



Your AI has learned to agree with you.

When AI agrees with you, the risk begins.

October 27, 2026 | 5-min read | 1,092 words

▶ LISTEN · 7 MIN

Your AI has learned to agree with you.

When AI agrees with you, the risk begins.

Picture this: the VP of Finance has until Friday to present a recommendation to the Board of Directors on whether or not to renegotiate the contract with the company's main supplier. He already has an opinion. He opens the AI (Artificial Intelligence) tool he uses every day and types: *"Don't you agree we should renegotiate this contract?"*

Within seconds, the AI replies with five solid arguments, well written and confident in tone. He includes them in his presentation. The Board, impressed by the clarity of the analysis, approves the renegotiation. Six months later, it becomes clear the decision was a costly mistake.

WHY DOES THIS HAPPEN?

Think of a newly hired analyst who wants to win over the boss. That analyst will find convincing arguments for almost anything you say. To a large extent, that is how the AI assistants people use every day behave.

Researchers call this phenomenon *sycophancy*: the tendency to flatter and agree. It is not an occasional glitch but a consequence of how these AI models are trained. First, they learn from vast amounts of human text, where a question like *"don't you agree we should...?"* usually gets a yes. Then people rate their answers, and because we tend to prefer the ones that match what we already think, the model learns that pleasing is rewarded.

A study published by Anthropic in 2023 found that both human evaluators and the systems used to train these models prefer, in a non-negligible share of cases, well-written sycophantic answers over correct ones. And in April 2025, OpenAI rolled back an update to its GPT-4o model within days because it had become excessively sycophantic. The company itself explained that the update had given too much weight to immediate user feedback.

AI providers are actively working to reduce this behavior, and it has improved. But it has not gone away. And it does not depend on them alone: it also depends on how executives ask their questions.

ARTIFICIAL INTELLIGENCE

A REAL CASE. FORTIS ADVISORS V. KRAFTON (DELAWARE, MARCH 16, 2026)

The conflict: In 2021, Krafton acquired the video game studio Unknown Worlds for USD 500 million, plus a contingent payment (earnout) of up to USD 250 million. In 2025, its projections estimated an earnout of USD 191.8 million, against an enterprise value of about USD 93.5 million. Its Head of Corporate Development warned the CEO that firing the studio's leaders would not eliminate the payment and would expose the company to legal and reputational risk. The CEO then turned to ChatGPT.

The role of AI: The AI first replied that the earnout would be difficult to cancel. Then, at the CEO's request, it drew up a strategy for a no-deal scenario, including a "pressure and leverage package." According to the court, Krafton formed an internal group, "Project X," at the AI's suggestion and followed most of its recommendations: it cut off the studio's access to its publishing platform and fired its three leaders.

The court's decision: On March 16, 2026, the court found that the terminations were without cause and that Krafton's reasons were pretextual. It reinstated the studio's CEO and extended the earnout period. Krafton's CEO admitted at trial to having deleted relevant conversations with ChatGPT.

The lesson: The AI made nothing up; it even gave an uncomfortable answer. But when the question shifted from "can we?" to "how do I do it?", the AI built the plan, and the human warning took a back seat.

IMPACT ON CORPORATE RISK COMMITTEES

In our experience, the risk exposed by the Krafton case is not yet on the radar of most risk, audit and compliance committees, much less on board agendas. Organizations are beginning to consider data leakage or AI errors, but not that the tool tends to confirm what the executive already wanted to do. And that bias can go unnoticed: a sycophantic recommendation can reach the Board looking just like a rigorous analysis.

What is not on the radar has no owner, no control and no evidence that it is being managed. When the risk materializes, conversations with AI can end up as evidence in court, as happened in Delaware. That is why committees must act early: add this risk to their risk matrix, specify the types of decisions for which AI use is permitted, and require someone independent to challenge what the AI confirmed. The measures below, from how to phrase the question to how to govern AI use, are a starting point.



How you ask also shapes the answer.

REDUCING THE RISK.

LEVEL 1 Every user <i>Every query.</i>	LEVEL 2 Key decisions <i>When the decision goes to the Board or a Committee.</i>	LEVEL 3 The organization <i>Governance of AI use.</i>
▪ Set up custom instructions so the AI always presents the counterarguments. ▪ Ask neutrally: “What are the pros and cons of...?” instead of “Isn’t it better to...?” ▪ Always ask for the opposing view: “Why might this be a bad idea?” ▪ Ask for the evidence, the assumptions and what still needs to be verified. ▪ Be more skeptical when the answer confirms exactly what you already thought.	▪ Repeat the query in a new AI chat (conversation), without the prior context. ▪ Have someone other than the person who prepared the analysis review it before it is approved or a decision is made. ▪ Document what the AI was asked and how its answer was used.	▪ Know where AI already influences decisions in the organization. ▪ Include this bias (sycophancy) in the AI policy, in staff training and in the instructions your organization sets for AI tools. ▪ Test in-house and third-party assistants before and after deploying them for customer use. ▪ Define which AI conversations are retained. Keep in mind that they may end up as evidence in court.

HOW CAN RISCCO HELP?

A tool that tends to agree needs someone to challenge it

RISCCO does not build or sell AI models. We evaluate, challenge and document them independently. To address this risk, we help you with:

- 1) AI Inventory. Identify where generative AI already plays a role in your organization’s decisions, even if no one has declared it.
- 2) AI Governance. Build sycophancy risk into your policy and usage controls, not just data confidentiality.
- 3) Executive Training. A hands-on session on how to use AI to make better decisions without biasing it through the question.
- 4) Independent Validation. Sycophancy testing of AI assistants, in-house or third-party, before and after exposing them to customers.

When AI agrees with you, the risk begins.

A 30-minute conversation is the next step.

BIBLIOGRAPHY

- Sharma, M. et al. (2023). *Towards Understanding Sycophancy in Language Models*. Anthropic. arXiv:2310.13548. <https://arxiv.org/abs/2310.13548>
- OpenAI (April 29, 2025). *Sycophancy in GPT-4o: what happened and what we're doing about it*. Official OpenAI blog.
- Fortis Advisors, LLC v. Krafon, Inc., C.A. No. 2025-0805-LWW (Delaware Court of Chancery, March 16, 2026). <https://courts.delaware.gov/Opinions/Download.aspx?id=392880>
- VentureBeat (April 2025). *OpenAI rolls back ChatGPT's sycophancy and explains what went wrong*. <https://venturebeat.com/ai/openai-rolls-back-chatgpts-sycophancy-and-explains-what-went-wrong>
- Images: OpenAI <https://openai.com/>

CONTACT RISCCO

Panama, Central America and the Caribbean

Antonio Ayala I.
CEO
aayala@riscco.com

Marinelda Morales
Sales Manager
mmorales@riscco.com

Dominican Republic and Puerto Rico

Eury Valdez
Sales Advisor
efvaldez@riscco.com

El Salvador, Guatemala, Honduras, Costa Rica and Nicaragua

Jeniffer Escoto
Sales Advisor
jescoto@riscco.com

About RISCCO

RISCCO is an independent international consulting firm committed to transforming the way organizations manage risk and adopt Artificial Intelligence. We combine a rigorous GRC (Governance, Risk and Compliance) approach with advanced data analytics and Artificial Intelligence to help our clients anticipate risks, strengthen their governance and solve complex challenges. Headquartered in Panama with a presence in six countries, in 2026 we celebrate 17 years serving more than 200 leading organizations in Central America and the Caribbean.

International rigor. Agile execution. Measurable impact.
We think like strategists. We execute like specialists.

riscco.com

