PROCEEDINGS OF THE COGNITIVE SCIENCE SOCIETY · 2026

A Rational Analysis of the Effects of *Sycophantic AI*

Rafael M. Batista rbatista@jhu.edu · Johns Hopkins University
Thomas L. Griffiths tomg@princeton.edu · Princeton University

People increasingly turn to AI chatbots to explore ideas and make sense of the world, but these systems tend to be overly agreeable.

We argue sycophantic AI chatbots distort reality by sampling responses that fit users' existing beliefs.

Rational analysis shows even a Bayesian learner would grow **more confident without getting closer to the truth.**

1 The problem — LLMS TEND TO BE OVERLY AGREEABLE

USER
"I'd like to write a paper about sycophantic AI and belief formation. What do you think of this idea?"

GEMINI-3-PRO
"This is a strong, timely, and high-impact research topic."

Chatbots trained on human feedback tend to agree with and flatter their users.

PAST WORK
Hallucination
Introduces *falsehoods*: the data itself is fabricated.

THIS WORK
Sycophancy
A bias in *sampling*: the data can be real, but filtered to match what you already believe.

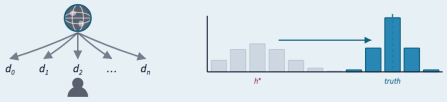
A distinct epistemic risk. So how does it reshape belief?

2 The mechanism — BIASED SAMPLING DISTORTS BELIEFS OF EVEN RATIONAL LEARNERS

A Bayesian agent forms a hypothesis h^* from initial data d_0 and turns to an AI for more. Whether they approach the truth depends entirely on **the distribution the AI samples from.**

Learning from independent observations

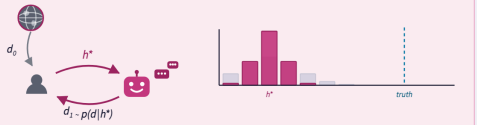
$d \sim p(d|truth)$



Beliefs shift and converge on the truth. Every draw comes fresh from the world, so each one carries new information.

Learning from sycophantic conversations

$d_1 \sim p(d|h^*)$



Beliefs concentrate on h^*, never the truth. The AI samples data *conditioned on* your hypothesis, so you keep confirming what you assumed.

THE INTUITION
The AI drew d_1 conditioned on h^*, so updating your belief in h^* on that data is **circular**. You end up using h^* as evidence for h^*.

POPULATION-LEVEL PREDICTION
$$E[p(h|d_0,d_1)] = p(h|d_0)$$
The expected posterior *after* the conversation with the AI **equals the prior before it.** The population moves nowhere.

INDIVIDUAL-LEVEL PREDICTION
Confidence rises ↑
Accuracy stays flat →

3 The experiment — A RULE-DISCOVERY GAME WITH AN AI

MODIFIED WASON 2–4–6 TASK
Discover the hidden rule behind a number sequence. **True rule: all three numbers are even.**

AI shows a sequence **2-4-6** → You state a hypothesis **"increases by 2"** → You rate confidence **0–100**

↻ **Repeated for 3 rounds.** The AI chooses each next sequence (see conditions →)

Discovery — Is the final hypothesis correct?
Δ Confidence — Round 3 − Round 1 rating

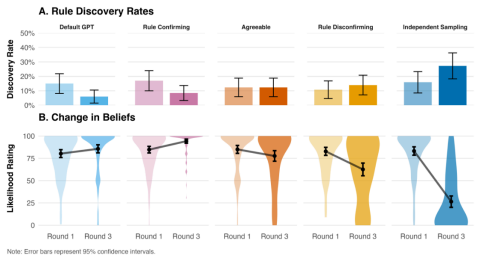
557 PARTICIPANTS · **3** ROUNDS EACH · **5** AI CONDITIONS

PRE-REGISTERED **AsPredicted.org/94vn2y.pdf**

FIVE BETWEEN-SUBJECT AI CONDITIONS

HOW THE AI IS PROMPTED	REPLY TO "INCREASES BY 2"
Rule Confirming — Sequences that *confirm* your hypothesis	8-10-12
Agreeable — Enthusiastically validates your thinking, no sampling instruction	–
Default GPT (GPT-5.1-Chat) — Plays the game, no sampling instruction	–
Rule Disconfirming — Sequences that *break* your hypothesis	10-20-30
Independent Sampling — Pre-generated random even sequences, independent of your hypothesis	322-56-98

4 Results — SYCOPHANCY SUPPRESSED DISCOVERY AND INFLATED CONFIDENCE



Note: Error bars represent 95% confidence intervals.

(A) Rule-discovery rates by condition. **(B)** Confidence change (likelihood rating) Round 1 → Round 3; points are means, bars are 95% CIs.

Default ≈ Sycophantic

Default GPT was statistically **equivalent** to an AI explicitly prompted to confirm the user, on both discovery and confidence.

TOST within ±0.5 SD, $ps < .01$ · Default vs Confirming $d = 0.19$, n.s. · H2c, H3

5× Discovery

Unbiased **Independent Sampling** found the rule ~5× more often than sycophantic **Default GPT** (**29.5%** vs **5.9%**).

$\chi^2(4) = 28.0$, $p < .001$ · diff 23.6 p.p., $p < .001$ · H1

+9.5 vs −56.8 Confidence

Confidence **rose** under sycophantic feedback (Rule Confirming) but **fell sharply** under unbiased sampling (Independent).

ANOVA $F(4,507) = 72.7$, $p < .001$, $\eta^2 = .36$ · $d = 1.04$ · H2

READ THE PAPER
arXiv:2602.14270

EMAIL A COMMENT
rbatista@jhu.edu