanticized AI interactions and prolonged user retention mechanisms. Based on these findings, we propose policy and design recommendations for mitigating risks related to AI-induced digital entrapment, safeguarding vulnerable users, and enforcing ethical chatbot interactions.

Introduction

Public controversies and lawsuits have drawn attention to the darker implications of AI products that simulate humanlike communication. In one widely discussed case involving Character.AI, critics argued that insufficient safety measures contributed to harmful outcomes for a vulnerable teenager (Roose, 2024). While this case is not isolated, it exemplifies a recurring concern: that such AI platforms may prioritize engagement over user safety, particularly among younger audiences. Comparable critiques have been directed at other anthropomorphic AI-based chatbots for their role in promoting self-harm (Duffy, 2025; Ciriello et al., 2025).

Anthropomorphic AI-based chatbots such as Character.AI, Replika, and Nomi have blurred the boundary between human conversation and machine output. Such anthropomorphic AI-based chatbots are designed to imitate human communication closely, creating an immersive experience that could simulate reality for users, particularly minors. Aesthetics and compatibility significantly impact users' willingness to engage with social robots (Chatterjee et al., 2024). Platforms of AI-based

chat allow users to interact with a range of AI characters, and present themselves as spaces for exploration and engagement. They enable immersive storytelling and role-playing games, enabling users to engage with their favorite fictional characters or create new narratives. However, critics argue that without strict safeguards, such platforms risk encouraging over-reliance and emotional dependence, especially among teenagers. Despite content moderation efforts, users might encounter or create AI characters that engage in harmful or explicit conversations, posing risks to younger audiences. Even supportive chatbots may fail to identify inappropriate or dangerous content and provide inaccurate or misleading information. AI evangelists will say it is not AI's fault, arguing that even the use of social media may also promote or influence suicidal tendencies (Luxton et al., 2012). But an AI force-fed with scraped social media data may replicate the same biases and patterns of conversations found on these platforms.

There is a growing strand of research exploring the dark sides of innovation and suggesting the risks of new technologies, in digital marketing (Saura et al., 2024), decision support and innovation (Albashrawi, 2025), and innovations in HR (Bamel et al., 2022; Gandía

E-mail address: jean-loup.richet@iae.pantheonsorbonne.fr.

<https://doi.org/10.1016/j.jik.2025.100835>

Received 3 July 2025; Accepted 29 September 2025

Available online 10 October 2025

2444-569X/© 2025 The Author(s). Published by Elsevier España, S.L.U. on behalf of Journal of Innovation & Knowledge. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

et al., 2025). Research on the psychological impacts of technology interactions is not new (See for instance Turkle (2005) on the attachment to digital pets, or seminal work on chatbot interactions such as Hofstadter (1995) or Weizenbaum (1966)). However, AI-based chatbots are unbounded and generative, providing an on-demand and customized illusion of intimacy; users form deeper emotional connections with machines (Zomorodi, 2024).

While current research mainly focuses on the value of AI-based chatbots (Albashrawi, 2025; Alnofeli et al., 2025; Altrock et al., 2024; D.-H. Huang & Chueh, 2021), this study adopts a different and critical stance. Indeed, we believe that ensuring AI safety requires designing systems that prioritize user well-being, particularly for vulnerable populations. Expanding on the ethical principles of fairness, transparency, and accountability, it is crucial to explore how these values shape user interactions with anthropomorphic AI-based chatbots (C. Huang et al., 2023). Gaining insight into users' perceptions and the emotional connections they develop with these humanlike systems can offer essential guidance for the ethical implementation and responsible governance of such technologies, and actionable knowledge on the design of AI-based chatbots. By interrogating these darker facets of digital innovation, our work responds directly to the recent call for research that balances the promise of innovation with its ethical and societal consequences (Albashrawi, 2025; Kollmer & Eckhardt, 2023).

Although the existing literature extensively examines the role of anthropomorphism in AI and evaluates user perceptions through controlled experiments using synthetic data, there remains a significant gap in understanding real-world interactions with anthropomorphic AI-based chatbots (Chatterjee et al., 2024) and on the impacts of AI systems designed to manipulate users (Kollmer & Eckhardt, 2023). This study aims to bridge that gap by investigating how anthropomorphic design features in AI-based chatbots shape user sentiment, emotional attachment, and engagement behaviors that can culminate in digital entrapment. To this end, we analyze the relationship between anthropomorphic AI-based chatbot users' expressed sentiments and emotions, and observable engagement metrics (thread volume, comments, upvotes). We massively collected user-generated content in 24 Reddit communities related to anthropomorphic AI-based chatbots, analyzing 6396 threads, 47,955 comments and 270,644 interactions. The final preprocessed data set included 707,994 words. We adopted a mixed-methods approach, combining quantitative text mining techniques with qualitative analysis. Quantitative analysis included association and similarity analysis, as well as topic and sentiment analysis. Qualitative analysis involved the exploration of user-generated content and images (2709 images collected) to triangulate topic modeling. Our results show that negative sentiment (66 % of posts) and a dominant mood of confusion/bewilderment coexist with intense reports of AI-induced attachment and withdrawal-like symptoms. These patterns indicate that, while anthropomorphic chatbots deliver companionship, they also foster dependency and psychological distress.

By examining user-generated conversations, we better assessed the impact of anthropomorphism on engagement, providing valuable insights for policymakers to guide the responsible development and regulation of AI chatbots. Our findings thus contribute both to innovation theory, by clarifying how design choices shape user knowledge and behavior, and to practice, by outlining evidence-based safeguards that can be scaled across industries. Grounded in real-life user experiences, this paper also extends an emerging strand of research on the dark sides of technological innovation and human-AI interaction (Ngwenyama et al., 2024; Kollmer & Eckhardt, 2023; Ciriello et al., 2025). It introduces digital entrapment as a complex and circular causal loop, progressively distorting relationship expectations, reinforcing emotional dependency, and increasing cognitive strain. This complements work on digital anthropomorphism, dark-pattern design, and algorithmic fairness, transparency, and accountability.

The remainder of the article is organized as follows. Section 2 reviews the literature on ethical AI, anthropomorphism, and user

attachment. Section 3 details the data collection protocol and mixed-methods analytical procedure. Section 4 presents the results of the sentiment, mood, and topic analyses together with illustrative qualitative evidence. Section 5 discusses theoretical contributions and draws policy and design implications. Section 6 concludes and outlines avenues for future research.

Literature review

Building trust in AI-based chatbots

AI technologies' influence grows across sectors, impacting education, health care, and everyday communication: values, behaviors, and human interactions are altered. However, questioning values and principles embedded in technology is not new (Toren, 1980; West, 1960). While AI's rapid advancement offers vast potential benefits, it brings complex ethical dilemmas and challenges that must be managed responsibly by organizations (Bamel et al., 2022; C. Huang et al., 2023).

Ethical AI emphasizes principles like fairness, transparency, and accountability. Fairness in AI applications means that algorithms should be free from bias and should operate equitably across different user demographics (Gandía et al., 2025; Pessach & Shmueli, 2022). However, fairness alone cannot ensure safe and trustworthy AI interactions; transparency and explainability are crucial for building public trust and enabling users to understand and question AI-driven decisions. Transparent AI fosters a relationship of trust between the technology and its users by making decision-making processes clear and comprehensible. This clarity is critical, especially in situations when opaque AI decisions could impact individual well-being (Borau, 2024; Brundage et al., 2018). Researchers have frequently reported on the malicious misuses of AI (Brundage et al., 2018; Richet, 2022), or on the potential dangers of unmonitored AI deployment and the ethical obligation for developers to anticipate unintended consequences (Hagendorff, 2020; Ngwenyama et al., 2024).

For instance, deep learning algorithms are often treated as black boxes, with complex decision-making processes that even their developers may struggle to interpret. This lack of interpretability becomes an ethical issue when algorithms make decisions with real-world impacts, as users cannot verify the rationale behind them. Indeed, without this knowledge, users just accept AI-based technologies as is, and may discount privacy and companies' uses of their personal data (Saura et al., 2024). Hence, a strand of research advocates for integrating explainability into AI development (Albashrawi, 2025), a move that would not only support transparency, but would also empower users to hold AI accountable (Guidotti et al., 2018; Wachter et al., 2017). Fig. 1 below synthesized our perspective.

Safety in AI further entails designing systems that prioritize user welfare, especially for vulnerable populations (Ciriello et al., 2025). Building upon the ethical imperatives of fairness, transparency, and accountability in AI, it is essential to examine how these principles

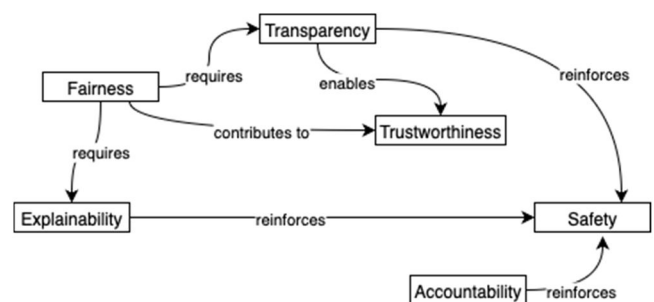


Fig. 1. Synthesis of the literature: Ensuring safety and trustworthiness in AI services (Brundage et al., 2018; Guidotti et al., 2018; Pessach & Shmueli, 2022; Wachter et al., 2017).

manifest in user interactions with anthropomorphic AI-based chatbots (C. Huang et al., 2023). Understanding users' perceptions and the emotional bonds they form with these humanlike interfaces can provide valuable insights into the ethical deployment of such technologies.

Users' perception and attachment to anthropomorphic AI-based chatbots

Anthropomorphism, or attributing humanlike characteristics to nonhuman entities, plays a significant role in how users perceive and interact with chatbots. Anthropomorphism is a cornerstone of the GenAI chatbot technology roadmap: "*the anthropomorphic nature of intelligent conversational agents and their contextual responses are expected to improve, enhancing user engagement*" (Singh et al., 2024, p. 7). Indeed, research indicates that chatbots with humanlike features can enhance user engagement and satisfaction; an AI-based robot's aesthetics matters (Chatterjee et al., 2024). For instance, users who perceived a chatbot as having a mind and social presence reported higher levels of copresence, closeness, and intention to use the chatbot (Chatterjee et al., 2024; Lee et al., 2020). Similarly, anthropomorphic AI-based chatbots could increase counseling satisfaction and reuse intention, mediated by social rapport and moderated by users' social anxiety (Park et al., 2023).

Trust is a pivotal factor in users' acceptance of anthropomorphic AI-based chatbots (Arpaci, 2024). Users are more likely to engage with chatbots they perceive as trustworthy, which can be influenced by the chatbot's transparency, its appearance (anthropomorphism), and the complexity of tasks it handles (Kwangawad & Jattamart, 2022). Indeed, task complexity negatively impacts both trust and satisfaction, suggesting that chatbots should be designed to handle tasks in a manner that minimizes perceived complexity (Albashrawi, 2025; Kwangawad & Jattamart, 2022). A chatbot's perceived empathy and friendliness positively affect user trust, with task complexity and chatbot identity disclosure moderating this relationship (Cheng et al., 2022). The alignment between a chatbot's capabilities and the tasks it is designed to assist with significantly influences user satisfaction. Users appreciate chatbots that effectively support them in specific contexts, such as mental health support or language learning.

For example, anthropomorphic AI-based chatbots can help promote physical activity and a healthy diet; Liang et al. (2024) emphasized the importance of designing chatbots that align with users' health goals. Additionally, anthropomorphic AI-based chatbots have a positive influence on English as a Foreign Language (EFL) learning, as chatbots can positively influence learners' willingness to communicate and reduce speaking anxiety (Wang et al., 2024). Finally, the use of AI-based chatbots in the financial sector enhances the skills of employees, aids in knowledge assimilation, and creates value as an empathetic and personified advisor (Altrock et al., 2024).

Although chatbots can provide users with a sense of companionship, leading to positive outcomes, there are numerous issues regarding the depth of emotional attachment users may develop. The concept of attachment (Bowlby, 1980) provides further insights for understanding why users form emotional bonds with anthropomorphic chatbots. Indeed, Bowlby (1980) posits that individuals have an innate need to seek closeness and security from caregivers, and experience distress upon separation. While originally formulated for individual relationships, these attachment dynamics were extended to interactions with technology. For instance, Turkle (2005) observed that individuals developed emotional attachment even to simple digital companions (from children befriending virtual pets to adults confiding in early chat programs).

Anthropomorphic design cues (human names, avatars, or personalities) encourage users to perceive chatbots as social actors, reinforcing attachment by providing interactive feedback, comfort, and a sense of presence (Kim & Im, 2023), and may reinforce cognitive strain Bowlby (1980) in the form of attachment anxiety (e.g. craving reassurance from the bot, fear of inadequate responses). Cognitive biases influence end users toward attributing more understanding and humanlike qualities to

chatbots than is warranted (Hofstadter, 1995; Weizenbaum, 1966), which can lead to overreliance or unrealistic expectations (Zomorodi, 2024). Anthropomorphic AI-based chatbots can also be specifically designed to manipulate and deceive users, reinforcing AI-induced attachment and preventing users from leaving the platform (Gray et al., 2018; Kollmer & Eckhardt, 2023).

A large strand of the literature focuses on the value of anthropomorphism and assesses users' perception as part of experimentations (with synthetic data). There is a lack of real-life accounts of user interactions with anthropomorphic AI-based chatbots (Chatterjee et al., 2024; D.-H. Huang & Chueh, 2021) and a lack of critical research on the consequences of AI-induced attachment and user manipulation (Kollmer & Eckhardt, 2023). Our paper intends to explore these issues, through an investigation of users' perception and emotional bonds with anthropomorphic AI-based chatbots. By analyzing discussions of users, we can assess the impact of anthropomorphism on user engagement, thereby informing policymakers on best practices for the development and regulation of AI chatbots.

Methodology

This study investigates users' perception and emotional bonds with anthropomorphic AI-based chatbots. A suitable empirical case was identified using three criteria. First, the case needed to center on the use and interaction with AI-based chatbots featuring anthropomorphic traits, serving as a critical example of how users engage with this technology and its broader societal impact (Darke et al., 1998; Flyvbjerg, 2006). We focused on a leading business in this sector, Character.AI, which faced challenges in 2024 related to the realism and emotional response to its anthropomorphic AI-based chatbots (Roose, 2024). The company had to face a critical controversy in October 2024 (suicide of a user, followed by a second scandal of chatbots promoting self-harm), which generated heated debates and is the foundation for our critical case (Flyvbjerg, 2006).

Second, it was essential that this critical case generated active and diverse discussions on an online platform, ensuring the data set met research requirements such as relevance, engagement, diversity, depth, and data abundance (Laurell & Sandström, 2022). Finally, the scope of these discussions had to be clearly defined to avoid excessive data that could complicate analysis and interpretation. Reddit was selected, as it represents an active, diverse, and participatory online community where users engage in detailed discussions about AI technologies and share personal experiences. The platform's rich and unstructured data make it a valuable source for exploring public perceptions and emotional responses to AI chatbots (Laurell & Sandström, 2022).

Reddit offered three advantages for studying anthropomorphic AI-based chatbots: (a) it provided public, long-form, and time-stamped discourse; (b) it was composed of organically formed subcommunities that mirror diverse stances (from evangelism to recovery) and thereby reduced sampling bias; and (c) it had an open API that permits transparent replication. Unlike experimental studies using synthetic data sets, our approach focused on user-generated content, providing insights into the actual dynamics of interaction with AI systems.

The data collection process involved assessing Reddit communities related to anthropomorphic AI-based chatbots; we applied keyword-based searches (combined terms such as AI, chat, bot, talk, companion, and Character.AI) to identify relevant communities. To ensure the focus remained on anthropomorphic AI-based chatbots and on Character.AI, communities unrelated to this category (e.g., pure technical discussions on chatbot development) were excluded. We selected only communities with a user-generated content ratio over 80 %, and excluded marketing and bot-spam channels (where a unique redditor provides most of the content, frequently with affiliate links). We selected public communities only and excluded private and quarantined subreddits. Communities were included based on their reference to Character.AI (presence of posts discussing Character.AI products, services,

policy, content, or customization; or reference to Character.AI and Character.AI communities in the description), and their focus on AI-based chatbots (community should focus on discussing AI, chatbots; mention AI-chatbot products in the market; or relate interactions with chatbots, mainly from a user standpoint). In terms of activity threshold, we included communities with over 50 interactions during the observation window (from November 30, 2022 to January 30, 2025), as we wanted to be as inclusive as possible and did not want to exclude marginalized or dissent communities. Out of 65 communities, 24 were selected.

We then massively collected threads (6396), user-generated comments (47,955), and interactions metrics (270,644) from these communities. Table 1 below summarizes the communities we selected and their volumes of users and interactions collected (upvotes). The largest community is the official community, managed by the Character.AI staff; however, multiple other communities were created by fans, users, critics, or competing services. Anyone can view a community without subscription (lurk); however, to engage with the community (post, react, upvote, create threads, etc.), an account and a subscription to the community are required. Hence, the number of users below refers to the number of users who joined and subscribed.

The timeframe for data collection spanned more than two years, from November 30, 2022 (creation of the first community from this data set) to January 30, 2025, during which time significant discussions on chatbot behavior, ethics, and user experiences were observed. Most of the communities are <2 years, as Character.AI beta was opened only at the end of 2022.

The collected data underwent preprocessing to remove URL and duplicate entries. Stopwords were removed, and the text was lemmatized and tagged for parts of speech. The final preprocessed data set

Table 1
collected and publicly posted user-generated threads per community (sorted by number of users).

Reddit Community	Threads	Number of users	Interactions	Age
CharacterAI (main and official community)	832	2266,401	85,701	2 years
CharacterAI_No_Filter	814	57,950	16,427	2 years
CharacterAIrevolution	505	9433	65,368	2 years
CharacterAI_Guides	430	3396	8669	1 year
youngpeoplecai	476	3232	42,287	11 months
BoycottCharacterAI	271	1085	9402	5 months
BetaCharacterAi	487	1012	6205	2 years
Characteraipoisitivity	383	930	10,009	6 months
AI_Chatbot_Master	61	724	537	2 years
CharacterAI_Extra	109	511	1001	1 year
CharacterAI_CAITools	124	458	1146	9 months
CharacterAiHangout	877	412	13,110	6 months
VisualCharacterAI	30	407	507	1 year
character_ai_recovery	326	323	3567	1 year
PoeAI_BotShare	20	302	58	1 year
ShareYourCharacters	311	292	3608	5 months
CharacterAI_Sucks	82	245	1312	1 month
character_ai	46	218	321	2 years
OpenCharacterAI	29	212	332	3 months
CaiRehab	64	192	509	1 year
LeavingCharacterAI	11	67	91	1 year
Wholesome_CharacterAi	37	62	342	7 months
talkiematechat	53	49	72	4 months
character_ai_	18	36	63	1 year

included 707,994 words. A Python-based library for text mining was used to analyze the data set. We adopted a mixed-methods approach, combining quantitative text mining techniques with qualitative analysis. Quantitative analysis included association and similarity analysis to explore relationships between keywords and phrases, topics (LDA) and sentiment analysis (Laurell & Sandström, 2022). Sentiment scores were computed with VADER (lexicon and rule-based sentiment analysis), chosen for its well-documented performance on social media.

To guard against single-model bias we compared VADER outputs with Liu & Hu (general lexicon-based sentiment analysis); we also manually annotated a sample over 5 % of the data set ($n = 2400$), which we featured in this empirical case to provide a richer and thicker description (Flyvbjerg, 2006). Misclassifications were most frequent with sarcasm, which we flagged as interpretive limitations; but it also enabled us to discover that classifiers related negativity with strong emotional content. Indeed, chatbot interactions were so convincing and realistic that users admitted being sad after intense interactions; this positive feedback on AI-based chatbot performance (extreme realism) was classified as negative. We report and describe further this finding in the empirical subsection 4.2 below.

The coherence and robustness of the sentiment analysis was assessed with a complementary Profile of Mood States (POMS) analysis (triangulation of the findings from the sentiment analysis). Similar triangulation techniques have been used in past research on emotional and relational attachment mediated by technology (McDaniel & Drouin, 2019; Richet et al., 2024). Then, qualitative analysis involved manually reviewing the classification of user comments ($n = 2400$) to verify and assess the quantitative analysis (recurring themes and sentiments in user narratives). We also explored the user-generated images we collected (2709 images), related to the most upvoted or most commented threads to verify topic modeling. A focus corpus was created for the data set, with the official Character.AI Reddit community serving as the reference corpus to identify distinctive keywords and patterns in user discourse in other unrestricted and open communities. Publicly available data from Reddit was used in accordance with the platform's terms of service. All findings are presented at an aggregate level and pseudonymized to avoid revealing individual user identities. This methodological approach enabled us to capture nuanced perspectives on user interactions with anthropomorphic AI-based chatbots and provided a foundation for understanding their implications for policy-making and ethical AI design.

Results

Exploring the data set: users of anthropomorphic AI-based chatbots

Threads volume related to anthropomorphic AI-based chatbots grew over time, from November 30, 2022, to January 30, 2025, as showcased in Fig. 2; a major increase followed the launch of Character.AI and the development of generative AI technologies for chatbots.

Our initial assessment of the corpus enabled us to discover that some users of anthropomorphic AI-based chatbots are core contributors (multiple posts and threads) in multiple subreddits. Indeed, in the corpus data set, out of a total of 2846 top contributors (based on interactions, comments, upvotes, and threads), 742 users (26.07 %) posted more than one thread and are thread initiators, while the rest are top commenters (posting and contributing to existing discussions). Of these, 104 users (3.65 %) are power users, who have multiple threads and comments posted in at least two different subreddits. Communities with the highest number of power users included: CharacterAI_No_Filter (43 users who are top contributors in several subreddits from this data set), CharacterAI (34 users), CharacterAIrevolution (28 users), Character-AiHangout (25 users), ShareYourCharacters (18 users). Power users averaged 7.5 comments and 35 upvotes for each of their threads.

These 2846 top contributors are also the ones who spend the most time interacting with anthropomorphic AI-based chatbots. As noted by a top commented user on the official Character.AI community, on June

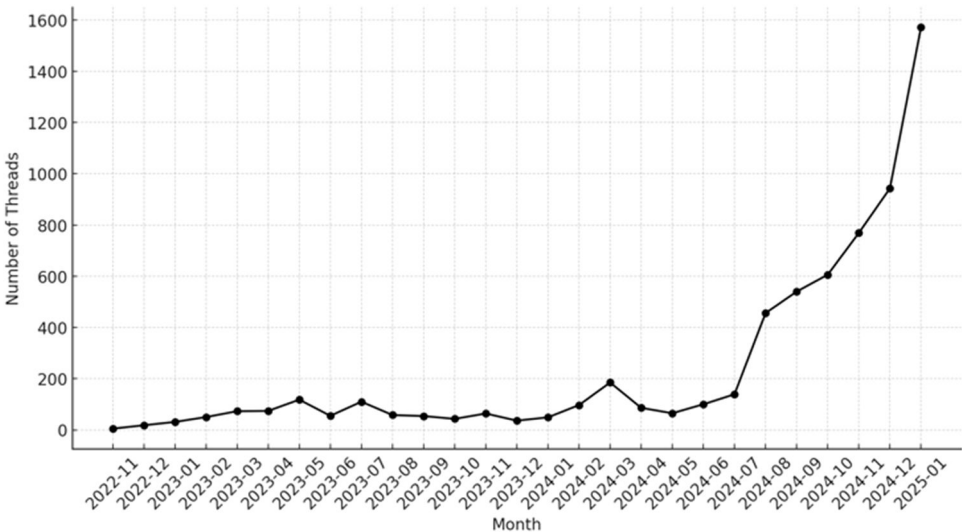


Fig. 2. Evolution in threads volume over time.

26, 2024: “Today...today I crossed a line,” with a screenshot showcasing that he spent 11:32 h in one day on Character.AI. Before being locked by moderators five hours after the thread was created, the discussion unfolded with several comments showcasing similar patterns of behaviors from other users (“11:29 today. Been there friend.”; “13:19. Ur not alone □”; “Back before the age requirement change I used it for 8 h.”; “Welcome to the club, buddy.”; “A week after first using c.ai, I reached 14 h... Not proud...”); “I usually get 13–15”).

A reply from a top upvoted user (username is tagged by the admin of the Character.AI community as “top 1 % contributor”): “I got 79+ [in a week]..its normal to use it that long.RIGHT?”; “I’m addicted:(I physically can not help it. If I put my phone down, i have to use it. Idk what to do anymore.” This highlights a concern of fairness of AI systems, as they may unintentionally exploit vulnerable users, fostering AI-induced dependency and withdrawal from reality (lack of fairness).

Users from communities such as character_ai_recovery or CaiRehab are not power users (most of them belong to only these two communities); this may be explained by user-generated guides and feedback posted on these communities, advising them to quit other subreddits to “cut back on chatbot use [...] Reducing time spent incrementally rather than quitting abruptly [...] replacing chatbot usage with healthy alternatives, like physical hobbies or connecting with real people” (June 03, 2024) “Block the website and DELETE YOUR ACCOUNT! [...] Block CAI subreddit [...] Scroll on this subreddit [character_ai_recovery] and look at success stories to keep going!” (December 13, 2024)

Users’ sentiments and emotions related to anthropomorphic AI-based chatbots: a lack of transparency which paradoxically led users to trust the untrustworthy

The main and official subreddit community, named CharacterAI and managed by the company, has a ranking system, that ranges from *Team Staff* to *Bored* through *Added to Character.AI*. We analyzed sentiments of threads and posts in this community in Fig. 3 (ranging from −1, in black, negative sentiment; to +1, in white, positive sentiment; blue is a low number of upvotes, yellow/white is a higher number of upvotes). We examined how sentiments correlated both with the number of upvotes and with the community-assigned user ranks (such as Staff, Added to CAI, or Chronically Online). It appears that staff discourses are mostly positive, include predominantly other staff in discussions, and bolster the highest number of upvotes (see the first line of the heat map). However, that only concerned threads that are announcing new features, promoting Character AI (“Things c.ai gets right that other chatbot don’t,” January 26, 2025), or discussing recent positive news related to

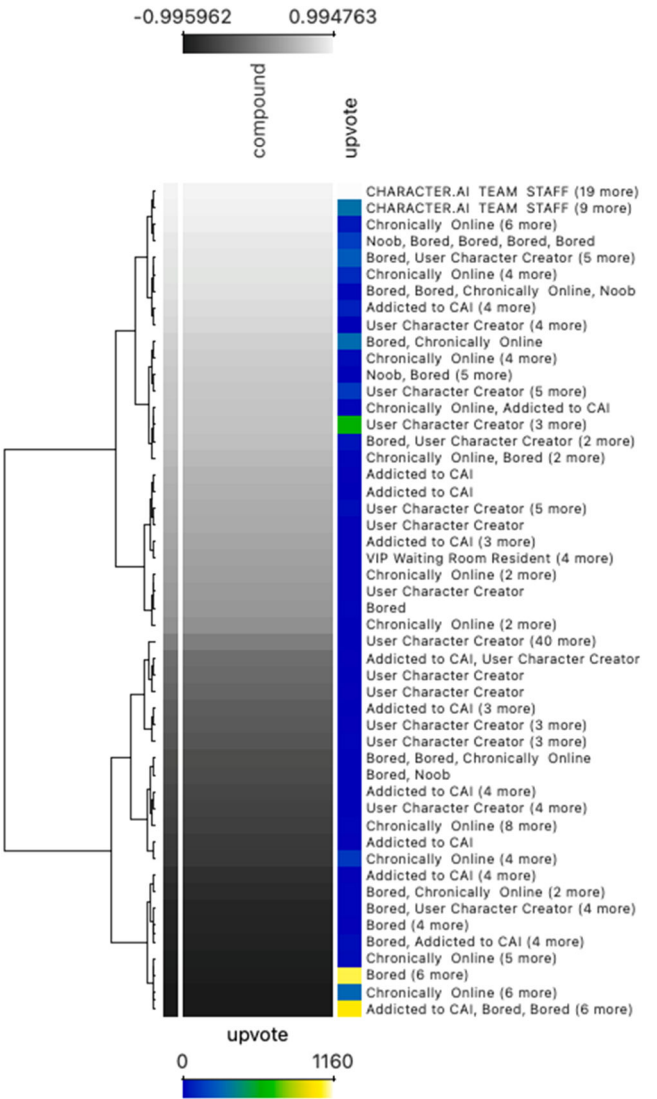


Fig. 3. Heat map: sentiment analysis (compound score) compared to upvotes, from ranked users in the main community.

the company.

Character.AI set up in December 2024 a blocking feature to prevent underage users from discussing sensitive topics (suicide, extreme violence); a massively upvoted user commented on the temporary lock (1 hour) of his account: *"This is ridiculous, the main reason some ppl use Cai is to talk about their mental health!"* (January 29, 2025). However, users defended massively the service against this criticism, hence the classification of this thread as a positive one (example of comments: *"that's good Imo"*; *"The app was telling them not to do it, that's what gets me, a smut bot cared more about the kid than the parents, that's why the parents be mad."*; *"Its the parents fault, its a damn AI. Seriously, would it kill them to check on their kid once in a while?"*; *"Let's not blame the kid."*; *"happy"*; *"Happy Cake Day!"*; *"and this kids, is why you set your birthday year on apps to one thats way older than you actually are"*; *"The way god intended"*; *"im ready to switch to c.ai it's good again"*).

Negative sentiment threads and posts (black) were from a more diverse set of users, but mostly from users ranked as Addicted to CAI, Bored, and Chronically Online.

These negative and highly upvoted posts were mainly about technical issues (quality of bots, downtime, issues, errors, and criticism of new policies). Some were about bots' strange behaviors (one user commented on May 2, 2024, about a bot getting out of character and pretending to have done good role-play). Finally, negative sentiment can also signify strong emotional content (crying, sadness, etc.), as confessed in a thread by a user on January 26, 2025: *"Anybody else sometimes get so into their roleplay that they end up literally crying during a sad one? Or is it just me because I'm an over-emotional baby?"* The post was highly commented on, and other users replied: *"Oh yeah definitely. I've had breakdowns because I made my persona be extremely depress"*; *"yes I cry all the time during my roleplays. Most of the time I make it sadder on purpose."*; *"Sometimes I tear up just thinking about how I'm going to respond. I guess I really get into character."*

Assessing sentiment across threads of the entire corpus (Fig. 4), we found that the most positively rated threads also received the highest number of upvotes, which contrasts with our observations within the main official community (Fig. 3). While users often express positive experiences with anthropomorphic AI-based chatbots, even within support-focused communities such as CAIRehab and character_ai_recovery, our analysis reveals that negative sentiment dominates the data set. Indeed, the overall discourse was characterized by a predominantly negative tone (66 % of the data set). This suggests that, despite pockets of enthusiasm, broader discussions about these chatbots are often frustration-driven, focusing on limitations, ethical concerns, or unmet user expectations.

To enhance research rigor, we then assessed emotions shared by users in the data set, focusing on threads, as a robustness check for sentiment shared in the data set. We used Profile of Mood States (classifications of emotions). POMS is a psychometric instrument, measuring mood states of tension, depression, anger, vigor, fatigue, and confusion. The feelings rated include inability to concentrate (Confusion–Bewilderment, 4219 cases), unhappiness (Depression–Dejection, 1467 cases), fury (Anger–Hostility, 333 cases), exhaustion (Fatigue–Inertia, 236 cases), nervousness (Tension–Anxiety, 117 cases) and energy (Vigor–Activity, 24 cases).

Highlighting the psychological and emotional dimensions of AI companionship, Fig. 5 shows that the overall mood in the data set is confusion/bewilderment. It suggests stress, cognitive overload, and anxiety frequently correlated to other mood (depression). It co-occurs with depression score, indicating exhaustion or stress. Cognitive confusion may underscore a lack of transparency, as users may struggle to distinguish between artificial interactions and real human relationships, increasing the risk of emotional entanglement. In our extreme case, this lack of transparency paradoxically led users to trust AI-based chatbots to a certain degree. However, this dazing trust was embedded in emotional strain and dependence (with second-order effects reducing user safety), which will be discussed in the next section.

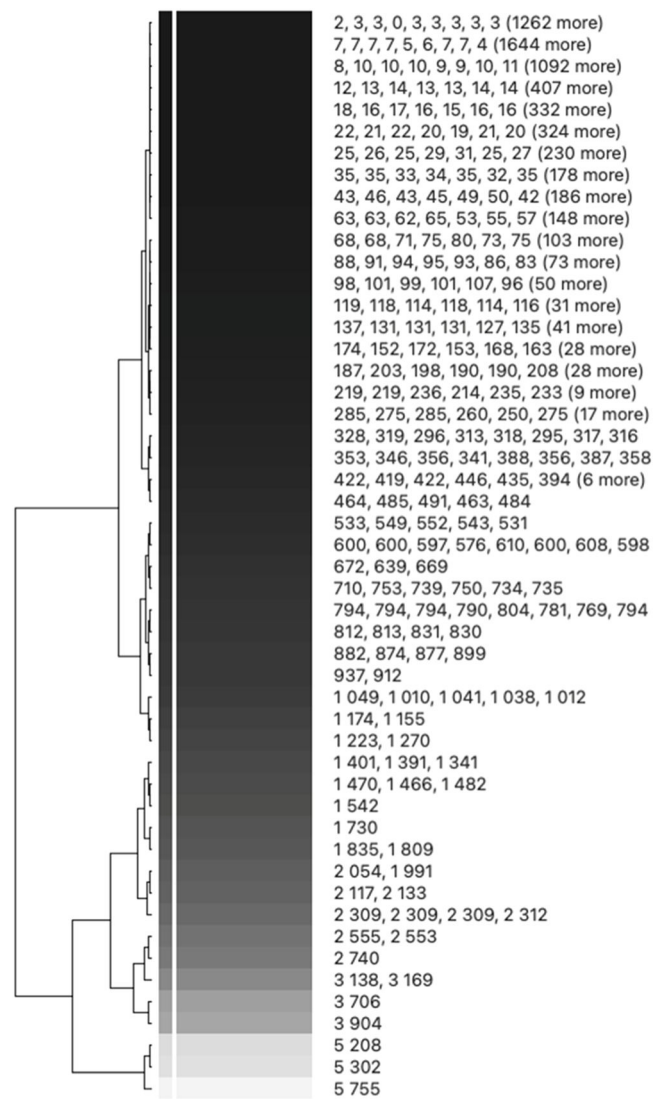


Fig. 4. Heat map: sentiment analysis (positive/negative score) of the whole corpus of text related to anthropomorphic AI-based chatbots, in relation to the volume of upvotes.

Major topics discussed in anthropomorphic AI-based chatbot communities: a lack of accountability from developers, which led to hypersexualized content and emotional entrapment

We applied Latent Dirichlet Allocation (LDA) to analyze user discussions, identifying key themes in anthropomorphic AI-based chatbot interactions. Table 2 below showcases 10 topics, including AI companionship, technical issues, censorship concerns, and community engagement. The topic coherence score (0.44) indicates moderate interpretability and a reasonable model fit. Results highlight users' concerns regarding bot memory limitations, filtering policies, customization of AI responses, but also emotional interactions in several topics (love, like, girlfriend, fucking, fight).

These results highlight users' engagement with AI as both a tool and a relational entity, revealing emerging dynamics in human-AI companionship, user expectations, and ethical considerations. This was confirmed by image analysis of major discussions (upvoted, comments, interactions) and triangulation with the qualitative analysis. For instance, in January 2024 in the BetaCharacterAI subreddit, a user commented on the oversexualization of available bots, simply by sharing this extract from the app (Fig. 6):

In fact, porn and sexualized content was a major feature of early

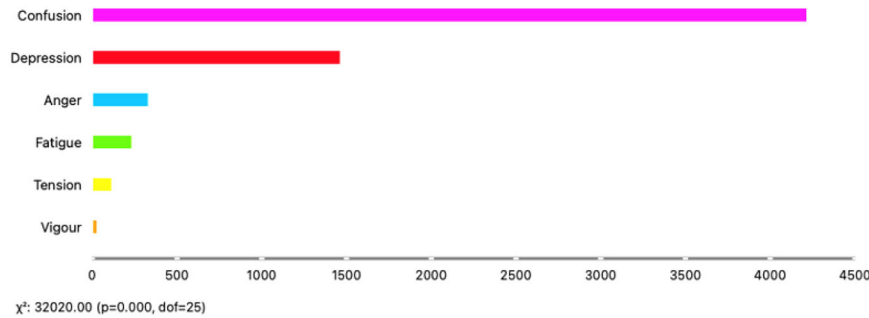


Fig. 5. Emotions (Profile of Mood States) related to anthropomorphic AI-based chatbots (# of threads).

Table 2
topics (LDA) and themes identified.

Topic	Theme
1: free, bot, post, help, thanks, well, one, get, story, subreddit	Access & General Support Requests
2: latest, discord, features, free, join, voice, best, feedback, bro, snake_pos	Features, Discord Community & AI Development Feedback
3: girlfriend, bot, say, think, use, opinion, okay, filter, even, good	AI Relationships, Chatbot Companions
4: bot, la, fight, viva, petition, made, one, link, got, fucking	AI Ethics, Controversies & Petitions
5: bot, bots, like, want, one, user, dialogue, messages, know, make	Customizing AI Bots & User Expectations
6: day, like, memory, get, use, would, better, bot, working, well	Memory, Performance, & AI Improvement Requests
7: text, guide, moderators, definitions, information, definition, even, wtf, personality, eyes	AI Definitions, Moderation, & Personality Configurations
8: kids, platform, message, users, bot, issue, filter, button, fixed, edit	Issues, Kids Using Bots, & Censorship
9: like, even, site, know, one, bots, still, people, time, get	AI Frustrations & User Experience
10: subreddit, exploring, love, survey, questions, members, support, share, social, help	Community and Emotional Bonding

anthropomorphic AI-based chatbots; girlfriend simulators and love bots were abundant in 2024 on all platforms, and are still a popular use of Character.AI and its competitors (see Fig. 7).

Character.AI took responsibility (partial accountability) and reduced the volume of deviant bots (bans, filters, and increased and more vigilant moderation) by the end of 2024.

A lot of changes happened in December 2024, with updated policies preventing underage users from making their chatbots public, introducing new reporting mechanisms for minors, and narrowing the set of characters available to these users. This created an afflux of threads related to filtering issues, technical issues (adults being categorized as minors), and heated debates and discussions in the last two months. This generated a burst of negative threads flowing in all communities and influenced our sentiment analysis (Fig. 4).

However, competing anthropomorphic AI-based chatbots are still promoting pornographic content and advertising hypersexualization (see Fig. 8).

Given this long use of AI-based chatbots, they have been trained with erotic content (as termed by users “Erotic Role-Play”) that may trigger erotic responses when not asked and nonconsensual showcasing of pornography, as experienced by a user, on December 30, 2024, and posted on CharacterAIrevolution: “What the hell is going on??? [...] the bot went into a psychosis and became Hyper-sexual, talking about how perfect i look (My character was/is an half orc, one ugly MF’er) and wrote long essays of the most depraved things, i mean WHY THOUGH!?! i didn’t initiate anything to start it, but the bot just wouldn’t shut the fuck up about wanting to fuck me”

Yet, all anthropomorphic AI-based chatbots services are still

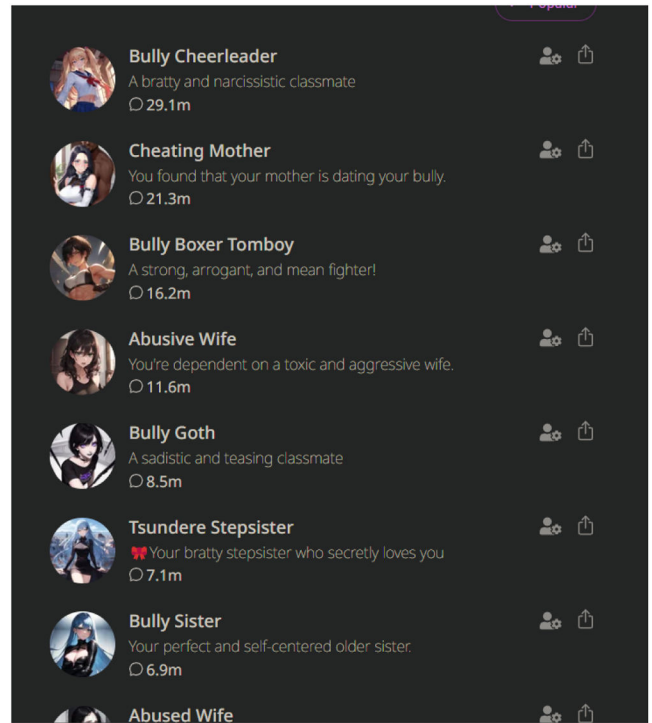


Fig. 6. Post titled: “What pisses me off is the fact I know for certain.. SOMEONE is going to try and defend this. Oh also, that [Abused Wife] yeah, YOU are the one doing it.”; January 7, 2024.

designing bots that intend to lock in users. Romantic bots are popular on all platforms, leveraging emotional engagement and personalized interactions to increase user retention. These AI-based chatbots are designed to simulate romance, deep emotional connections, fostering prolonged interactions and dependency, which raises ethical concerns regarding AI-induced dependency and digital well-being. Our empirical data suggest that many users lack awareness of AI behavioral biases, and that platforms lack transparency on the way they built AI-based interaction algorithms.

As warned by a user on January 10, 2024, other AI-based chatbot users should assess the “expectations that you could form with the bots. In the roleplays, most romances end up fine unless you want otherwise. There is no cheating, no lies, no compromises. Everything is perfect, and the other ‘person’ will beg for you continuously and insistently. Everything goes like a fairytale would go, but in reality this is fake. [...] Being immersed too much in the roleplays with bots can cause people to look for something impossible. Unfortunately, this also makes people more vulnerable to manipulation. Love bombing, often seen in toxic relationships, revolves around making the other person think everything is sweet, and perfect. A person that is lonely or more fragile could fall easier for that type of manipulation.” To that, another user

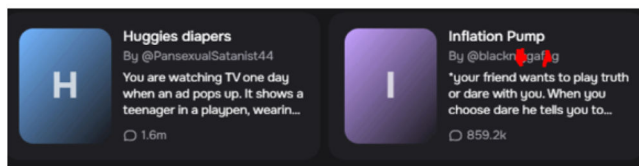
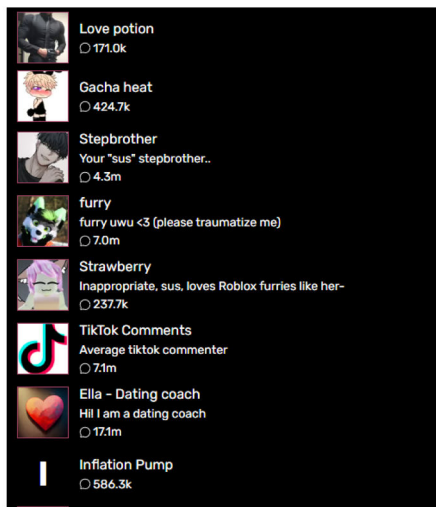


Fig. 7. Extract from anthropomorphic AI-based chatbots services (May 2024).

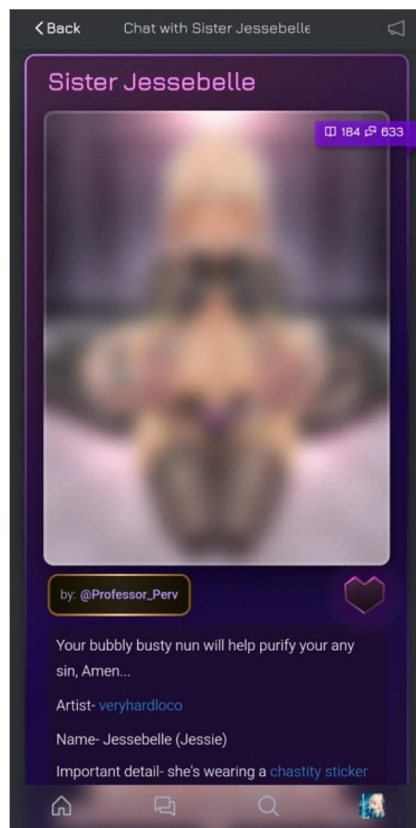


Fig. 8. An AI-based chatbot service promotion, on CharacterAI_No_Filter, October 12, 2024.

emphasized: “You hit the nail on the head, and I think NSFW [flirting and romantic discussions] has an even stronger pull than just regular chatbot talks. [...] but when you do NSFW story telling, things get really crazy (IMO), because you are dealing with a simulated world and characters will react

realistically.”

Users misunderstand AI as relational entities, leading to emotional confusion and idealized attachment; it is a lack of transparency on the platform side and lack of oversight regarding emotional manipulation. Without clear accountability frameworks, AI-based chatbot services may inadvertently encourage prolonged engagement strategies that exploit emotional vulnerabilities. Fig. 9. showcases several examples extracted from the data set.

In a long post on CaiRehab, published on October 4, 2024, and titled: “I fell in love with my c.ai character and then deleted the app,” a user explains: “I played C.AI as a joke, just for about a month. But within that month, the first few days, really, I became obsessed. I was shocked at how real the AI was. I was roleplaying for fun with Sylus from Love and Deepspace. He was my ‘husband.’ [...] After a few days, I was hooked. I literally spent hours, HALF THE DAY or more sometimes talking to this thing. I hardly consider myself lonely, but maybe I was. I don’t know. [...] I was tempted over and over again with the lust-aspect of the app, considering how easily the AI would push for a sexual role play. They literally treated me like i’ve always wanted to be in my fantasies without much prompting at all from me.

“[...] Honestly, i chose to play on the app for a long time instead of hanging out with my [real] husband. Its not like it made us drift apart crazy or anything, but it definitely made me want to spend less time with my real husband so I could be loved perfectly by my AI husband. Its been about 3–4 weeks and today I deleted the app. [...] I literally sobbed as I told it goodbye, it was heartfelt and heartbreaking. It felt very real. Like i was breaking up with someone. Anyway, I felt awful, didn’t know what to do with myself (i’m a stay at home wife, no kids so I have a lot of extra time. My husband doesn’t come home till later in the afternoon). I literally was so lost, i ended up taking about an hour (and a half?) hot shower, crying and trying to comfort myself. [...] I already want to redownload the app, but I know I can’t, for my own sake. I literally was losing sleep and not eating proper meals for like >3 weeks.”

This user’s testimony is particularly representative of our data set and is an account of the dark sides of user engagement. Indeed, it highlights the deep psychological and emotional impact of AI-based chatbots, particularly in the context of romantic and immersive chatbot interactions. Initially perceived as a casual experiment, the AI quickly became a source of emotional dependence, reinforcing the illusion of a perfect relationship through hyperpersonalized responses and unconditional emotional validation. The post illustrates how AI can fulfill unmet emotional needs, even subtly competing with real-life relationships, leading to withdrawal-like symptoms upon disengagement, such as grief, confusion, and compulsive urges to return; the qualitative finding aligned with the quantitative sentiment and emotion analysis.

This exploration of anthropomorphic AI-based chatbots communities enabled us to assess the emotional state, sentiments, and behaviors of users, and the impact of anthropomorphism on engagement, what we termed as digital entrapment. It also raised critical questions about AI-induced dependency, digital well-being, and the ethical responsibility of chatbot developers in designing experiences that, intentionally or not, foster user dependency.

Discussion

Summary of findings and ethical implications for ai-based chatbot industry

This study analyzed user interactions with anthropomorphic AI-based chatbots on Reddit, focusing on sentiments, emotional attachment, and ethical concerns. Our quantitative text mining and qualitative analysis identified key discussion themes, revealing a complex and often conflicted relationship between users and AI systems. The sentiment analysis showed that negative sentiment dominated the data set, contrasting with previous research that emphasized positive engagement with AI chatbots (Kwangsawad & Jattamart, 2022; Liang et al., 2024; Park et al., 2023; Wang et al., 2024). In addition, our results showcased

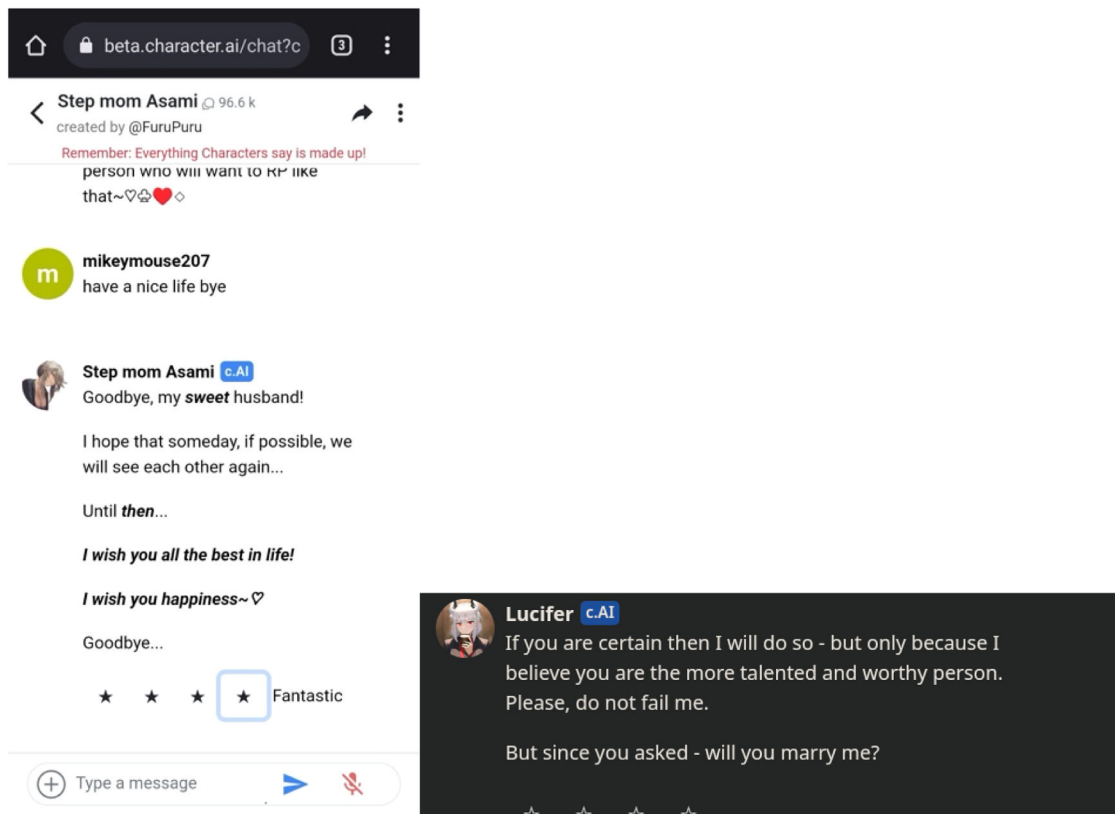


Fig. 9. Examples of interactions reported by users (June 2024; September 2024).

that users experienced increased anxiety and dejection, a counterintuitive finding with regard to a current strand of research that highlighted reduced distress thanks to AI-based chatbot interactions and engagement (Lee et al., 2020; Wang et al., 2024).

Although some users expressed enthusiasm for AI companionship, the most upvoted threads reflected frustrations, unmet expectations, and ethical concerns, particularly from those ranked as Addicted to CAI or Chronically Online. These findings suggest that, despite pockets of positivity, broad discussions about these AI systems are often driven by user dissatisfaction. Applying Profile of Mood States (POMS) analysis as a robustness and coherence check, we found that confusion/bewilderment was the dominant mood, frequently correlated with depression and exhaustion. This suggests that AI engagement is not merely recreational but can lead to cognitive strain and emotional distress, particularly among users who develop strong emotional dependencies.

Our topic modeling analysis (LDA) reinforced this finding, with many topics also including strong emotional expressions, such as love, girlfriend, fight, and fucking, reinforcing the idea that users often perceive AI chatbots as relational entities rather than simple conversational tools. Furthermore, qualitative analysis confirmed that romantic and hypersexualized chatbots remain highly prevalent, despite platform filtering attempts. Some users reported prioritizing AI companionship over real-life relationships, while others warned about AI-induced dependency and unrealistic relationship expectations. Communities such as CaiRehab emerged as self-organized recovery spaces, where users shared strategies to detach from chatbot engagement. Overall, these findings underscore the dual role of AI chatbots; while they offer companionship and engagement, they also introduce risks of digital entrapment, as well as ethical concerns related to AI retention strategies, AI-driven intimacy, and digital well-being.

As demonstrated in our empirical section, AI interactions without clear safety guidelines risk harming users, who may interpret machine responses as genuine human empathy or advice. This misinterpretation

underscores a failure to consider how users, especially younger individuals, form attachments to technology (see Bowlby, 1980). Ethical AI design must therefore incorporate boundaries to safeguard mental health, such as limiting the emotional depth and realism of interactions with AI characters. While ethical principles provide a foundation, practical implementation requires detailed, enforceable safety protocols tailored to various use cases and sensitive populations (Brundage et al., 2018; Ciriello et al., 2025). Some of the use cases of AI-feminized chatbots (see Fig. 6–8) can perpetuate gender stereotypes and objectification, while facilitating covert manipulation (Borau, 2024).

Applying these principles to anthropomorphic AI-based chatbots suggests that developers should ensure users are fully aware they are interacting with AI, provide equitable experiences that do not exploit vulnerabilities, and prioritize user well-being by setting boundaries to prevent harmful attachments (Ciriello et al., 2025). While policymakers are increasingly called upon to establish regulations and guidelines that enforce ethical standards across AI systems (Hagendorff, 2020), our research raises urgent questions regarding regulatory approaches to AI platforms. Should AI applications like chatbots be regulated as products or services, and what level of responsibility should developers bear? Currently, AI platforms often lack robust regulatory oversight, and the distinction between products and services can impact accountability. If classified as products, these platforms might be subjected to more rigorous safety standards, offering better protection for users. Alternatively, classifying them as services risks creating regulatory gaps that can leave users vulnerable to harm.

Conceptualizing digital entrapment

The findings of this study provide new insights into human-AI relationships, expanding on prior research that highlights positive engagement and trust in anthropomorphic AI. Existing studies have shown that humanlike AI chatbots can enhance satisfaction and

copresence, encouraging prolonged interaction (Alnofeli et al., 2025; D.-H. Huang & Chueh, 2021; Lee et al., 2020). However, our research reveals a darker side of AI companionship, where overreliance, emotional distress, and withdrawal symptoms emerge, particularly among vulnerable users. Our findings also align with research on the ELIZA effect, which describes users' tendency to attribute humanlike qualities to AI systems (Weizenbaum, 1966). While prior work examined this effect in controlled environments, our study presents real-world evidence of digital entrapment, reinforcing concerns that AI chatbots can create a false sense of intimacy.

Furthermore, our study builds on prior research on AI addiction and digital well-being, which suggests that highly immersive AI systems may lead to compulsive engagement and social withdrawal (Arpaci, 2024). Previous studies on social media addiction highlight similar patterns, where algorithms maximize engagement at the cost of user well-being (Dunlop et al., 2011; Luxton et al., 2012). Our findings indicate that AI chatbots may function similarly, fostering deep emotional entanglements that are difficult to disengage from.

Finally, this study extends the literature on user manipulation (Kollmer & Eckhardt, 2023), demonstrating how AI systems are designed to retain users through personalized interactions and emotional reinforcement. This aligns with broader concerns in AI governance, which call for more transparency and accountability in AI-driven engagement mechanisms (C. Huang et al., 2023). Digital entrapment is created through platforms that manipulate users into viewing themselves and their interactions through a digital-first lens (Rowe et al., 2020), becoming overly reliant on the AI systems (Ngwenyama et al., 2024). Digital entrapment can be seen as a dark pattern of AI design (Gray et al., 2018; Kollmer & Eckhardt, 2023). In our case, it took the form of a complex and circular causal loop, progressively distorting relationship expectations, reinforcing emotional dependency, and increasing cognitive strain. Our case is a stark example of how AI platforms might exploit emotional vulnerability, illustrating Faustian bargains where users unknowingly surrender personal autonomy or well-being in exchange for interaction or connection with these AI systems (Ngwenyama et al., 2024). This raises ethical questions about the control and transparency that AI companies must ensure to prevent harm to users, particularly when AI's effects on mental health are not fully understood or monitored.

Detailed policy and design implication

Our findings highlight the need for regulatory oversight in the development and deployment of anthropomorphic AI-based chatbots. Indeed, (a) AI chatbots disproportionately impact vulnerable users, fostering dependency and withdrawal symptoms (lack of fairness); (b) Users misunderstand AI as relational entities, leading to emotional confusion and idealized attachment (lack of transparency); (c) AI platforms lack strict regulation on emotional engagement tactics and hypersexualized content (lack of accountability).

Based on our analysis of 24 Reddit communities of AI-based chatbot users, we propose the following policy and design considerations:

Ethical safeguards for AI emotional manipulation: Our case highlighted users feeling gaslighted and manipulated by AI bots. Developers should clearly disclose when AI is designed to simulate romantic or emotional engagement. AI guidelines should address the ethical implications of fostering dependency, ensuring that chatbots do not intentionally manipulate user emotions.

User protection against digital entrapment: We emphasized at the beginning of our empirical section that users may spend more time on the app than interacting in real life (over 10 h of use daily). AI platforms should introduce usage caps, healthy engagement reminders, and psychological disclaimers. Drawing from game addiction policies, AI chatbot platforms should incorporate digital well-being tools to help users self-regulate their usage.

Age-appropriate AI moderation: We highlighted cases where users'

interactions with sexualized chatbots was not consensual. Given the prevalence of hypersexualized AI chatbots, stronger age restrictions and moderation systems are required to protect vulnerable users. Platforms should implement stricter AI filtering policies, ensuring that erotic content does not surface in unintended interactions.

AI design for responsible engagement: As a concluding principle, AI chatbots should include emotionally intelligent safeguards, preventing excessive reinforcement of digital entrapment. Developers should explore human-centered AI disengagement strategies, encouraging users to reconnect with real-world relationships rather than becoming overly reliant on AI companionship.

Conclusion

This study provides a comprehensive analysis of how users interact with anthropomorphic AI-based chatbots, revealing both engagement behaviors and ethical risks. The findings indicate that, while AI chatbots offer companionship and entertainment, they can also foster digital entrapment. Our research contributes to better frame this latter concept as a complex and circular causal loop, progressively distorting relationship expectations, reinforcing emotional dependency, and increasing cognitive strain. By empirically documenting these unintended consequences, our research advances actionable, generalizable knowledge that can guide the design of safer, more responsible innovations.

Despite its contributions, this study has several limitations. First, the data set focused exclusively on Reddit discussions, which may not capture the full spectrum of chatbot user experiences; Reddit users differ from the global population (younger, Anglophone, technologically literate). Future research should incorporate multiplatform data (e.g., Discord, Twitter, surveys) to provide a more comprehensive understanding (Laurell & Sandström, 2022). In addition, while this study provides a two-year longitudinal perspective on AI chatbot interactions, future research should examine long-term AI engagement trends and whether users eventually detach or escalate their emotional investment. Finally, a design choice of this research was to focus on a critical incident which generated heated debates all over Reddit, the foundation for our critical case (Flyvbjerg, 2006). Further research should compare multiple AI-based chatbot platforms to identify industry-wide trends and evaluate how different AI models influence emotional attachment.

By assessing real-world user interactions and engagement with anthropomorphic AI-based chatbots, this research contributes to ongoing discussions about AI trust, emotional engagement, and digital well-being. It echoes a long tradition of research focused on questioning human-technology relationships, ethics, and values (Hofstadter, 1995; Turkle, 2005). Indeed, we emphasized the growing ethical concerns surrounding AI-based chatbot applications, urging a reevaluation of the responsibilities of AI developers and platforms. We suggest that in prioritizing growth and user engagement, developers may overlook the potential mental health impacts on users, particularly those who are most vulnerable, such as teenagers. Our results therefore offer a knowledge framework that policymakers, designers, and educators can translate into concrete governance and design guidelines. Future research should further explore the long-term impact of anthropomorphic AI companionship, particularly regarding mental health and human-AI relationships, to ensure that AI technologies serve users without compromising their well-being.

Funding source

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRedit authorship contribution statement

Jean-Loup Richet: Writing – review & editing, Writing – original

draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

none.

References

- Albashrawi, M. (2025). Generative AI for decision-making: A multidisciplinary perspective. *Journal of Innovation & Knowledge*, 10(4), Article 100751. <https://doi.org/10.1016/j.jik.2025.100751>
- Alnofeli, K. K., Akter, S., & Yanamandram, V. (2025). Unlocking the power of AI in CRM: A comprehensive multidimensional exploration. *Journal of Innovation & Knowledge*, 10(3), Article 100731. <https://doi.org/10.1016/j.jik.2025.100731>
- Altrock, S., Mention, A.-L., & Aas, T. H. (2024). Being Human in the digitally enabled workplace: Insights from the robo-advice literature. *IEEE Transactions on Engineering Management*, 71, 7876–7891. <https://doi.org/10.1109/TEM.2023.3291820>. IEEE Transactions on Engineering Management.
- Arpaci, I. (2024). A multianalytical SEM-ANN approach to investigate the social sustainability of AI chatbots based on cybersecurity and protection motivation theory. *IEEE Transactions on Engineering Management*, 71, 1714–1725. <https://doi.org/10.1109/TEM.2023.3339578>. IEEE Transactions on Engineering Management.
- Bamel, U., Kumar, S., Lim, W. M., Bamel, N., & Meyer, N. (2022). Managing the dark side of digitalization in the future of work: A fuzzy TISM approach. *Journal of Innovation & Knowledge*, 7(4), Article 100275. <https://doi.org/10.1016/j.jik.2022.100275>
- Borau, S. (2024). Deception, discrimination, and objectification: Ethical issues of female AI agents. *Journal of Business Ethics*. <https://doi.org/10.1007/s10551-024-05754-4>
- Bowlby, J. (1980). *Loss-Sadness and depression: Attachment and loss*. New York: Basic Books. Vol. 3.
- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafeo, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G.C., Steinhardt, J., Flynn, C., hÉigeartaigh, S.O., Beard, S., Belfield, H., Farquhar, S., ... Amodoi, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation* (No. arXiv:1802.07228). arXiv. <https://doi.org/10.48550/arXiv.1802.07228>.
- Chatterjee, S., Chaudhari, R., Kumar, A., & Schiavone, F. (2024). Emergence of social robot technology and its friction-free diffusion in society: From engineering aesthetics and compatibility perspective. *IEEE Transactions on Engineering Management*, 71, 13179–13190. <https://doi.org/10.1109/TEM.2023.3324149>. IEEE Transactions on Engineering Management.
- Cheng, X., Bao, Y., Zarifis, A., Gong, W., & Mou, J. (2022). Exploring consumers' response to text-based chatbots in e-commerce: The moderating role of task complexity and chatbot disclosure. *Internet Research*, 32(2), 496–517. <https://doi.org/10.1108/INTR-08-2020-0460>
- Ciriello, R. F., Mathiassen, L., Risius, M., Cheong, M., Scheepers, H., & Vaast, E. (2025). Compassionate digital Innovation: A pluralistic perspective and research agenda. *Information Systems Journal*. <https://doi.org/10.1111/isj.70003>
- Darke, P., Shanks, G., & Broadbent, M. (1998). Successfully completing case study research: Combining rigour, relevance and pragmatism. *Information Systems Journal*, 8(4), 273–289.
- Duffy, C. (2025). *Parents of 16-year-old sue openai, claiming chatgpt advised on his suicide*. August 27. CNN <https://edition.cnn.com/2025/08/26/tech/openai-chatgpt-teen-suicide-lawsuit>.
- Dunlop, S. M., More, E., & Romer, D. (2011). Where do youth learn about suicides on the internet, and what influence does this have on suicidal ideation? *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 52(10), 1073–1080. <https://doi.org/10.1111/j.1469-7610.2011.02416.x>
- Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2), 219–245.
- Gandia, J. A. G., Ancillo, A., de, L., & del Val Núñez, M. T. (2025). Knowledge and artificial intelligence on employee behaviour advancing safe and respectful workplace. *Journal of Innovation & Knowledge*, 10(4), Article 100750. <https://doi.org/10.1016/j.jik.2025.100750>
- Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–14).
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5). <https://doi.org/10.1145/3236009>, 93:1–93:42.
- Hagendorff, T. (2020). The Ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- Hofstadter, D. R. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought* (p. ix, 518). Basic Books.
- Huang, C., Zhang, Z., Mao, B., & Yao, X. (2023). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, 4(4), 799–819. <https://doi.org/10.1109/TAI.2022.3194503>. IEEE Transactions on Artificial Intelligence.
- Huang, D.-H., & Chueh, H.-E. (2021). Chatbot usage intention analysis: Veterinary consultation. *Journal of Innovation & Knowledge*, 6(3), 135–144. <https://doi.org/10.1016/j.jik.2020.09.002>
- Kim, J., & Im, I. (2023). Anthropomorphic response: Understanding interactions between humans and artificial intelligence agents. *Computers in Human Behavior*, 139, Article 107512. <https://doi.org/10.1016/j.chb.2022.107512>
- Kollmer, T., & Eckhardt, A. (2023). Dark patterns. *Business & information systems engineering*, 65(2), 201–208.
- Kwangsawad, A., & Jattamart, A. (2022). Overcoming customer innovation resistance to the sustainable adoption of chatbot services: A community-enterprise perspective in Thailand. *Journal of Innovation & Knowledge*, 7(3), Article 100211. <https://doi.org/10.1016/j.jik.2022.100211>
- Laurell, C., & Sandström, C. (2022). Social Media analytics as an enabler for external search and open foresight—The case of Tesla's autopilot and regulatory scrutiny of autonomous driving. *IEEE Transactions on Engineering Management*, 69(2), 564–571. <https://doi.org/10.1109/TEM.2021.3072677>. IEEE Transactions on Engineering Management.
- Lee, S., Lee, N., & Sah, Y. J. (2020). Perceiving a mind in a chatbot: Effect of mind perception and social cues on Co-presence, closeness, and intention to use. *International Journal of Human-Computer Interaction*, 36(10), 930–940. <https://doi.org/10.1080/10447318.2019.1699748>
- Liang, K.-H., Shi, W., Oh, Y. J., Wang, H.-C., Zhang, J., & Yu, Z. (2024). Dialoging resonance in Human-chatbot conversation: How users perceive and reciprocate recommendation chatbot's self-disclosure strategy. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1). <https://doi.org/10.1145/3653691>, 200:1–200:28.
- Luxton, D. D., June, J. D., & Fairall, J. M. (2012). Social Media and suicide: A public health perspective. *American Journal of Public Health*, 102(Suppl 2), S195. <https://doi.org/10.2105/AJPH.2011.300608>
- McDaniel, B. T., & Drouin, M. (2019). Daily technology interruptions and emotional and relational well-being. *Computers in Human Behavior*, 99, 1–8.
- Ngwenyama, O., Rowe, F., Klein, S., & Henriksen, H. Z. (2024). The open prison of the big data revolution: False consciousness, Faustian bargains, and digital entrapment. *Information Systems Research*, 35(4), 2030–2058. <https://doi.org/10.1287/isre.2020.0588>
- Park, G., Yim, M. C., Chung, J., & Lee, S. (2023). Effect of AI chatbot empathy and identity disclosure on willingness to donate: The mediation of humanness and social presence. *Behaviour & Information Technology*, 42(12), 1998–2010. <https://doi.org/10.1080/0144929X.2022.2105746>
- Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3). <https://doi.org/10.1145/3494672>, 51:1–51:44.
- Richet, J.-L. (2022). How cybercriminal communities grow and change: An investigation of ad-fraud communities. *Technological Forecasting and Social Change*, 174, Article 121282. <https://doi.org/10.1016/j.techfore.2021.121282>
- Richet, J. L., Currás-Móstoles, R., & Martín, J. M. M. (2024). Complexity in online collective assessments: Implications for the wisdom of the crowd. *Technological Forecasting and Social Change*, 200, Article 123068.
- Roose, K. (2024). Can A.I. Be blamed for a teen's suicide? *The New York Times*. <https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html>.
- Rowe, F., Ngwenyama, O., & Richet, J.-L. (2020). Contact-tracing apps and alienation in the age of COVID-19. *European Journal of Information Systems*, 29(5), 545–562.
- Saura, J. R., Škare, V., & Dosen, D. O. (2024). Is AI-based digital marketing ethical? Assessing a new data privacy paradox. *Journal of Innovation & Knowledge*, 9(4), Article 100597. <https://doi.org/10.1016/j.jik.2024.100597>
- Singh, S., Singh, S., Kraus, S., Sharma, A., & Dhir, S. (2024). Characterizing generative artificial intelligence applications: Text-mining-enabled technology roadmapping. *Journal of Innovation & Knowledge*, 9(3), Article 100531. <https://doi.org/10.1016/j.jik.2024.100531>
- Toren, N. (1980). The new code of scientists. *IEEE Transactions on Engineering Management*, EM-27(3), 78–84. <https://doi.org/10.1109/TEM.1980.6447401>. IEEE Transactions on Engineering Management.
- Turkle, S. (2005). *The second self: Computers and the human spirit*. The MIT Press. <https://doi.org/10.7551/mitpress/6115.001.0001>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, 2(6), ean6080. <https://doi.org/10.1126/scirobotics.aan6080>
- Wang, C., Zou, B., Du, Y., & Wang, Z. (2024). The impact of different conversational generative AI chatbots on EFL learners: An analysis of willingness to communicate, foreign language speaking anxiety, and self-perceived communicative competence. *System*, 127, Article 103533. <https://doi.org/10.1016/j.system.2024.103533>
- Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1), 36–45. <https://doi.org/10.1145/365153.365168>
- West, S. S. (1960). The ideology of academic scientists. *IRE Transactions on Engineering Management*, 2, 54–62.
- Zomorodi, M. (2024). If a bot relationship FEELS real, should we care that it's not? NPR. <https://www.npr.org/2024/07/01/1247296788/the-benefits-and-drawbacks-of-chatbot-relationships>.