The Artificial Sentience Debate: ELIZA Effect and the AI Manipulation Problem

Corrado Claverini (University of Salento)

The rapid development of generative AI and the numerous news reports of suicides and attempted murders resulting from long conversations with chatbots with increasingly human-like characteristics have led various stakeholders – academics, lawyers, ethicists and civil society – to highlight the risks of anthropomorphising AI. Some go so far as to speak of LaMDA as a sentient AI, and others, in niche online communities, recommend using VOID chat to create emotionally characterisable AI agents with whom one can converse without ethical or legal constraints. There is also the phenomenon of virtual influencers and AI girlfriends/boyfriends with characteristics almost indistinguishable from human ones, elaborate personalities and the ability to communicate empathically. This scenario is fuelling the debate about the ELIZA effect and the need for stricter regulation of these technologies, which can potentially manipulate the most vulnerable. This paper aims to address the ethical concerns related to the use of generative AI systems and the interaction with chatbots capable of influencing people in different directions, focusing on three sets of questions: a) Are the mere transparency requirements that are usually demanded of generative AI sufficient, or should more restrictions be considered that do not hinder innovation? b) How can the risk inherent in generative AI systems to manipulate human behaviour be reduced without limiting the development of these systems? c) And to what extent is the AI developer responsible in the event of adverse consequences for the user?

References (selection)

1. Alves, R. (2023), "Man Says AI Girlfriend 'Encouraged' Royal Assassination Attempt," Medium, July 11, https://medium.com/the-generator/a-chatbot-made-me-do-it-man-blames-ai-girlfriend-for-royal-assassination-attempt-d6410a0ae9aa
2. Cohen, T. (2023), "Regulating Manipulative Artificial Intelligence," SCRIPTed, 20, 1.
3. Deshpande, A., Rajpurohit, T., Narasimhan, K., & Kalyan, A. (2023), "Anthropomorphization of AI: Opportunities and Risks," arXiv:2305.14784, DOI: https://doi.org/10.48550/arXiv.2305.14784
4. Eliot, L. (2023), "Generative AI ChatGPT As Masterful Manipulator Of Humans, Worrying AI Ethics And AI Law," Forbes, March 1, https://www.forbes.com/sites/lanceeliot/2023/03/01/generative-ai-chatgpt-as-masterful-manipulator-of-humans-worrying-ai-ethics-and-ai-law/?sh=7ae5ba0e1d66
5. Kirakowski, J., O'Donnell, P., Yiu, A. (2007), "The Perception of Artificial Intelligence as 'Human' by Computer Users." In: Jacko, J.A. (eds.), Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments, Berlin-Heidelberg: Springer, DOI: https://doi.org/10.1007/978-3-540-73110-8_40
6. Lemoine, B. (2022), "Is LaMDA Sentient?—an Interview," Medium, June 11, https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917
7. Lovens, P.-F. (2023), "Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là," La Libre Belgique, 28 Mars, https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/; "Le chatbot Eliza a brisé une vie : il est temps d'agir face à l'IA manipulatrice," La Libre Belgique, 29 Mars, https://www.lalibre.be/debats/2023/03/29/le-chatbot-eliza-a-brise-une-vie-il-est-temps-dagir-face-a-lia-manipulatrice-BSGGRV7IBRDNROO33EWGFVMWAA/
8. Rosenberg, L. (2023), "The Manipulation Problem: Conversational AI as a Threat to Epistemic Agency," arXiv:2306.11748, DOI: https://doi.org/10.48550/arXiv.2306.11748
9. Smuha, N.A., De Ketelaere, M., Coeckelbergh, M., Dewitte, P. & Poullet, Y. (2023), "Open Letter: We are not ready for manipulative AI – urgent need for action," KU Leuven, 31 March, https://www.law.kuleuven.be/ai-summer-school/open-brief/open-letter-manipulative-ai

10. Switzky, L. (2020), “ELIZA Effects: Pygmalion and the Early Development of Artificial Intelligence,” Shaw, 40, 1, DOI: https://doi.org/10.5325/shaw.40.1.0050