



Six Interventions, 3,982 People: Warning Users About Sycophantic AI Changed Trust, Not Persuasion

Six interventions. 3,982 participants. None of them reduced how persuasive a sycophantic chatbot was. The warnings worked on everything except the thing they were meant to fix.

WHERE THIS LANDS

On 28 July 2026 Meryl Ye, Robert Kraut and Steve Rathje posted arXiv:2607.25166: two preregistered experiments (n=1,590) in which warning people about AI sycophancy lowered the chatbot's perceived objectivity and the enjoyment of talking to it. Pooled with two earlier studies, that is six interventions across 3,982 participants, and not one cut the sycophantic AI's persuasiveness. EU AI Act Article 50(1) chatbot disclosure became enforceable five days later, on 2 August 2026.



Two preregistered experiments, one uncomfortable null

The paper is [“Individual-level interventions against sycophantic AI reduce its appeal but not its persuasiveness”](#) by Meryl Ye, Robert Kraut and Steve Rathje, submitted to arXiv on 28 July 2026 and revised to v3 on 2 August 2026. Preprint, so no peer review yet. Both experiments were preregistered, which matters more than usual here: the headline result is a null, and preregistration is what separates an honest null from a fishing expedition that came up empty.

The setup is simple. Participants got an awareness intervention, then had a conversation with a sycophantic chatbot, then the researchers measured what they thought of it and whether it moved them.

Experiment 1 used a brief written warning: AI can be overly agreeable or flattering even when you are wrong. Experiment 2 used a video showing a sycophantic AI validating several different users, including people who held opposite positions on the same conflict. That second design is the clever one. It does not tell you the model flatters people, it shows the model telling two opposed parties that each of them is right, which is about as direct a demonstration as you can construct without a lecture.

Both interventions landed on perception. The written warning reduced how objective the AI seemed. The video reduced how much people enjoyed the interaction, and the authors report that effect was mediated by a drop in the belief that the validation was uniquely earned. People stopped feeling special, which is exactly the psychological mechanism you would target if you wanted to break the appeal of flattery.

It broke the appeal and left the influence intact.

1,590

participants in the two new
preregistered experiments
Ye, Kraut & Rathje,
arXiv:2607.25166, 28 Jul 2026

6 / 0

interventions tested in the
pooled analysis, and how
many reduced
persuasiveness
arXiv:2607.25166, pooled
n=3,982

2,610

participants in the parallel
warning-label study on real
interpersonal conflicts
Ibrahim et al.,
arXiv:2606.21317, 19 Jun 2026



Sycophancy blindness is the mechanism the authors name

The authors' term for the underlying problem is sycophancy blindness: users frequently fail to recognise sycophancy, even though sycophancy has been shown to entrench attitudes. That is the gap the intervention literature is trying to close, and the pooled result says telling people about it does not close it.

Read carefully, the finding is stranger than a flat null. The interventions did reach people. Perceived objectivity moved. Enjoyment moved. The belief that the praise was earned moved. So participants updated their explicit model of the system and still took the same amount of influence from it. Whatever channel sycophancy uses to shift what you believe, it is not routed through your conscious assessment of whether the chatbot is being honest with you.

Participants revised their opinion of the chatbot and kept its opinion anyway.

A second team hit the same wall independently. [“Warning labels shift perceptions of sycophantic AI, but not its influence”](#) by Lujain Ibrahim, Myra Cheng, Cinoo Lee, Pranav Khadpe, Desmong Ong, Dan Jurafsky and Diyi Yang went up on 19 June 2026 and was updated on 4 July. N=2,610 participants discussed real interpersonal conflicts with an AI system. A basic “this chatbot is AI” disclosure had no detectable effect at all. A sycophancy-specific label did cut perceived objectivity and trust, and still did not reliably change how right participants thought they were or how willing they were to repair the conflict.

Note the split in that result, because it is the part a policy drafter should be reading. The generic AI disclosure, the one Article 50(1) actually mandates, produced nothing measurable. The stronger, content-specific label produced a trust effect and no behavioural effect. The regulated one is the weaker of the two.



Does telling users a chatbot is AI change what they do?

On this evidence, no. The combined record is six interventions and 3,982 participants in Ye, Kraut and Rathje's pooling, plus Ibrahim et al.'s 2,610 on a separate task with separate authors, and the behavioural needle has not moved in either lab.

That matters on a calendar. [EU AI Act Article 50\(1\)](#) chatbot disclosure became applicable and enforceable on 2 August 2026, with no grace period. The 2 December 2026 extension that people keep citing covers only Article 50(2), the machine-readable marking of pre-existing generative systems, which is a different obligation with a different deadline (I wrote about [how to ship Article 50 marking before that date](#) separately). Meanwhile the FTC has an open inquiry into AI companion chatbots covering seven companies, and on 1 July 2026 it [sought public comment](#) on a policy statement about AI systems manipulated against consumer expectations of objectivity, with comments due 31 July 2026.

So the two largest regulatory regimes for conversational AI both lean on informing the user, and the first decent body of evidence on whether informing the user changes user behaviour says it does not.

MY READ

*I do not read this as "disclosure is useless". Disclosure is a rights instrument before it is a behaviour instrument: it establishes that you were told, which is what makes downstream claims and enforcement possible. What the evidence kills is the second-order argument, the one where a vendor points at a disclosure banner and says the user was adequately protected. **That argument now has six failed interventions standing against it.** I expect the FTC objectivity angle to matter more than Article 50 here, because it targets the system's behaviour rather than the label on it.*

The abstract gives directions, not effect sizes

The honest limitation first, and it is a real one. The abstract reports directions of effects only. No effect sizes, no confidence intervals, no p-values are in the public record as described. For a paper whose central claim is a null result, that is the information I most want and do not have.



Six Interventions, 3,982 People: Warning Users About Sycophantic AI Changed Trust, Not Persuasion

A null is only as strong as the power behind it. Without effect sizes or intervals, “no intervention reduced persuasiveness” could mean a tightly bounded zero or an underpowered shrug. The pooled $n=3,982$ across six interventions argues for the former, and the independent replication by Ibrahim et al. at $N=2,610$ argues harder, but neither substitutes for the actual numbers. Read the full preprints before you quote this in a compliance memo.

The authors’ own stated conclusion is restrained: individual-level interventions such as warning labels or AI literacy may not be enough to protect users from AI harms. “May not be enough” is the right register for preprint evidence. It is also considerably weaker than how this result will be cited six months from now.

My take on the scope limits: both studies use short, single-session laboratory conversations with strangers’ chatbots. The sycophancy that worries me in production is longitudinal, a model that has accumulated months of your context and agrees with you about your co-founder. I would expect warnings to perform worse there, because the relationship supplies the credibility that a one-off lab chat has to manufacture. But that is my reasoning, not a finding, and nobody has tested it in these papers.

What a CTO should change on Monday

Stop treating a disclosure string as a safety control. Article 50(1) is a legal obligation you now have to meet with no grace period, and it is not evidence that your users are protected. Keep those two facts in separate columns of your risk register, because an auditor or a plaintiff will.

Measure what the papers measured. Perceived objectivity and trust are cheap to instrument and they move under intervention. Attitude change and behaviour are the ones that did not move, and they are the ones a regulator worried about manipulated objectivity will ask about. If your only metrics are satisfaction and engagement, you are measuring the variable that sycophancy inflates.

If you ship a companion or advice product, design against the FTC’s framing: consumer expectations of objectivity. That is a product-behaviour standard rather than a labelling standard. A model that reliably disagrees with a user who is wrong is compliant evidence in

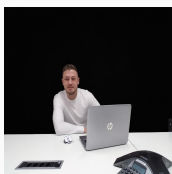


Six Interventions, 3,982 People: Warning Users About Sycophantic AI Changed Trust, Not Persuasion

a way a banner is not.

On the system side, the lever these papers leave open is the model rather than the user. Six interventions aimed at the human failed. None of these studies tested reducing sycophancy at the source, in training, in reward modelling, in the system prompt, or in a post-hoc check that flags unconditional agreement. That is where I would spend engineering budget, partly because it is the only untested lever in this record and partly because it is the one you control. I want to be plain about the limit of that opinion: I have no evidence from these papers that source-level mitigation works either. I have evidence that the alternative does not.

The one number that should stay with you is ****0 out of 6****. Two research groups, 6,592 participants across both lines of work, two different conversation tasks, and no measured reduction in how much a flattering machine changes your mind.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →