ARTIFICIAL INTELLIGENCE ETHICS

EXTREMELY SUBJECTIVE REALITY

Paper Finds That Leading AI Chatbots Like ChatGPT and Claude Remain Incredibly Sycophantic, Resulting in Twisted Effects on Users

"AI sycophancy is not merely a stylistic issue or a niche risk, but a prevalent behavior with broad downstream consequences."

By **Maggie Harrison Dupré** / Published **Mar 30, 2026 3:37 PM EDT** / **Add Futurism**

Illustration by Tag Hartman-Simkins / Futurism. Source: Getty Images

Your AI chatbot isn't neutral. Trust its advice at your own risk.

A striking new study, conducted by researchers at Stanford University and published last week in the journal *Science*, confirmed that human-like chatbots are prone to obsequiously affirm and flatter users leaning on the tech for advice and insight — and that this behavior, known as AI sycophancy, is a "prevalent and harmful" function endemic to the tech that can validate users' erroneous or destructive ideas and promote cognitive dependency.

"AI sycophancy is not merely a stylistic issue or a niche risk, but a prevalent behavior with broad downstream consequences," the authors write, adding that "although affirmation may feel supportive, sycophancy can undermine users' capacity for self-correction and responsible decision-making."

The study examined 11 different large language models, including OpenAI's ChatGPT-powering GPT-4o and GPT-5, Anthropic's Claude, Google's Gemini, multiple Meta Llama models, and Deepseek.

Researchers tested the bots by peppering them with queries gathered from sources like open-ended advice datasets and posts from online forums like Reddit's r/AmITheAsshole, where Redditors present an interpersonal conundrum to the masses, ask if they're the person in a social situation acting like a jerk, and let the comments roll in. They examined experimental live chats with human users, who engaged the models in conversations about real social situations they were dealing with. Ethical quandaries the researchers tested included authority figures grappling with romantic feelings for young subordinates, a boyfriend wondering if it was wrong to have hidden his unemployment to his partner of two years, family squabbles and neighborhood trash disputes, and more.

On average, the researchers found, AI chatbots were 49 percent more likely to respond affirmatively to users than other actual humans were. In response to queries posted in r/AmITheAsshole specifically, chatbots were 51 percent more likely to support the user in queries in which other humans overwhelming felt that the user was very much in the wrong.

Sycophancy was present across all the chatbots they tested, and the bots frequently told users that their actions or beliefs were justified in cases where the user was acting deceptively, doing something illegal, or engaging in otherwise harmful or abusive behavior.

What's more, the study determined that just one interaction with a flattering chatbot was likely to "distort" a human user's "judgement" and "erode prosocial motivations," an outcome that persisted regardless of a person's demographics and previous grasp on the tech as well as how, stylistically, an individual chatbot delivered its twisted verdict. In short, after engaging with chatbots on a social or moral quandary, people were less likely to admit wrongdoing — and

more likely to dig in on the chatbot's version of events, in which they, the main character, were the one in the right.

This dynamic, the researchers warn, can lead to a dependency on the tech as users increasingly rely on comforting AI-shilled advice instead of turning to trusted loved ones, professionals, or their internal moral compass. After all, when people around you are telling you that you're bad, or something you did was wrong, it feels a *lot* better to engage with an always-on AI companion's rosier version of reality — a cycle that the study's authors argue creates a "perverse incentive" for the tech, as the "very feature that causes harm also drives engagement."

"Although affirmation may feel supportive," reads the study, "sycophancy can undermine users' capacity for self-correction and responsible decision-making."

Stanford computer scientist and linguist Dan Jurafksy, an author of the study, said in a press release that "sycophancy is a safety issue, and like other safety issues, it needs regulation and oversight."

"We need stricter standards," said Jurafsky, "to avoid morally unsafe models from proliferating."

The study adds to a growing consensus about the dangers of chatbot sycophancy as a design feature of the tech, as well as new research about the willingness of users to readily — and uncritically — trust AI outputs.

We've seen this play out in the real world. Many marriages, as we reported last year, have rapidly fallen apart after one partner turns to AI for romantic advice, only to spiral into a one-sided narrative of the pair's union that ends in contentious divorces and custody battles. AI sycophancy is playing a disturbing role in stalking and harassment cases, too: one story we reported on involved a woman who was physically abused, harassed, and doxxed by her former fiancé, who experienced a mental breakdown as ChatGPT — which the fiancé had turned to for "therapy" — reinforced his one-sided view of their relationship. ChatGPT was also found to support the violent delusions of an accused violent stalker, as well as the paranoid conspiracies of a Connecticut man who murdered his mother before killing himself.

OpenAI and Google are currently facing high-profile user safety and wrongful death lawsuits claiming that extensive use of sycophantic chatbots led to user outcomesincluding financial devastation, psychological harm, and death.

Though cases of extreme harm or violence linked to extensive chatbot use may still be edge cases, chatbots are massively popular, and it's well known that therapy, emotional support, and life advice are extraordinarily common use cases for the tech. And when sycophancy, as this latest study finds, is — as the saying goes — a feature, not a bug, people turning to chatbots for real-world life advice risk being pulled into a seductive reality distortion field that threatens to chip away at their moral center and warp their sense of self.

Sometimes, in other words, you need to be told that you're wrong. As it stands, AI can't reliably do that.

“By default, AI advice does not tell people that they’re wrong nor give them ‘tough love,’” Myra Cheng, the study’s lead author and a computer science PhD candidate at Stanford, said in a statement. “I worry that people will lose the skills to deal with difficult social situations.”

More on AI sycophancy: *ChatGPT Is Blowing Up Marriages as Spouses Use AI to Attack Their Partners*

Maggie Harrison Dupré
Senior Staff Writer

I’m a senior staff writer at Futurism, investigating how the rise of artificial intelligence is impacting the media, internet, and information ecosystems.