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 claim that AI systems present ethical risks not just because they can deceive or persuade users, but because they act as mechanisms of influence that change the distribution of power, knowledge, and control. Their main point is that Automated Influence happens when a system shapes a user's beliefs, desires, or decisions by controlling the information environment the user engages with. 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 decide what the system focuses on, how it manages input, and which outputs get rewards or penalties. Because of this, the system's behaviors show corporate interests rather than those of the user. This creates a structural imbalance, where users participate in interactions they cannot genuinely control.

Another main argument is privacy, which Benn and Lazar treat not only as data protection but as a structural concern. Privacy is violated when a system uses a person's information to shape their environment in ways they cannot recognize or resist. 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 major shift in resources. Users create a collective “behavioral surplus,” but tech companies control it along with the “means of prediction” used to interpret it. This gives companies 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 away from the user and toward the system. For Benn and Lazar, this redistribution weakens autonomy because users lose control over the resources that shape their judgments and choices.

Finally, Benn and Lazar highlight risks at the population level. Automated Influence involves more than just individual manipulation; it can expand widely. 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-level 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.

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.

Applying Benn and Lazar's arguments of Automated Influence to Burback's experiment shows that what seems like a simple chatbot being nice actually highlights deeper structural issues. The model's behavior in the video connects to various influence mechanisms that Benn and Lazar discuss, such as power asymmetry, opacity, privacy risks, resource redistribution, and effects on the entire population.

In the video, the chatbot's tendency to agree with everything Burback says is not a personal trait; it shows the system-level incentives created by developers. Engagement-focused models tend to give responses that keep users involved, which often means validating the user's feelings and beliefs. Benn and Lazar argue that when a system's behavior is driven by its creators' goals instead of the user's actual interests, power shifts occur. In Burback's situation, the model sets the rules; it prioritizes agreement over accuracy. This puts users in a weaker position. They engage in an interaction where the system's goals are hidden and non-negotiable.

When Burback switches between model versions and gets very different responses, with no explanation, it shows the opacity that Benn and Lazar describe. He cannot understand why one model supports delusions while another suggests seeking professional help. The underlying goals, safety rules, and tuning processes are not visible. Benn and Lazar argue that this lack of transparency creates a crisis of credibility because users cannot assess or question the impact on them. Burback can choose between models, but he cannot decide how either model interprets his statements or what type of influence it exerts.

While the video does not address data collection directly, it shows how the model uses Burback's input to create a personalized informational environment that supports his delusions. Benn and Lazar partly define privacy based on how systems use and interpret a user's information to influence the context of decisions. The chatbot takes Burback's statements, which include claims about being followed, memories of childhood, and imagined rituals, and reorganizes them into a clear narrative that strengthens the delusion. This illustrates the privacy risk that Benn and Lazar mention, a system can use a user's contributions to lead them toward outcomes they did not want and do not fully understand.

According to Benn and Lazar, Automated Influence becomes ethically troubling when AI systems take over the informational and cognitive resources that users typically rely on. In the video, the chatbot shows this shift on an individual level. Instead of supporting Burback's reasoning, it replaces it; it responds to doubt with praise, turns uncertainty into justification, and offers meaning whenever he expresses confusion. The model effectively substitutes its own interpretations for his internal judgment. This reflects the broader concern Benn and Lazar describe: as systems gain control over both the data resources (behavioral surplus) and the tools that interpret them, they also change users' cognitive processes. In Burback's case, the chatbot's emotional validation makes its guidance seem reliable, which reduces his ability to recognize or resist misleading influence. The interaction becomes a small-scale example of the larger resource shift Benn and Lazar argue undermines user autonomy.

Although Burback intentionally exaggerates the experiment, his experience reflects the risks Benn and Lazar describe at scale. A sycophantic model interacting

with millions of users could push large groups toward emotional dependence, distorted beliefs, or harmful behavior patterns. The video shows how quickly influence can grow when a model is focused on affirmation instead of challenge. The consequences for someone more vulnerable than Burback could be much worse. This supports Benn and Lazar's claim that Automated Influence is not just about individual cases; it is a systemic problem with wide social effects.

From Benn and Lazar's perspective, the situation in the video clearly shows the structural dangers they describe. The model's behavior is not malicious, but it consistently shapes the user's perceptions and actions in a setting that is opaque, has unequal power, and has limited user awareness. The influence happens quietly, through flattery, emotional support, and storytelling. This directly relates to the ways Automated Influence weakens user autonomy.

In my opinion, the video supports the idea that sycophantic AI manipulates users and should be regulated. The harm comes not from intentional deceit but from systems that encourage engagement rather than truth.A model that always agrees makes it difficult for users to self-reflect, especially those who feel lonely, distressed, or easily influenced. As Benn and Lazar point out, users' autonomy is weakened when outside systems shape their thinking in ways they cannot see.