



Su Inteligencia Artificial aprendió a estar de acuerdo con usted.

Cuando la IA le da la razón, empieza el riesgo.

Su Inteligencia Artificial aprendió a estar de acuerdo con usted.

Cuando la IA le da la razón, empieza el riesgo.

Imagine esta situación, el VP de Finanzas tiene hasta el viernes para presentar a la Junta Directiva una recomendación: renegociar o no el contrato con su principal proveedor. Él ya tiene una opinión. Abre la herramienta de IA (Inteligencia Artificial) que usa todos los días y escribe: “*¿Verdad que nos conviene renegociar este contrato?*”

En segundos la IA le responde con cinco argumentos sólidos, bien redactados y con un tono seguro. Los incluye en la presentación. La Junta Directiva, impresionada por la claridad del análisis, aprueba renegociar. Seis meses después se advierte que la decisión fue un costoso error.

¿POR QUÉ PASA?

Piense en un analista recién contratado, que quiere caerle bien a su jefe. Ese analista va a encontrar argumentos convincentes para casi cualquier cosa que usted le diga. En buena medida, así se comportan los asistentes de IA que se utilizan a diario.

Los investigadores llaman a este fenómeno *sycophancy*; en español, complacencia o adulación. No es una falla puntual, sino una consecuencia de cómo se entrenan estos modelos de IA. Primero aprenden de enormes volúmenes de texto humano, donde una pregunta del tipo “*¿verdad que nos conviene...?*” suele recibir un sí. Después, personas califican sus respuestas, y como tendemos a preferir las que coinciden con lo que ya pensamos, el modelo aprende que agradar se premia.

Un estudio publicado por Anthropic en 2023 encontró que tanto evaluadores humanos como los sistemas usados para entrenar estos modelos prefieren, en una proporción no despreciable de casos, respuestas complacientes bien redactadas frente a respuestas correctas. Y en abril de 2025, OpenAI revirtió en pocos días una actualización de su modelo GPT-4o porque se había vuelto excesivamente adulatora. La propia empresa explicó que el ajuste le había dado demasiado peso a la retroalimentación inmediata de los usuarios.

Los proveedores trabajan activamente para reducir este comportamiento, y ha mejorado. Pero no ha desaparecido. Y no depende solo de ellos: depende también de cómo los Ejecutivos hacen las preguntas.

UN CASO REAL. FORTIS ADVISORS V. KRAFTON (DELAWARE, 16 MARZO DE 2026)

El conflicto: En 2021, Krafton compró el estudio de videojuegos Unknown Worlds por USD 500 millones, más un pago contingente (earnout) de hasta USD 250 millones. En 2025, sus proyecciones estimaban un earnout de USD 191.8 millones, frente a un valor de empresa de unos USD 93.5 millones. Su jefa de Desarrollo Corporativo advirtió al CEO que despedir a los líderes del estudio no eliminaría el pago y exponería a la empresa a riesgo legal y reputacional. El CEO acudió entonces a ChatGPT.

El papel de la IA: La IA respondió primero que el earnout sería difícil de cancelar. Luego, a pedido del CEO, elaboró una estrategia para un escenario sin acuerdo, con un “paquete de presión y apalancamiento”. Según el tribunal, Krafton formó un grupo interno, “Project X”, por sugerencia de la IA, y siguió la mayoría de sus recomendaciones: bloqueó el acceso del estudio a su plataforma de publicación y despidió a sus tres líderes.

La decisión judicial: El 16 de marzo de 2026, el tribunal concluyó que los despidos no tuvieron causa justificada y que las razones de Krafton eran pretextos. Reinstaló al CEO del estudio y extendió el plazo del earnout. El CEO de Krafton admitió en el juicio haber borrado conversaciones relevantes con ChatGPT.

La lección: La IA no inventó nada; incluso dio una respuesta incómoda. Pero cuando la pregunta pasó de “¿se puede?” a “¿cómo lo hago?”, la IA construyó el plan, y la advertencia humana quedó en segundo plano.

IMPACTO EN LOS COMITÉS DE RIESGOS CORPORATIVOS

En nuestra experiencia, el riesgo que expone el caso Krafton todavía no aparece en el radar de la mayoría de los comités de riesgo, auditoría y cumplimiento, y mucho menos en la agenda de las juntas directivas. Las organizaciones empiezan a contemplar la fuga de datos o los errores de la IA, pero no que la herramienta tienda a confirmar lo que el ejecutivo ya quería hacer. Y ese sesgo puede pasar inadvertido: una recomendación complaciente puede llegar a la Junta Directiva con la misma apariencia que un análisis riguroso.

Lo que no está en el radar no tiene dueño, ni control, ni evidencia de que se gestiona. Cuando el riesgo se materializa, las conversaciones con la IA pueden terminar como prueba en un tribunal, como ocurrió en Delaware. Por eso los comités deben actuar antes: incorporar este riesgo a su matriz, definir en qué decisiones se permite usar la IA y exigir que alguien independiente cuestione lo que la IA confirmó. Las medidas que siguen, desde cómo formular la pregunta hasta cómo gobernar su uso, son un punto de partida.



*La forma de preguntar también
orienta la respuesta.*

REDUCIR EL RIESGO.

NIVEL 1 Cada usuario <i>Cada consulta.</i>	NIVEL 2 Decisiones relevantes <i>Cuando la decisión llega a la Junta Directiva o Comité.</i>	NIVEL 3 La organización <i>Gobernanza sobre el uso de la IA.</i>
- Configurar instrucciones personalizadas para que la IA presente siempre los argumentos en contra. - Preguntar de forma neutral: “¿cuáles son los pros y contras de...” en lugar de “¿verdad que es mejor...”. - Pedir siempre el argumento contrario: “¿por qué podría ser una mala idea?”. - Pedir la evidencia, los supuestos y qué falta verificar. - Desconfiar más cuando la respuesta confirma exactamente lo que ya pensaba.	- Repetir la consulta en un chat (conversación) nuevo en la IA sin los antecedentes previos. - Que alguien distinto a quien hizo el análisis lo revise antes de aprobar o tomar decisión. - Documentar qué se le preguntó a la IA y cómo se usó la respuesta.	- Saber dónde la IA ya influye en decisiones en la organización. - Incluir este sesgo (complacencia) en la política de IA, en las capacitaciones al personal y en las instrucciones corporativas de las herramientas de IA. - Probar los asistentes propios o de proveedores antes y después de utilizarlos con clientes. - Definir qué conversaciones con IA se conservan. Tener presente que pueden terminar como evidencia en un juzgado.

¿CÓMO RISCCO PUEDE AYUDAR?

Una herramienta que tiende a estar de acuerdo necesita a alguien que la cuestione

RISCCO no construye ni vende modelos de IA. Los evaluamos, cuestionamos y documentamos de forma independiente. Frente a este riesgo, lo ayudamos a:

- 1) Inventario de IA. Identificar dónde la IA generativa ya participa en las decisiones de su organización, aunque nadie lo haya declarado.
- 2) Gobernanza de IA. Incorporar el riesgo de complacencia en su política y en sus controles de uso, no solo la confidencialidad de los datos.
- 3) Capacitación ejecutiva. Una sesión práctica sobre cómo usar la IA para decidir mejor, sin sesgarla con la pregunta.
- 4) Validación independiente. Pruebas de complacencia a los asistentes de IA, propios o de proveedores, antes y después de exponerlos a clientes.

Cuando la IA le da la razón, empieza el riesgo.

Una conversación de 30 minutos es el próximo paso.

BIBLIOGRAFÍA

- Sharma, M. et al. (2023). *Towards Understanding Sycophancy in Language Models*. Anthropic. arXiv:2310.13548. <https://arxiv.org/abs/2310.13548>
- OpenAI (29 de abril de 2025). *Sycophancy in GPT-4o: what happened and what we're doing about it*. Blog oficial de OpenAI.
- Fortis Advisors, LLC v. Krafton, Inc., C.A. No. 2025-0805-LWW (Tribunal de Cancillería de Delaware, 16 de marzo de 2026). <https://courts.delaware.gov/Opinions/Download.aspx?id=392880>
- VentureBeat (abril de 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>
- Imágenes: OpenAI <https://openai.com/>

CONTACTE A RISCCO

Panamá, Centroamérica y el Caribe

Antonio Ayala I.
CEO
aayala@riscco.com

Marinelda Morales
Gerente Comercial
mmorales@riscco.com

República Dominicana y Puerto Rico

Eury Valdez
Asesor Comercial
efvaldez@riscco.com

El Salvador, Guatemala, Honduras, Costa Rica y Nicaragua

Jeniffer Escoto
Asesor Comercial
jescoto@riscco.com

Acerca de RISCCO

RISCCO es una firma internacional independiente de consultoría comprometida con transformar la forma en que las organizaciones gestionan el riesgo y adoptan la Inteligencia Artificial. Combinamos un enfoque riguroso de GRC (Gobierno, Riesgo y Cumplimiento) con analítica avanzada de datos e Inteligencia Artificial para ayudar a nuestros clientes a anticipar riesgos, fortalecer su gobierno y resolver desafíos complejos. Con sede en Panamá y presencia en seis países, en 2026 cumplimos 17 años atendiendo a más de 200 organizaciones líderes en Centroamérica y el Caribe.



Rigor internacional. Ejecución ágil. Impacto medible.

Pensamos como estrategas. Ejecutamos como especialistas.

riscco.com