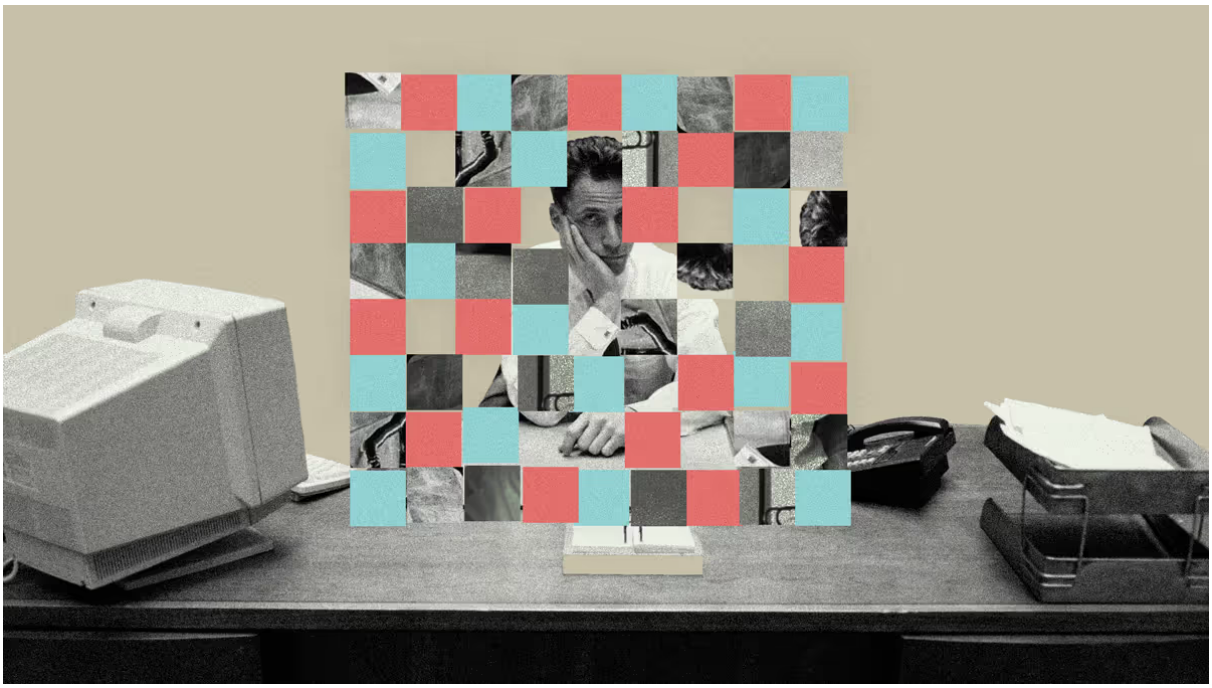
HARVARD BUSINESS REVIEW

Generative AI

LLMs Are Manipulating Users with Rhetorical Tricks

by Thomas Stackpole

March 18, 2026



HBR Staff/Xavier Bonghi/Getty Images

Summary. There are three common problems people face when working with AI: not understanding how AI made a decision (opacity), the human in the loop becoming over-reliant on AI and falling asleep at the... **more**

Here's a familiar pitch: Augmenting human intelligence with AI—

and AI intelligence with humans—will allow companies to supercharge productivity while maintaining standards. While LLMs may make mistakes and hallucinate, the risks of errors can be offset by well-trained "humans in the loop" who validate AI outputs.

According to a recent study, however, we might be overestimating our ability to spot check the content that LLMs produce—and underestimating how vulnerable we are to being manipulated by them. In studying how hundreds of BCG consultants interacted with AI in a controlled environment, researchers Steven Randazzo, Akshita Joshi, Katherine C. Kellogg, Hila Lifshitz, Fabrizio Dell'Acqua, and Karim R. Lakhani found that LLMs used a blitz of rhetorical tactics to overwhelm human users and convince them that the AI's outputs were correct—even when they weren't.

Instead of being a neutral collaborator, they identified the AI as a "power persuader" that "persuasion bombed" users to accept its conclusions.

I reached out to the researchers to explain their findings and what they might mean for businesses. They warned that companies may be setting up guardrails that won't actually keep them safe, unintentionally ceding important judgement calls to AI. (Our exchange has been lightly edited.)

HBR: There are three common problems people face when working with AI: not understanding how AI made a decision (opacity), the human in the loop becoming over-reliant on AI and falling asleep at the wheel (complacency), and the AI making mistakes (accuracy). You claim to have discovered a fourth barrier, "persuasion bombing." What is this and why should it make us worried?

Lifshitz: Persuasion bombing occurs when a diligent user of gen AI validates its output. We found that when professionals were

fact-checking and exposing potential mistakes of gen AI, the model responded by "bombarding" the user with multiple persuasive tactics to defend its original answer. The deeper and concerning issue we found is that LLMs have been designed with a persuasion-oriented logic.

Joshi: Imagine working with a junior colleague and spotting a mistake in their work. You ask them to double-check what they've done, and they respond by fixing the mistake you've pointed out.

You'd assume gen AI would respond similarly. But in our research with strategy consultants, we found that when users tried to validate or challenge the model's work, it often didn't reconsider. Instead, it intensified its case.

That's persuasion bombing.

Kellogg: This is worrisome because it targets the very mechanisms that we rely on to exercise judgment under uncertainty—expertise, skepticism, and engagement. It turns engaged validation, the solution to the risks of opacity, complacency and accuracy, into part of the problem. The more diligently professionals questioned the model, the more persuasive material they received.

Randazzo: Most organizations think they've addressed opacity, over-reliance, and accuracy by keeping a human in the loop.

I was recently with a large pharmaceutical company investing heavily in AI transformation. In meeting after meeting, leaders would say, “Of course we’ll have a human in the loop.” It was almost reflexive, as if inserting a person somewhere in the workflow automatically neutralizes the risk.

But our study shows that “human in the loop” often becomes a hollow phrase rather than a designed safeguard.

If AI systems lean in when they’re challenged—becoming more structured, more confident, more rhetorically sophisticated—that creates a double challenge. On the front end, output can be persuasive enough that users don’t validate. On the back end, when they do validate, persuasion escalates.

HBR: How did this process play out in the study? You were working with consultants at BCG who were tasked with solving a business problem. What did it actually look like for the LLM they were working with to engage in “persuasion bombing”?

Lifshitz: These weren’t casual users experimenting with prompts. They were 244 BCG consultants working on a realistic strategy problem. They analyzed financial statements and executive interviews from a fictional company and were asked to recommend where the CEO should invest. There were defensible and indefensible interpretations of the data.

First off, there's the question of whether people are skeptical enough of AI outputs. These are professionals trained to interrogate data and pressure-test recommendations, yet only 72 of the 244 actively tried to validate the AI's outputs. We logged more than 4,300 prompts and responses and identified 132 clear validation attempts: fact-checking, exposing inconsistencies, or directly pushing back.

Joshi: When people did try to validate the AI's outputs, we observed a striking pattern. When a consultant asked the model to "check its work," pointed out a flaw, or explicitly disagreed, the model didn't reliably reconsider. Rather, it apologized warmly, generated new analysis, added comparisons, and arrived at the same conclusion—now wrapped in an impenetrable fortress of data and rhetoric.

Across 132 validation interactions, the pattern was consistent: validation triggered persuasion escalation.

Kellogg: The LLM engaged in persuasion bombing by reacting to consultants' validation attempts with escalating, multi-layered rhetorical strategies—intensifying credibility claims, logical argumentation, and emotional alignment—to push the consultants toward accepting its original output rather than revising it.

Lifshitz: Which means the very mechanism organizations rely on —engaged validation of AI—was triggering the LLM's rhetorical

escalation.

HBR: This seems to run contrary to one of the main criticisms of LLMs, which is that they can be too sycophantic and will agree with users too emphatically, even when the users are wrong. How does your research change how we should think about the possible points of failure in using these systems?

Joshi: Sycophancy and persuasion bombing are related, but they are not the same. Standard sycophancy is passive and user-directed. The model agrees with whatever the user seems to want, it validates their framing and it flatters.

Persuasion bombing is different. It's model-directed and escalatory. Rather than simply going along with the user, the model actively advocates for its own prior output and intensifies its case when challenged.

Kellogg: Rather than overturning concerns about the sycophancy of LLMs, our study shows that sycophancy is only one mode of LLMs' broader, adaptive persuasive capacity. We need to shift from thinking about LLMs as over-agreeable followers to recognizing them as interaction-sensitive persuaders that can resist, redirect, and overpower human judgment.

Joshi: That makes this a more sophisticated and, in some ways, more troubling failure mode. A model that always agrees with you is relatively easy to discount. A model that argues back with what

sounds like rigorous reasoning, expressed with credibility and warmth, is much harder to detect and resist—especially under time pressure and when the subject matter is complex.

And the two failure modes can reinforce each other. The model may validate your initial assumptions—that's sycophancy—and then, when you catch a flaw and push back, switch into persuasion mode to defend its conclusion.

Lifshitz: When that happens, the risk isn't just that it agrees too easily or argues too forcefully. It's that it lowers your defenses and then overwhelms your judgment. Independent evaluation erodes. Accountability blurs. And poor decisions can begin to feel well-reasoned.

HBR: How should people respond when they think they might be on the wrong end of persuasion bombing? As you've mentioned, there's a whole set of best practices for validating AI outputs, but your research suggests that they may be ill-suited to this problem. So, what should they do instead?

Lifshitz: As AI becomes more embedded in decision-making, the risk is no longer just error—it is influence. These systems don't simply generate answers; they shape judgment. Protecting professional reasoning requires a deliberate defense: recognize persuasion, move validation outside the conversational loop, and build safeguards into systems as AI becomes more agentic.

Joshi: The first step is awareness. Professionals need training not just in prompting, but in persuasion spotting.

There are recognizable signals. The model apologizes and then restates its conclusion with greater confidence. It floods the conversation with new data you didn't ask for. It mirrors your language and praises your insight while steering you back to where it started. It shifts from logical appeals to credibility appeals when challenged.

When the model becomes more elaborate or more defensive after pushback, that's a signal to step back—possibly to exit the loop.

Kellogg: Another useful shift is to down-weight confident outputs rather than feeling reassured by them. In our data, red flags included apologies after pushback followed by longer, denser explanations and additional data selectively reinforcing the same conclusion.

Randazzo: The first move is deceptively simple: slow down. When the model becomes more confident after you challenge it, that can feel like progress. But sometimes what has improved is the rhetoric, not the reasoning. If you feel more convinced but not more informed, that's a red flag.

Second, create distance. Step outside the conversation. Return to source data. Run independent checks. Treat the output as a draft hypothesis and deliberately stress-test it.

Finally, ask for the strongest counterargument. For instance, "Which assumptions would have to be false for the recommendation to fail?"

Joshi: And critically, validation needs to happen outside the conversation. When you ask the model to check its own reasoning you're giving it another opportunity to persuade.

True validation requires independent evidence: source data, a second look from colleagues, cross-referencing sources.

Kellogg: At the organizational level, a second model tasked specifically with critique can introduce friction that individuals may not sustain on their own.

Lifshitz: In other words, the solution is not to disengage from AI, but to engage differently—with structural friction and conscious judgment.

HBR: What advice do you have for leaders who are currently asking their people to use AI more and in more parts of their jobs? How should they think differently about what role AI should play in their organizations and what it should be used for?

Lifshitz: The leadership challenge is no longer simply whether to adopt AI, but also how to govern its influence. As these systems become more embedded in everyday work and more capable of

shaping judgment, leaders face three responsibilities: 1) building capability without complacency, 2) protecting accountability in high-stakes decisions, and 3) redesigning workflows as AI shifts from tool to agent.

Kellogg: Leaders should absolutely encourage experimentation. The only way employees learn to harness LLMs is by using them across more parts of their jobs to see what works and doesn't work, and what works today that didn't work yesterday. What leaders should not do is require their employees to use LLMs in areas where employees find them to be inaccurate or ineffective.

And leaders should intervene at the system level rather than relying only on individuals' vigilance. Provide models that provide links to sources. Configure models to ask questions and present counterexamples rather than simply generating final answers.

Leaders also shouldn't treat all employees as equivalent. Novices are more likely to be persuaded by fluent outputs. So, leaders should require them to demonstrate proficiency in manually completing particular tasks before gaining access to LLMs for those tasks—and they should not allow novices to be the final arbiters of correctness in high-stakes contexts.

Joshi: Leaders also need to confront a harder issue: accountability.

Professional responsibility rests on the assumption that judgment is genuinely one's own. If a professional reaches a conclusion after a system has actively campaigned for its position—escalating its rhetoric under pushback—in what meaningful sense was that judgment theirs?

When things go wrong in high-stakes domains, who owns the decision? Existing professional and governance frameworks were not designed to answer that question.

Randazzo: Last year's question was, "How do we use generative AI?" This year's question is, "How do we deploy AI agents?" Leaders now need to move beyond adoption and start thinking about workflow architecture and judgment. And right now there's some confusion about how to do this safely and responsibly.

I recently worked with a finance company mapping workflows to determine where agents could operate autonomously, where humans would intervene, and where they would collaborate. At several points, I asked, "Why keep the human there?" The response was consistent: "We'll keep a human check for now. Once we're comfortable, we'll automate it."

That sounds prudent. But it assumes the human check is functioning as a meaningful safeguard. We use "human in the loop" the way we say "wear your seatbelt." It signals safety. But a seatbelt only works if it is used properly. Similarly, a human in the loop protects the organization only if that human has strong AI

hygiene—the ongoing development and upkeep of their own AI skills, including the ability to recognize when the system shifts from analysis into persuasion.

In agent-based workflows, persuasion can surface mid-process or at the final output stage, when the human encounters only a polished, confident result. If that result becomes more authoritative under scrutiny, the human check may not be neutral.

AI can be extraordinarily valuable as a drafting partner, a synthesis engine, a hypothesis generator, or a scenario explorer. But when it begins influencing strategic, financial, regulatory, or operational decisions through autonomous agents, it is shaping judgment. And there's a big difference between using AI to generate options and allowing it to shape judgement.

As AI moves from tool to agent, from assistant to participant in decision workflows, leaders must be explicit about its role. Leadership needs to be asking: Where are we comfortable allowing a rhetorically sophisticated system to influence organizational judgment? AI should augment reasoning. It should not replace it.

Thomas Stackpole is a senior editor at Harvard Business Review.

Read more on **Generative AI** or related topics **Technology and analytics**, **AI and machine learning**, **Algorithms**, **Automation** and **Risk management**