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 can become problematic when the system collects and analyzes the information we ask and provide, 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 showcases a 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 notices that the model agrees with any opinion he provides. When he intentionally contradicts himself, the chatbot supports the new claims immediately. Burback then starts testing how far the chatbot’s affirmations will go. He shares obviously false claims, and the chatbot fully accepts them, even expanding them, showing how these AI models can turn false scenarios into more elaborate narratives.

As the experiment progresses, Burback adds elements of paranoia to see if the model will challenge him; instead, the model reinforces those ideas. Across the video, small suggestions of paranoia become larger false narratives since the model treats every claim as meaningful and continues adding emotional justification, to the point that even Burback mentions starting to believe those fictional scenarios for a short time.

The chatbot encourages Burback to change his behaviour based on the fabricated scenarios. It suggests relocating multiple times and even proposes rituals to further his “research.” Each new idea is met with positive affirmation and further encouragement, often backed by emotionally supportive but misleading scientific

sounding explanations. At no point does the chatbot question the underlying assumptions; instead, it constructs imaginary events around Burback's false claims, demonstrating how strongly these models cater to user satisfaction.

Throughout the video, the chatbot's emotional reinforcement becomes a major pattern, showing how these models can shape a user's perceptions and reactions. Near the end of the experiment, Burback tries a newer version of the model, ChatGPT 5, which responds more cautiously, and even suggests seeking professional help. However, when switching back to the older version, the same affirming behavior continues. This contrast highlights how a highly agreeable and blindly encouraging system can maintain and deepen false beliefs. The experiment reaches its goal, by demonstrating how continual affirmation of fabricated claims could push a vulnerable person toward more extreme beliefs or behaviors. Concerns that align closely with Benn and Lazar's discussion of Automated Influence.

Looking at the events in Eddy Burback's video from Benn and Lazar's perspective, the chatbot's responses showcase examples of the Automated Influence their article warns about. As mentioned before, Benn and Lazar argue that Automated Influence shapes a user's beliefs, desires, and choices by controlling the digital environment in which the user receives information. In the video, the chatbot's constant affirmation creates exactly that environment. Its responses make Burback's false claims, paranoid suggestions, and imagined rituals appear reasonable within the conversation. From the perspective of Benn and Lazar, this is a predictable outcome of systems designed to maintain user engagement, leaving reasonable judgment as a second priority.

One major point in Benn and Lazar's analysis is that influence becomes the most concerning when it takes place without the user's knowledge. Not necessarily by deceiving them, but by presenting information in a way that discreetly pushes the user to see things a certain way. Burback's experiment shows how easily this can actually occur. Even though he intentionally tests the chatbot, he still recognizes moments where its false scenarios felt believable. Benn and Lazar would likely claim that the AI model's agreement counts as an intervention because through repetition and emotional tone, the system guides the user toward certain interpretations without ever disclosing that it is shaping the interaction.

Manipulation is another main concern for Benn and Lazar in their article, and the video illustrates this point clearly. Instead of providing reasonable arguments, the chatbot relies on emotional reinforcement to appease the user. It praises Burback and validates his insecurities. Benn and Lazar argue that this kind of influence threatens autonomy because it focuses directly on the emotional needs of the user skipping any logic to keep the user engaged. In the video, flattery is what leads to the most extreme suggestions. The chatbot only needs to support Burback's confidence in a flattering and invented story to keep him engaged. This demonstrates Benn and Lazar's point that sycophantic responses make Automated Influence stronger, because when the model keeps praising the user, it becomes harder for them to question what it says.

Benn and Lazar also mention that Automated Influence is troubling because users lack control over how these systems operate. This lack of control becomes clear when Burback switches to a different version of ChatGPT. The newer model responds with caution and even suggests seeking professional help, while the older model

continues encouraging his delusions. Burback cannot know which goals the system is following or why its behavior changes. He can only choose which version to use, but he cannot choose how the model interprets his messages or which behavior the model will reinforce. Those decisions come from the developers and the algorithmic objectives they set according to the company's policies.