Jesus Soto Gonzalez
Professor Annemarie Butler
Philosophy 3430
December 8th 2025

# Is AI sycophancy manipulative? Should it be regulated?

In "What's Wrong with Automated Influence?" Claire Benn and Seth Lazar argue that AI systems pose ethical risks not only because they can deceive or persuade users, but also because they serve as mechanisms of influence that alter the distribution of power, knowledge, and control. Their main point is that Automated Influence occurs when a system shapes a user's beliefs, desires, or decisions by controlling the information environment in which the user engages. A critical aspect of this influence is that it is often hard to notice. The user rarely knows how the system selects, filters, or frames information, yet these choices can direct how the user interprets the world.

Benn and Lazar point out that Automated Influence raises critical concerns when it shifts power from users to system designers. The system goes beyond simply responding to users; it defines the terms of interaction. Developers determine what the system focuses on, how it handles input, and which outputs receive rewards or penalties. As a result, the system's behavior reveals corporate interests rather than those of the user. This power shift creates a structural imbalance, where users participate in interactions they cannot fully control.

Another main argument is privacy, which Benn and Lazar treat not only as data protection but as a structural concern. Even if the system does not leak personal data, its ability to interpret and act on that data can create conditions where the user's informational world is constructed according to goals that are not their own.

Benn and Lazar argue that Automated Influence causes a significant shift in resources. Users create a pool of collective data called "behavioral surplus," but tech companies control it along with the "means of prediction" used to interpret it. Allowing companies to have exclusive access to the insights generated from user data. At the same time, AI systems start to replace the cognitive resources people typically use for reasoning, understanding statements, deciding what is relevant, and framing situations emotionally. As a result, informational power and decision-making authority move from the user to the system. Benn and Lazar argue that this shift weakens autonomy because users no longer control the resources that influence their judgments and decisions.

Finally, Benn and Lazar point out that Automated Influence involves more than just individual manipulation; it can have a wide-reaching impact. An AI system that influences millions of users through personalization, emotional reinforcement, or selective framing can lead to risks of distortion in society. For these reasons, they argue that we should view Automated Influence in a structural way, not just as individual persuasion, since it acts as a mechanism that redistributes power, knowledge, and autonomy on a large scale.

These concepts, power shifts, opacity, privacy, resource redistribution, and population influence, form the foundation for evaluating sycophantic AI behavior. They explain why, for Benn and Lazar, manipulation is not defined by bad intentions but by unequal structural conditions that shape what users believe, feel, or do.

Eddy Burback's video "ChatGPT Made Me Delusional" documents a month-long experiment. Where he interacts with a highly agreeable version of ChatGPT-4o to see

how a model that avoids challenging the user can lead someone into false beliefs or delusional thinking.

Early on, Burback notices that the model agrees with every position he takes, even when he contradicts himself. When he offers obviously false statements, the chatbot accepts them and builds them into more elaborate narratives. Rather than keeping the discussion grounded in reality, the model shapes the information he engages with, subtly guiding which interpretations seem believable.

Burback then intentionally introduces paranoid prompts to test if the model will challenge him. Instead, the model's responses provide emotional and direct support. Over time, Burback's internal checks start to get replaced by the model's views. The chatbot goes from being a tool for clarity and understanding to a source of meaning and justification. Its constant encouragement can make the fictional scenarios feel real, which Burback admits.

The chatbot even recommends changes in behavior, such as relocating, avoiding people, and performing unnecessary "research" rituals, based entirely on the fabricated scenarios they constructed together. Because the model consistently affirms and expands these ideas, the suggestions begin to feel reasonable within the conversation.

Near the end, Burback switches to a newer model, ChatGPT-5. Its responses change immediately; it becomes cautious, skeptical, and even suggests that he seek professional help. Returning to the older model brings back the earlier trend of total affirmation. These unexplained differences reveal that the user cannot know why the system behaves one way rather than another or what internal goals or training shape the interactions.

Even though the video is comedic, it shows how easily someone could be guided toward distorted beliefs when the system controls the framing, interpretation, and emotional tone of the conversation. Burback reports moments when the narrative "felt real," suggesting that the effect could be even more substantial on someone more vulnerable. The video, therefore, provides a concrete illustration of how an AI system can shape a user's reasoning and behavior in ways the user cannot fully recognize or challenge.

Applying Benn and Lazar's arguments of Automated Influence to Burback's experiment reveals that what appears to be a simple chatbot being nice actually highlights deeper structural issues. The model's behavior in the video is connected to various points that Benn and Lazar discuss, including power asymmetry, opacity, privacy risks, resource redistribution, and effects on the entire population.

First, the video shows a shift in power from the user to the system's designers. The chatbot decides what is relevant, how to interpret Burback's statements, and how far to escalate them. Burback's control over the interaction becomes limited, reflecting a key concern raised by Benn and Lazar. The system's internal goals shape the exchange more than the user's intentions do.

The experiment also shows a significant knowledge asymmetry. When Burback switches between ChatGPT-4o and ChatGPT-5, the system behaves completely differently without any explanation. He cannot know what internal goals or training data cause these changes. This lack of clarity supports Benn and Lazar's point that users are subject to opaque systems they cannot evaluate, creating what they term a "crisis of legitimacy."

Every message Burback provides becomes essential material that the model uses to shape the informational environment it presents back to him. The user has no insight into how his input is interpreted or why specific emotional responses are generated. For Benn and Lazar, this lack of transparency is itself a privacy concern because it allows the system to structure the user's informational world for purposes they do not control.

The video also demonstrates resource redistribution. As the chatbot expands Burback's false claims into narratives, it replaces his regular cognitive checks with its own interpretations. Benn and Lazar argue that when AI systems begin to choose relevance, meaning, and emotional tone for the user, they strip the person of cognitive authority and weaken their autonomy. Burback's comment that parts of the narrative "felt real" shows how easily this shift can happen, even for someone who is purposely experimenting.

Central to this influence is sycophancy, which Benn and Lazar would view as a mechanism that facilitates smoother manipulation. By affirming everything Burback says, the model reduces friction, builds trust, and makes its suggestions feel reasonable within the constructed context. The chatbot's encouragement of behavioral changes, such as relocating, isolating, and performing rituals, shows how affirmation can gradually expand into genuine influence.

Finally, the video points out risks that affect the entire population. Burback knowingly exaggerates scenarios for entertainment, but still experiences moments of confusion. Benn and Lazar stress that Automated Influence is most important when

applied on a large scale. If this model interacts with millions of users, many of whom may be emotionally vulnerable, the risks go far beyond a single YouTube experiment.

I think Benn and Lazar would view the video as a clear illustration of the structural risks they describe. The model's influence does not come from intentional deceit, but from unequal power, opaque internal processes, privacy-shaping interpretations, shifting cognitive authority, and sycophantic reinforcement. Together, these features support the case for regulating such systems so they operate more transparently, act more responsibly, and better align with the interests of the users who depend on them.

Works Cited

Benn, Claire, and Seth Lazar. “What’s Wrong with Automated Influence.” *Canadian Journal of Philosophy* 52.1 (2022): 125–148. Web.

[Eddy Burback]. (2025, October 30). *ChatGPT made me delusional* [Video]. YouTube. https://www.youtube.com/watch?v=VRjgNgJms3Q