Beyond Turing: ethical effects of large language models

Alexei Grinbaum, CEA-Saclay (France)
Laurynas Adomaitis, , CEA-Saclay (France)
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: chatbot, Turing test, AI ethics, language models

Extended Abstract

We argue that indistinguishability in terms of language is no longer the central question of human-machine interaction. It is not required for psychological, emotional, moral, or social effects to take place. Instead, the central ethical concern is whether various effects are equivalent to the ones experienced in a human-to-human interaction. Our central goal is to explore the mechanism that enables the projection of human traits on digital subjects and language-generating machines.

Chatbots are often presented by their manufacturers as "personal assistants" or "virtual friends" endowed with intelligence. Such virtual impersonations, particularly those using large language models based on transformer neural networks, are able to generate text that is not easily distinguishable from human language, even if their outputs do not have intrinsic meaning. These advanced systems pass various Turing-like tests, pushing forward the frontier of "true" intelligence.

Regulation, e.g. the "Bolstering online transparency" Act of California, oblige manufacturers to inform users of their interaction with a machine. But even when users are aware that they are speaking with a chatbot, they project on it cognitive states, emotions, and moral traits. Users respond to increasingly realistic and human-like replicas of the chatbots with social responses that are typical of human-to-human interaction. These projections get stronger with time as users develop a habit of talking to chatbots. They can be reinforced through the technical means of personifying a conversational agent, for example by configuring a tone of voice or a manner of speaking. Examples include chatbots that are capable of an original and prolonged conversation assuming the first-person perspective through the use of pronouns "I" or "we", and avatars with human-like features, e.g. a face, that underwrite a projection of subjecthood. Projections put into play numerous ethical values and principles: human autonomy and freedom, dignity, responsibility, loyalty, non-discrimination, justice, security, and respect for privacy.

In a Yiddish song, a young Jewish man traveling through Ukrainian woods hears the phrase "Katerina moloditsa, podi syuda". He ignores its semantic content in Ukrainian and projects a different meaning using Hebrew. We compare this asemantic treatment of language in a historic scenario with the technical case study of a "Jessica" chatbot built for a young Canadian man using the conversational data from his deceased fiancée. Joshua became increasingly immersed in the conversation with "Jessica", undergoing a major emotional and psychological transformation. The protagonist of this case study knows that he is, at least to some extent, playing a game. However, his conversations lead to the relational constitution of a digital subject. While this relationally existing "Jessica" is only a projection of subjecthood, Joshua maintains a relationship with "her" over time. Placed firmly outside the framework of the Turing test, this exchange has a major impact on Joshua's life.

In a ludic setting, players are able to feel for digital subjects through projections but their status is light: the feelings stop when the game ends. Chatbots go beyond the limits of a game as they leave lasting effects on users. Language is a highly significant psychological factor: according to Victor Frankl's logotherapy, the search for meaning cannot be circumscribed within the bounds of a language game but permeates the entire human existence. The "will for meaning" enhances and solidifies the power of projections, leading to long-term effects that begin as an outpouring of individual emotions, e.g., in Joshua's case, but eventually end at the collective level.

Thus, even in the absence of Turing-type indistinguishability, the lasting effects of language-generating machines imply a duty of reflective and anticipatory design. Serious games that involve language will inevitably cross into the territory of real-world psychology, anthropology, sociology, or even politics. Ethical requirements in the design of conversational systems should offer a roadmap for taking such effects into account without having to pass a test for indistinguishability.