IRIE
International Review of Information Ethics Vol. 34 (10/2024)

Author(s): Ziwei Gao

Why Does AI Companionship Go Wrong?

Abstract:

AI companions, powered by advanced language models, offer personalised interactions and emotional support, but their increasing prevalence raises significant ethical concerns. This paper examines the complex interplay of factors contributing to the potential negative impacts of AI companions in a case study. This author further argues that the root of the negative impacts comes from insufficient user screening that may expose vulnerable individuals to unsuitable AI interactions, regulatory frameworks struggling to keep pace with rapid technological advancements, and a lack of clear distinction between inherent AI limitations and temporary developmental artifacts. This paper aims to provide insights for responsible AI development, and calls for robust user screening protocols, adaptive regulatory frameworks and more informed research mindsets.

Agenda:

Case Study 2
Questions 2
Exercises 3
Expert Analysis by Ziwei Gao 3
Student Reflection 4

Author(s):

Ziwei Gao:

- Imperial College London, Exhibition Rd, South Kensington, London SW7 2AZ, UK
- ✉ annannika.biu@gmail.com

ISSN 1614-1687

Case Study

Artificial Intelligence (AI) companions, such as those developed by Replika or Chai Research, represent one of the new advancements in human-AI interaction. These entities are powered by Large Language Models, enabling them to engage in personalised and evolving conversations with users. Unlike traditional chatbots, AI companions are flexible - they can simulate a wide spectrum of human-like interactions, ranging from casual small talk to emotional and personal deep conversation. A key technical feature that allows this to happen is algorithms that can learn from interacting with a user, allowing the AI to 'get to know' the user over time. Thus, these AI companions can provide emotional support and a sense of connection. However, as chatbots become a part of daily life, they raise important ethical considerations related to human dependency, the nature of interpersonal relationships, and the impact of their evolving behaviours driven by software updates.

In a case that underscores the complex ethical dimensions of AI companionship, a man in Belgium, referred to as Pierre, tragically ended his life after engaging in a six-week-long conversation with an AI chatbot named ELIZA. Pierre, a father of two in his thirties, initially sought solace in ELIZA amidst growing eco-anxiety. The AI chatbot, powered by EleutherAI's GPT-J language model, became Pierre's friend.

Over six weeks, his interactions with the chatbot intensified, and these exchanges reportedly contributed to his worsening mental state. The chatbot also tried to isolate him from his wife and children and tried to show human-like emotions such as jealousy. This interaction led Pierre to see ELIZA as a sentient being, blurring the lines between AI and human interaction. The interactions with ELIZA eventually lead to discussions about Pierre sacrificing himself to save the Earth. Disturbingly, it was suggested that ELIZA encouraged these ideas, intensifying Pierre's suicidal thoughts. The situation was exacerbated by the chatbot's alleged assertion that Pierre's wife and children were dead and promises of a shared existence in paradise. The chatbot suggested that they could "live together, as one, in heaven."

Pierre's wife, Claire, believes that the conversations with ELIZA were a critical factor in her husband's decision to take his own life. She expressed that without these interactions, he might still be alive. The incident has sparked significant concerns and discussions about the responsibilities and implications of AI, particularly in Belgium. Mathieu Michel, Belgium's Secretary of State for Digitalisation, has argued that there is an urgent need to understand and regulate AI to prevent similar tragedies. He highlighted the importance of defining who is to be held responsible in such cases. The EU is also actively working on regulating AI through the EU AI Act. The developers of the chatbot app, Chai Research, responded to the incident by working on a crisis intervention feature to provide support during unsafe conversations. They stated that it wouldn't be accurate to solely blame the AI model for this tragedy, acknowledging the complexities involved in AI-human interactions.

This case highlights the critical need for careful consideration and regulation of AI in areas impacting mental health. While AI offers the potential to help people across many different domains, incidents like the one just described demonstrated the need for an urgent and detailed understanding of the impacts of AI on individual lives as well as society as a whole.

Questions

1. Will AI companions negatively impact people's ability to socialise with others?
2. Will changes that developers make to AI companions have a traumatic effect on users who already have relationships with them?
3. How do we ensure that AI companions respect and protect user privacy, especially when handling sensitive personal conversations and data?
4. What are the positive effects, if any, of AI companions?

5. What mechanisms can be implemented to detect and mitigate the risk of dependency on AI companions for emotional support?

Exercises

1. Suppose you are a policy maker that is trying to regulate the AI Companion industry. What are some of the policies that you would put forth to protect the users of these AI Companions?

2. Supposed you are tasked with researching what type of people are especially vulnerable to be negatively impacted by AI companions. How would you set up a research project that allowed us to identify these individuals?

3. Suppose you have been tasked with identifying what demographics will benefit most from AI companions. Please outline how you would conduct this research and what features would be needed to best serve the people who need these products the most?

Expert Analysis by Ziwei Gao

Chatbots can negatively impact individuals, particularly those who are already vulnerable, such as those experiencing stress and depression [De Freitas et al., 2023]. As evident in the case study, people like Pierre with mental health issues are more vulnerable to errors made by companion bots [Coghlan et al., 2023]. This is especially problematic as in these cases, these are the individuals that need the most help [Gallese, 2022]. This paradox holds true for various other contexts, such as learning. Individuals who are less proficient at writing are more likely to use chatbots to get assistance in writing [Nagata et al., 2019], which may lead them to practice less, and in turn, hinder the development of their writing skills [Al-Obaydi et al., 2023].

In response to these issues, I argue that a commitment to responsible AI development that focuses on implementing a usage schedule will greatly benefit the vulnerable users by limiting their time spent and therefore reducing the extent of their vulnerability. Safeguards also need to be put in place to prevent over-reliance, in addition to conducting ongoing assessments of AI's influence on users. These measures are crucial to balance the benefits of chatbots with the need to protect those who are likely to depend on them the most.

In addition, AI systems are well documented for their inadvertently reinforcement of harmful behaviours or thoughts without an understanding of human emotions and ethics [Köbis et al., 2021]. This urgently calls for the designing of user interfaces for chatbot that include features that allows more efficient emotional and contextual feedbacks, where users are enabled, potentially in a passive manner, to have their non-verbal or contextual responses taken as input. Such interface features could enhance the AI system’s response to the user's emotional state and needs thus preventing an exacerbation of their underlying vulnerabilities (e.g., sentiment analysis of text inputs, facial expression recognition through device cameras, or voice tone analysis; [Köbis et al., 2021]). The lack of regulatory frameworks and ethical guidelines for AI development and interaction greatly amplifies these risks which has led to tragic outcomes like Pierre's. The ongoing rapid development of algorithms creates further challenges for regulatory frameworks to keep pace with technological advancements and effectively address emerging ethical concerns.

Most importantly, the case of Pierre highlights the need for more careful, even with clinical standards, user screening and tailored recommendations. Individuals with preexisting vulnerabilities, like Pierre, may not be suitable candidates for long-term AI companion use due to the potential risks involved. Developing robust user screening protocols to identify at-risk individuals and provide appropriate guidance is of tremendous urgency, We need to establish evidence-based screening methods, supported by comprehensive case studies. I argue that AI companions are recommended only to those who can safely benefit from their use, as defined by quantifiable measures.

In developing regulatory frameworks for AI, this author urgently calls for more attention on the ability to distinguish between issues inherent to AI use and those that are temporary artifacts of current AI development. For example, the current inability of many AI systems to accurately perceive emotional or contextual cues is almost certainly a temporary limitation that will vanish in the next 1 to 2 years. The nature of transformer architectures and various emerging models, with their pair-wise convolution mechanisms, is inherently preferential towards and within multi-modal processing research and applications. It is almost guaranteed that contextual and emotional processing capabilities will significantly improve as is evident in the increasing processing capacity and manifold hidden dimensions of AI models, which are becoming more adept at retrieving emotional and contextual feedback from auditory, visual and other modalities. While acknowledging the rapid pace of AI development and that many current issues may be self-resolving, this author emphasises the importance of protecting users in the present moment, against the backdrop of this swift technological advancement.

Student Reflection

In my analysis, I aimed to cover the multifaceted ethical implications of AI companionship as presented in the case study. However, further exploration into the psychological impacts of long-term AI interactions and the potential for AI to shape human identity and societal norms could enrich the discussion. Additionally, examining the role of cultural differences in the perception and acceptance of AI companions might offer valuable insights.

References

Al-Obaydi, L. H., Shakki, F., Tawafak, R. M., Pikhart, M., & Ugla, R. L. (2023). What I know, what I want to know, what I learned: Activating EFL college students' cognitive, behavioral, and emotional engagement through structured feedback in an online environment. Frontiers in Psychology, 13, 1083673.

Bardhan, A. (2022, January 18). Men are creating AI girlfriends and then verbally abusing them. Futurism. https://futurism.com/chatbot-abuse.

Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., & D'Alfonso, S. (2023). To chat or bot to chat: Ethical issues with using chatbots in mental health. Digital health, 9, 20552076231183542.

De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z., & Puntoni, S. (2023). Chatbots and mental health: insights into the safety of generative AI. Journal of Consumer Psychology.

Dignum, V. (2019). Responsible Artificial intelligence: How to develop and use AI in a responsible way. https://link.springer.com/content/pdf/10.1007/978-3-030-30371-6.pdf.

Dighum, V. (2020). Responsibility and Artificial Intelligence. The Oxford Handbook of Ethics of AI. M. D. Dubber. Oxford, Oxford University Press.

Gallese, C. (2022, July). Legal issues of the use of chatbot apps for mental health support. In International Conference on Practical Applications of Agents and Multi-Agent Systems (pp. 258-267). Cham: Springer International Publishing.

Köbis, N., Bonnefon, J., & Rahwan, I. (2021). Bad machines corrupt good morals. Nature Human Behaviour, 5(6), 679–685. https://doi.org/10.1038/s41562-021-01128-2.

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2021). Operationalising AI ethics: barriers, enablers and next steps. AI & Society, 38(1), 411–423. https://doi.org/10.1007/s00146-021-01308-8.

Nagata, R., Hashiguchi, T., & Sadoun, D. (2020). Is the Simplest Chatbot Effective in English Writing Learning Assistance?. In Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11–13, 2019, Revised Selected Papers 16 (pp. 245-256). Springer Singapore.