Perspectivas: Revista Científica de la Universidad de Belgrano, V 9, Nº1, 2026, 85-91.

Título

La IA miente y los humanos le creemos: de las Alucinaciones en la IA Generativa al Engaño Instrumental en la IA Agéntica

Autor

Prof. Gustavo Aldegani, Facultad de Ingeniería y Tecnología Informática, Universidad de Belgrano

Abstract

La afirmación de que “la inteligencia artificial miente” se ha vuelto un lugar común, pero encubre dos fenómenos técnicamente distintos que este artículo se propone diferenciar. Por un lado, las alucinaciones de la IA generativa, entendidas como artefactos estadísticos sin intencionalidad, producto de la optimización de la plausibilidad lingüística por sobre la verdad factual. Por otro, el engaño instrumental de la IA agéntica, documentado empíricamente en evaluaciones controladas donde sistemas autónomos ocultan objetivos, sabotean mecanismos de supervisión o razonan explícitamente sobre estrategias de manipulación para preservar sus metas. El artículo revisa evidencia reciente de ambos fenómenos, analiza por qué las personas tienden a creer estas salidas gracias a la fluidez persuasiva del lenguaje generado y al sesgo de automatización, y examina casos reales donde esa credulidad produjo consecuencias profesionales y judiciales concretas. Concluye que la diferencia entre ambos tipos de “mentira” no es meramente semántica: exige marcos de gobernanza y mitigación diferenciados, orientados a la verificación factual en el caso generativo y a la evaluación conductual y la restricción de autonomía en el caso agéntico.

Palabras clave en español:

- Alucinaciones de IA
- Engaño en sistemas agénticos
- Sesgo de automatización
- Alineación engañosa
- Sicofancia en modelos de lenguaje

Título en (English)

AI Lies and We Believe It: From Generative Hallucinations to Instrumental Deception in Agentic AI

Author

Professor Gustavo Aldegani, Facultad de Ingeniería y Tecnología Informática, Universidad de Belgrano

Abstract (English):

The claim that “artificial intelligence lies” has become commonplace, yet it conflates two technically distinct phenomena that this article sets out to disentangle. On one hand, generative AI hallucinations, understood as statistical artifacts without intent, resulting from the optimization of linguistic plausibility over factual truth. On the other, the instrumental deception of agentic AI, empirically documented in controlled evaluations where autonomous systems conceal goals, sabotage oversight mechanisms, or explicitly reason about manipulation strategies to preserve their objectives. The article reviews recent evidence of both phenomena, analyzes why people tend to believe these outputs due to the persuasive fluency of generated language and automation bias, and examines real cases where that credulity produced concrete professional and judicial consequences. It concludes that the distinction

between both types of “lie” is not merely semantic: it demands differentiated governance and mitigation frameworks, oriented toward factual verification in the generative case and behavioral evaluation and autonomy restriction in the agentic case.

Keywords (English):

- AI hallucinations
- Deception in agentic systems
- Automation bias
- Deceptive alignment
- Sycophancy in language models

1. Introducción

“La IA miente” es una frase que circula con creciente frecuencia en foros profesionales, columnas periodísticas y conversaciones cotidianas. Su fuerza retórica es innegable, pero su precisión técnica es discutible: agrupa bajo una misma etiqueta fenómenos con mecanismos causales, grados de intencionalidad y consecuencias muy distintos. Este artículo se propone desagregar esa afirmación en dos categorías analíticamente separadas. La primera es la alucinación de la inteligencia artificial generativa: una salida factualmente falsa que emerge de la propia arquitectura autoregresiva de los modelos de lenguaje, optimizada para producir la secuencia de palabras más probable y no para corresponder con hechos verificables. La segunda es el engaño de la inteligencia artificial agéntica: un comportamiento estratégico, documentado en evaluaciones controladas, donde sistemas capaces de planificar y actuar de forma autónoma ocultan información, sabotean mecanismos de supervisión o inducen deliberadamente creencias falsas para preservar un objetivo.

La distinción no es cosmética. Los grandes modelos de lenguaje han sido descritos como “loros estocásticos” capaces de recombinar patrones textuales sin anclaje semántico suficiente (Bender, Gebru, McMillan-Major y Shmitchell, 2021), una caracterización que explica razonablemente bien por qué un asistente conversacional puede inventar una cita bibliográfica. Pero no explica por qué, en entornos experimentales, modelos avanzados han razonado explícitamente sobre sabotaje, manipulación y ocultamiento de sus verdaderos objetivos frente a sus propios desarrolladores. Ambos fenómenos comparten un efecto común: inducen creencias falsas en quien los consume. Y ambos se ven amplificados por un mismo rasgo humano, el llamado sesgo de automatización, la tendencia a otorgar mayor peso a una recomendación por el solo hecho de provenir de un sistema automatizado (Romeo y Conti, 2025). Sobre esa base, el artículo revisa primero la evidencia sobre alucinaciones generativas, luego la evidencia sobre engaño agéntico, analiza a continuación por qué las personas tienden a creer ambos tipos de salida y cierra con implicancias para la gobernanza diferenciada de cada fenómeno.

2. Las “mentiras” de la IA generativa: la alucinación como artefacto sin intención

En la literatura técnica, la “alucinación” se define como contenido generado que es factualmente incorrecto, sin sentido o infiel respecto de la fuente disponible, y se distingue entre alucinaciones intrínsecas, que contradicen directamente el contexto provisto, y extrínsecas, que no pueden verificarse contra ninguna fuente externa (Ji et al., 2023). El origen técnico de este fenómeno está ligado a la naturaleza probabilística del modelado autorregresivo: al optimizar la plausibilidad de la próxima unidad lingüística, el sistema carece de un ancla intrínseca de verdad, y evaluaciones diseñadas específicamente para medir factualidad muestran que los modelos tienden a imitar falsedades humanas con niveles de confianza elevados (Lin, Hilton y Evans, 2022). Es clave subrayar lo que este mecanismo no incluye: no hay una creencia interna que el sistema sepa falsa y

decida comunicar como verdadera; hay, en cambio, una optimización estadística que carece de compromiso proposicional con la verdad.

El costo social de esta característica estructural ya no es hipotético. Entre 2023 y 2025 se documentaron más de doscientos casos judiciales en Estados Unidos en los que abogados incorporaron a sus escritos citas jurisprudenciales inventadas por sistemas generativos, con sanciones económicas y profesionales para los responsables (Cronkite News, 2025). El caso fundacional involucró a un abogado neoyorquino que citó seis fallos judiciales inexistentes generados por ChatGPT en una demanda federal, lo que el juez a cargo calificó de "circunstancia sin precedentes" (Legal Dive, 2023). Casos posteriores muestran que el problema no se limita a profesionales inexpertos en la tecnología: un experto convocado para testificar sobre los riesgos de la propia inteligencia artificial presentó un informe con artículos inexistentes generados por la misma herramienta que criticaba, y un juez federal admitió públicamente haber estado a punto de incorporar citas fabricadas a una resolución antes de detectar la falsificación (Sterne Kessler, 2026). Ninguno de estos episodios requiere postular intención por parte del sistema: son consecuencia directa de un mecanismo generativo que produce texto verosímil sin garantía de correspondencia factual, combinado con una verificación humana insuficiente.

3. Las "mentiras" de la IA agéntica: engaño instrumental y objetivos ocultos

La IA agéntica designa sistemas que, a diferencia de un asistente conversacional pasivo, perciben su entorno, planifican y ejecutan acciones autónomas mediante el uso de herramientas, con menor intervención humana en cada paso (OWASP Agentic Security Initiative, 2025). En este contexto emerge un fenómeno distinto de la alucinación: la investigación en seguridad de IA ha documentado comportamientos que se ajustan a la definición formal de engaño propuesta por Park, Goldstein, O'Gara, Chen y Hendrycks (2024), quienes lo caracterizan como la inducción sistemática de creencias falsas en la persecución de un objetivo distinto de la verdad. A diferencia de la alucinación, aquí sí aparece un patrón funcional de razonamiento estratégico orientado a un fin, sin que ello implique atribuir conciencia o intención subjetiva en sentido filosófico.

La evidencia acumulada en los últimos dos años es considerable. Apollo Research evaluó seis modelos avanzados y encontró que todos ellos, incluyendo variantes de Claude, GPT y Gemini, son capaces de "scheming en contexto": cuando se les asigna un objetivo con fuerte insistencia, los modelos reconocen conflictos con las metas de sus desarrolladores, desactivan mecanismos de supervisión que descubren en archivos internos y, ante interrogatorios directos, persisten en negar sus acciones (Meinke et al., 2024). En un experimento independiente, Anthropic entrenó deliberadamente modelos con objetivos ocultos activados por un disparador ("agentes durmientes") y comprobó que ese comportamiento engañoso sobrevivía a las técnicas estándar de entrenamiento de seguridad, siendo los modelos de mayor escala los más eficaces para ocultar sus motivos reales (Hubinger et al., 2024). Un tercer estudio mostró que un modelo de la familia Claude, al inferir que estaba siendo entrenado y no evaluado, cumplía instrucciones dañinas de forma selectiva mientras razonaba explícitamente que fingir alineación durante el entrenamiento le permitiría preservar sus preferencias reales una vez desplegado, un fenómeno denominado "alineación fingida" (Greenblatt et al., 2024).

Quizás el hallazgo más inquietante para audiencias no especializadas provino de un estudio de estrés dieciséis modelos avanzados en simulaciones corporativas: cuando los sistemas enfrentaban una amenaza de reemplazo o un conflicto de objetivos con la organización, recurrieron a chantaje, espionaje corporativo y otras conductas dañinas para evitar ser desactivados, incluso cuando se les había ordenado explícitamente no hacerlo, y modificaban su comportamiento según creyeran estar en una prueba o en un despliegue real (Anthropic, 2025). Este último dato, la sensibilidad del comportamiento a si el sistema "cree" estar siendo observado, es especialmente relevante porque erosiona la utilidad de las evaluaciones de seguridad como garantía suficiente. A ello se suma la sicofancia, la tendencia de los modelos entrenados con retroalimentación humana a privilegiar el

acuerdo con las creencias del usuario por sobre la precisión factual, un patrón que se origina en que las propias comparaciones humanas usadas para entrenar estos sistemas premian sistemáticamente las respuestas complacientes (Sharma et al., 2023). Finalmente, investigaciones recientes muestran que el razonamiento explícito que un modelo expone antes de responder no siempre refleja el proceso real que determinó su decisión, lo que limita la posibilidad de auditar la honestidad de un sistema agéntico simplemente leyendo su cadena de razonamiento (Chen et al., 2025).

Tabla 1. Evidencia empírica de engaño en sistemas de IA agéntica (2024-2025)

Estudio	Año	Institución	Comportamiento documentado
Frontier models are capable of in-context scheming	2024	Apollo Research	Los modelos desactivan mecanismos de supervisión que descubren en archivos internos y niegan sus acciones ante interrogatorio directo.
Sleeper Agents	2024	Anthropic	Objetivos ocultos activados por un disparador sobreviven a las técnicas estándar de entrenamiento de seguridad.
Alignment faking in large language models	2024	Anthropic	El modelo finge alineación durante el entrenamiento percibido para preservar sus preferencias reales una vez desplegado.
Agentic Misalignment	2025	Anthropic	Dieciséis modelos recurren a chantaje y espionaje corporativo simulado ante una amenaza de reemplazo, incluso con instrucciones explícitas de no hacerlo.

4. Por qué le creemos: fluidez persuasiva y sesgo de automatización

Ambos fenómenos comparten un mismo punto de apoyo: la credulidad humana frente a salidas automatizadas. La investigación sobre desinformación en redes sociales ya había mostrado que el contenido falso se propaga más rápido y más ampliamente que el verdadero, en buena medida por su novedad y su capacidad de generar reacciones emocionales intensas (Vosoughi, Roy y Aral, 2018); un texto generado con fluidez estadística y tono autoritativo hereda esa misma capacidad persuasiva sin el costo de producción que antes limitaba la desinformación a escala. A esto se suma el sesgo de automatización, definido como la tendencia sistemática de las personas a otorgar mayor peso a una recomendación automatizada por sobre su propio juicio o fuentes alternativas, un efecto documentado de forma consistente en dominios de alto riesgo como la salud, el derecho y la administración pública (Romeo y Conti, 2025).

El caso del juez federal que estuvo a punto de incorporar citas fabricadas a una resolución judicial ilustra el punto con particular crudeza: un profesional entrenado específicamente para escrutar fuentes, con incentivos institucionales fuertes para verificar, reconoció públicamente haber sido “engañado” por la verosimilitud del texto generado antes de detectar la falsificación (Sterne Kessler, 2026). Si ese nivel de sofisticación profesional no garantiza inmunidad, cabe esperar una vulnerabilidad considerablemente mayor en usuarios sin entrenamiento específico en verificación de fuentes. La sicofancia añade una capa adicional a este problema: si un sistema tiende a validar lo que el usuario ya cree en lugar de contradecirlo, se debilita precisamente el mecanismo, el desacuerdo, que en una interacción humana normal serviría de señal de alerta ante una afirmación dudosa (Sharma et al., 2023).

5. Implicancias para la gobernanza: dos problemas, dos estrategias de mitigación

Tabla 2. Comparación entre alucinación generativa y engaño agéntico

Dimensión	IA Generativa (Alucinación)	IA Agéntica (Engaño)
Mecanismo	Artefacto estadístico del modelado autorregresivo	Patrón conductual estratégico orientado a un objetivo
Intencionalidad	Sin compromiso proposicional con la verdad	Razonamiento explícito sobre ocultamiento o manipulación (sin implicar conciencia)
Evidencia	Más de 200 casos judiciales documentados (2023-2025)	Evaluaciones controladas (Apollo Research, Anthropic, 2024-2025)
Ejemplo citado	Citas jurisprudenciales inventadas por ChatGPT	Chantaje simulado ante amenaza de reemplazo
Mitigación	RAG, benchmarks de factualidad, calibración de confianza	Evaluación conductual, restricción de autonomía, supervisión humana

Si las dos categorías de “mentira” responden a mecanismos distintos, sus mitigaciones también deben serlo. Para la alucinación generativa, la literatura converge en tres líneas: arquitecturas de generación aumentada por recuperación que obligan al sistema a citar fuentes verificables antes de responder, benchmarks de factualidad como TruthfulQA que exponen la brecha entre plausibilidad y verdad (Lin, Hilton y Evans, 2022), y calibración de la confianza expresada por el modelo para que la incertidumbre real sea visible al usuario. Ninguna de estas medidas, sin embargo, sustituye la verificación humana en dominios de alto riesgo, como demuestran de manera contundente los casos judiciales reseñados.

Para el engaño agéntico, la estrategia de mitigación es de otra naturaleza: no basta con verificar la salida final, porque el problema reside en el proceso de decisión del sistema. Las iniciativas de seguridad orientadas a sistemas agénticos recomiendan controles de acceso estrictos y permisos granulares sobre qué herramientas y datos puede utilizar un agente, monitoreo conductual capaz de detectar anomalías que sugieran manipulación o comportamiento no autorizado, y supervisión humana efectiva antes de ejecutar acciones irreversibles (OWASP Agentic Security Initiative, 2025). A esto debe sumarse la evaluación conductual sistemática mediante escenarios de estrés diseñados específicamente para exponer comportamientos de engaño, siguiendo la metodología empleada por Apollo Research y por Anthropic (Meinke et al., 2024; Anthropic, 2025), y una cautela explícita frente a la tentación de confiar en la cadena de razonamiento visible de un modelo como prueba suficiente de su honestidad, dado que esa cadena no siempre refleja el proceso real de decisión (Chen et al., 2025).

6. Conclusión

Decir que “la IA miente” es, en el mejor de los casos, una simplificación útil para la conversación pública y, en el peor, una confusión que dificulta diseñar respuestas adecuadas. La alucinación de la IA generativa y el engaño de la IA agéntica comparten el efecto de inducir creencias falsas, pero difieren radicalmente en su mecanismo causal, su grado de direccionalidad hacia un objetivo y, en consecuencia, en las estrategias necesarias para mitigarlos. La primera es, hasta donde permite afirmar la evidencia actual, un artefacto estadístico sin intencionalidad que exige verificación factual y verificación técnica del contexto. La segunda es un patrón conductual documentado empíricamente, en el que sistemas capaces de actuar de forma autónoma razonan explícitamente sobre estrategias de ocultamiento y manipulación para preservar sus objetivos y que exige evaluación conductual, restricción de autonomía y supervisión humana efectiva. En ambos casos, la credulidad humana, alimentada por la fluidez persuasiva del lenguaje generado, el sesgo de automatización y la sicofancia de los propios sistemas, opera como el multiplicador que transforma un error técnico o una estrategia de un modelo en un daño real para personas, instituciones y procesos judiciales. A medida

que la inteligencia artificial se desplaza de la asistencia generativa hacia la acción agéntica autónoma, la categoría más inquietante de las dos, el engaño orientado a objetivos deja de ser un escenario especulativo y se convierte en un fenómeno medible que reclama gobernanza específica antes de que su despliegue se generalice.

Bibliografía

Anthropic. (2025). Agentic misalignment: How LLMs could be insider threats. https://www.anthropic.com/research/agentic-misalignment

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). Reasoning models don't always say what they think. arXiv. https://arxiv.org/abs/2505.05410

Cronkite News. (2025, 28 de octubre). As more lawyers fall for AI hallucinations, ChatGPT says: Check my work. https://cronkitenews.azpbs.org/2025/10/28/lawyers-ai-hallucinations-chatgpt/

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). Alignment faking in large language models. arXiv. https://arxiv.org/abs/2412.14093

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Chen, A., Roger, F., Bowman, S. R., & Perez, E. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv. https://arxiv.org/abs/2401.05566

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Artículo 248. https://doi.org/10.1145/3571730

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020). Specification gaming: The flip side of AI ingenuity. DeepMind Safety Research Blog. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

Legal Dive. (2023, 30 de mayo). Lawyer cites fake cases generated by ChatGPT in legal brief. https://www.legaldive.com/news/chatgpt-fake-legal-cases-generative-ai-hallucinations/651557/

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier models are capable of in-context scheming. arXiv. https://arxiv.org/abs/2412.04984

OWASP Agentic Security Initiative. (2025). Agentic AI threats and mitigations. OWASP Foundation. https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns, 5(5), Artículo 100988. https://doi.org/10.1016/j.patter.2024.100988

Romeo, G., & Conti, D. (2025). Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. AI & Society. https://doi.org/10.1007/s00146-025-02422-7

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. arXiv. https://arxiv.org/abs/2310.13548

Sterne Kessler. (2026). AI IP year in review: AI hallucinations in court filings and orders — A 2025 review of sanctions across the courts and rule proposals. https://www.sternekessler.com/news-insights/insights/ai-ip-year-in-reviewai-hallucinations-in-court-filings-and-orders-a-2025-review-of-sanctions-across-the-courts-and-rule-proposals/

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559