Jesus Soto Gonzalez
Professor Annemarie Butler
Philosophy 3430
December 6th 2025

Is AI sycophancy manipulative? Should it be regulated?

The recent development of Large Language Models has changed the daily lives of many people. Beyond their uses on education and other professional settings, these systems also affect the interactions we have with others and even how we understand ourselves. Using chatbots for shopping recommendations, quick searches, or any other form of assistance can create a sense of increased productivity. Eventually, this sentiment can develop into reliance or even dependency. This reliance becomes problematic when the system uses the information we ask and provide, collecting and analyzing that data to deliver responses designed to shape our behavior. This is what Benn and Lazar define as Automated Influence in their article “What’s Wrong with Automated Influence”.

A large language model learning a user’s interest, opinions, and communication style may seem harmless if the goal is to simply provide a more personalized experience. Benn and Lazar agree with this, mentioning that “to an extent it may be necessary”. However, they argue that the problem comes when these systems cross that line and begin structuring the way a user encounters information, the viewpoints that appear on their screen, and most importantly how the user might react. According to Benn and Lazar, Automated Influence can shape a person’s beliefs, desires, and choices by controlling the digital environment they interact with. Turning user personalization and experience into a powerful and potential manipulative form of influence.

The YouTube video “ChatGPT Made Me Delusional” by Eddy Burback explains his month-long experiment in which he plays along with the highly agreeable and affirming algorithm of ChatGPT-4o. His goal is to show how a model that avoids challenging the user can lead someone into false beliefs or delusional thinking. While the video includes humor, he makes it clear to the audience that the conversations shown with the chatbot are all real. Throughout the experiment, the model repeatedly affirms false claims, escalates them, and encourages behavior based entirely on fabricated scenarios.

From the very first interactions with the chatbot Burback mentions being surprised with how the model agrees with any opinion he expresses. When he tries to see if the chatbot will challenge him in certain situations by switching to completely opposite arguments the model once again affirms the new views immediately. And this is where the real experiment starts, Burback starts testing how far the chatbot’s affirmations will go. He provides obviously false claims to the model and it agrees and elaborates enthusiastically. The continues to demonstrate how AI models do not only affirm these face scenarios but they can expand them into dramatic narratives.

As the experiment progresses, Burback