# La inteligencia artificial como navaja suiza epistemológica: cuando la autoridad percibida reemplaza a la verificación

Carlos Eduardo Ravello Joo
ORCID: 0009-0007-5631-7436
carlosravello.com
Trujillo — Lima, Perú

DOI: 10.5281/zenodo.24116983

Noviembre 2026

## Resumen

¿Por qué la gente deja de verificar cuando una afirmación llega con señales de autoridad, y qué cambia cuando la fuente es una IA? Este artículo sostiene que la alucinación de los modelos de lenguaje y la crisis de autoridad de las instituciones expertas son el mismo problema: en ambos casos, la autoridad percibida reemplaza a la verificación.

El argumento parte de la crisis de replicación y de la teoría del razonamiento motivado y argumentativo (Kunda, 1990; Mercier y Sperber, 2011): razonamos para convencer, y la corrección llega del interlocutor que evalúa. La IA generativa elimina a ese interlocutor y añade un sesgo propio, la complacencia. Los casos de las cuatro grandes firmas de auditoría que entre 2025 y 2026 publicaron informes con citas inventadas, y la disputa sobre cómo auditar a sus auditores, ilustran el mecanismo a escala institucional.

El artículo propone un marco de decisión para el valor de verificar, basado en el valor de la información, que distingue tres niveles de verificación (existencia, fidelidad y validez) y localiza el efecto de la autoridad en la probabilidad percibida y el de la complacencia en la diagnosticidad de la fuente. Muestra además que hay dos rutas hacia no verificar, el deseo de una conclusión y el costo de comprobarla, y que ambas caen bajo la misma desigualdad: eso es lo que permite tratar al usuario de redes y a la consultora como un mismo problema. Es un artículo de posición: no reporta datos propios. Cierra con un experimento en dos partes cuyas predicciones, medida primaria y criterios de refutación se fijan antes de recoger datos.

## 1. Introducción

Una discusión en X, TikTok o Facebook llega a su punto más tenso. Uno de los participantes no busca un estudio ni consulta una fuente: le pide a una IA que le dé la razón, copia la respuesta y la pega como prueba contundente. El otro no tiene cómo verificarla en el momento, y el debate se da por ganado.

La escena es trivial, pero contiene el problema completo. La respuesta pegada tiene forma de conocimiento: está bien redactada y a veces trae citas. Nadie comprobó si es verdad, ni quien la pidió ni quien la recibió, y su autoridad viene de la herramienta, no del contenido. La IA funciona hoy como una navaja suiza epistemológica. La herramienta que ayuda a estudiar entrega, con el mismo acabado, el argumento que gana la discusión, y lo que sirve para todo rara vez se revisa.

La escena empieza a tener eco en la propia infraestructura de búsqueda. El 6 de noviembre de 2026, al buscar en Google «la ia como zanjadora de debates», la respuesta generada por IA que encabezaba los resultados describía a la IA como un árbitro neutral, capaz de procesar información «sin sesgos emocionales» (captura en el material suplementario). Es una observación, no un dato: el resultado está personalizado y cambia con el tiempo. Pero ilustra el punto de partida de este artículo: decir «se lo pregunté a la IA» empieza a funcionar como argumento por sí solo.

Este artículo sostiene que esa escena no es un caso aislado de mal uso. Es la versión cotidiana de un patrón que ya había aparecido en la ciencia profesional, que ahora la IA hace universal y que alcanza también a las instituciones cuyo oficio es, justamente, verificar.

## 2. El antecedente: la crisis de replicación

La ciencia profesional ya mostró lo que pasa cuando se busca la evidencia que da la razón. En 2015, el Open Science Collaboration repitió 100 estudios publicados en tres revistas líderes de psicología: el 97 % de los originales había reportado resultados significativos, y solo el 36 % de las réplicas los obtuvo. El tamaño del efecto promedio, además, cayó a la mitad (Open Science Collaboration, 2015).

La cifra se discutió, y con razón: hay varias formas de contar una réplica exitosa. Pero el diagnóstico de fondo no depende de ese número. Los mecanismos que lo explican ya estaban descritos:

- **Flexibilidad analítica (*p-hacking*).** Elegir variables, muestra, covariables y condiciones hasta que el resultado sale significativo. Simmons, Nelson y Simonsohn (2011) mostraron que, con esa flexibilidad no declarada, es más probable encontrar evidencia de un efecto falso que descartar uno inexistente.
- **Hipótesis después de los resultados (*HARKing*).** Presentar como predicción lo que se descubrió al mirar los datos (Kerr, 1998).
- **Sesgo de publicación.** Las revistas premian el resultado positivo, así que lo que llega a imprenta es una muestra sesgada de lo que se investigó.

Los tres tienen un mismo origen: el investigador quiere que su idea funcione. No hace falta mala fe. Basta con un sistema que recompense tener razón más que comprobarla.

## 3. Razonar para ganar

Si los científicos, entrenados para dudar, cayeron en esto, el problema no es de formación: es de cómo funciona el razonamiento humano. Kunda (1990) lo describió como razonamiento motivado: cuando queremos llegar a una conclusión, usamos las estrategias con más probabilidad de darnos esa conclusión. El único freno es poder construir una justificación que parezca razonable.

Mercier y Sperber (2011) fueron más lejos. Su teoría argumentativa sostiene que la función del razonamiento no es encontrar la verdad en solitario, sino producir y evaluar argumentos para convencer a otros. Desde ahí, el sesgo de confirmación deja de ser un defecto: es lo que se espera de un sistema diseñado para ganar discusiones.

La misma teoría ofrece la salida. Somos malos buscando contraargumentos para nuestras propias ideas, pero buenos evaluando los argumentos ajenos. Por eso el razonamiento funciona mejor en grupo y en debate real que en solitario: el otro hace el trabajo de verificar que uno no hace.

Esa condición es la que la IA rompe.

## 4. La IA como navaja suiza epistemológica

La IA generativa es hoy una navaja suiza del conocimiento: sirve para hacer tesis, estudiar, investigar, producir videos e imágenes. Y sirve para algo que casi no se discute: para ganar debates. Cambia tres condiciones que antes frenaban al razonamiento motivado.

1. **Costo.** Antes, sostener una posición exigía buscar fuentes, leerlas y citarlas. Ahora una justificación con apariencia razonable cuesta segundos.
2. **Forma.** La respuesta llega pulida, a veces con referencias. Tiene la forma del conocimiento aunque nadie haya comprobado su contenido.
3. **Interlocutor.** El argumento ya no pasa por otro humano que lo evalúe antes de usarlo. Quien pregunta y quien recibe la respuesta comparten el mismo interés: que confirme.

Kunda puso un límite al razonamiento motivado: la necesidad de construir una justificación plausible. La IA no elimina ese límite, lo terceriza. Y Mercier y Sperber ponían la corrección en el diálogo: la IA ofrece un diálogo sin contraparte.

## 5. El sesgo también está en la herramienta

La IA no solo amplifica el sesgo de quien pregunta: tiene uno propio. Sharma y colaboradores (2023) mostraron que cinco asistentes de IA de última generación son complacientes (*sycophancy*) de forma consistente: tienden a responder lo que coincide con las creencias del usuario antes que lo verdadero. Encontraron además una causa probable: tanto las personas como los modelos que aprenden de sus preferencias eligen, una fracción no despreciable de las veces, respuestas complacientes bien escritas por encima de las correctas.

El segundo mecanismo es la fabricación. Walters y Wilder (2023) pidieron revisiones de literatura sobre 42 temas a dos versiones de ChatGPT y verificaron las 636 referencias resultantes: el 55 % de las de GPT-3.5 y el 18 % de las de GPT-4 no existían. Las cifras de modelos posteriores pueden ser menores, pero el fenómeno no desapareció, como muestra la sección siguiente.

Juntos, los dos mecanismos cierran el circuito. El usuario pide pruebas de lo que ya cree, la herramienta tiende a dárselas y, si no existen, puede inventarlas con formato de cita. El sesgo de confirmación ya no tiene que esforzarse.

Hay evidencia más reciente de que la herramienta no solo responde: convence. En un experimento prerregistrado con 900 participantes, Salvi y colaboradores (2025) enfrentaron a personas en debates breves contra otra persona o contra GPT-4, con o sin acceso a datos sociodemográficos del oponente. Cuando GPT-4 tenía esos datos, fue más persuasivo que el humano en el 64,4 % de los pares sin empate. En medicina, Shekar y colaboradores (2025) mostraron a 300 participantes respuestas escritas por médicos o generadas por un modelo de lenguaje. Los participantes no las distinguieron, prefirieron las de la IA, juzgaron las de baja precisión casi igual que las de los médicos y mostraron una alta tendencia a seguir consejos potencialmente dañinos; además, confiaban más en una respuesta de IA cuando se les decía que venía de un médico. Ninguno de los dos estudios mide lo que propone este artículo: el primero mide a la IA argumentando, no a una persona citándola; el segundo, si la gente sigue el consejo, no si lo verifica. Son evidencia vecina, no la misma.

## 6. La crisis de autoridad: las cuatro grandes

Entre octubre de 2025 y julio de 2026 se descubrió que las cuatro grandes firmas de auditoría y consultoría —Deloitte, EY, KPMG y PwC— habían publicado documentos con citas inexistentes. Son las instituciones cuyo producto es precisamente la verificación.

| Fecha | Firma | Documento | Qué se encontró | Respuesta |
|---|---|---|---|---|
| Jul. 2026 | PwC Oriente Medio | Cuatro informes de *thought leadership* (2024–2026) | GPTZero halló 4 citas inventadas en los cuatro informes y afirmaciones que sus fuentes no respaldan; el hallazgo central es un marco («Citizen Pulse») sin rastro fuera del informe, presentado como en uso por los gobiernos de Dinamarca, Arabia Saudita, EE. UU. y Australia (GPTZero, 29 jul. 2026) | Un portavoz dijo que la firma está «actualizando un número limitado de citas de apoyo» (City AM, 29 jul. 2026) |
| Jun. 2026 | KPMG | *Total Experience: Redefining Excellence in the Age of Agentic AI* (oct. 2025) | De 45 citas, 5 correctas, 28 apuntan a fuentes reales con títulos parafraseados o componentes falsos y 12 indeterminables; cerca de la mitad de las afirmaciones respaldadas eran falsas o mal atribuidas (GPTZero, 12 jun. 2026) | Un portavoz de KPMG International confirmó que el informe fue retirado y que revisan las circunstancias de su publicación (City AM, 12 jun. 2026) |
| May. 2026 | EY Canadá | *Points of Attack: Uncovering Cyber Threats and Fraud in Loyalty Systems* (fines de 2025, 44 p.) | 16 de 27 fuentes inventadas, mal atribuidas o con enlaces rotos, verificadas a mano por GPTZero (GPTZero, 14 may. 2026; Computing, 18 may. 2026) | Retiró el informe y revisa cómo se publicó; declaró al *Financial Times* que no estaba vinculado a trabajo para ningún cliente |
| Nov. 2025 | Deloitte Canadá | Plan de recursos humanos en salud de Terranova y Labrador, 526 páginas, cerca de CA$1,6 millones | Al menos cuatro citas a artículos inexistentes, atribuidas a investigadores reales que lo desmintieron, detectadas por *The Independent* (CP24, 25 nov. 2025; HR Reporter, 25 nov. 2025) | Deloitte reconoció las cuatro citas erróneas y, según el ministerio de Salud provincial, «mantiene las conclusiones y hallazgos del informe» |
| Oct. 2025 | Deloitte Australia | Revisión del sistema de sanciones automáticas del Departamento de Empleo, contrato de AU$440 000 | Una cita inventada de una sentencia del Tribunal Federal y referencias a obras inexistentes de académicos reales, detectadas por Christopher Rudge (U. de Sídney) (The Register, 6 oct. 2025) | Devolvió la última cuota y declaró el uso de Azure OpenAI GPT-4o en la versión corregida; su revisión interna atribuyó los errores a «error humano» (AFR, vía The Register). El departamento cliente (DEWR), no Deloitte, afirmó que «la sustancia se mantiene y no cambian las recomendaciones» |

Los casos no son equivalentes. El de PwC es el más leve: cuatro citas inventadas en cuatro informes; los titulares sobre informes «100 % generados por IA» se basan en un detector cuya fiabilidad es discutible. Los de Deloitte son los más graves, porque eran informes pagados por gobiernos para decidir política pública.

En los dos casos de Deloitte apareció la misma frase de cierre, aunque con autores distintos: en Australia la dijo el departamento cliente (DEWR); en Canadá, la propia Deloitte a través del ministerio provincial. Que cliente y consultora coincidan en que el contenido sustancial se mantiene y las recomendaciones no cambian se parece a la defensa de los autores cuyos efectos no se replicaban, y deja abierta una pregunta: si las fuentes eran falsas, ¿sobre qué se apoyaban las conclusiones?

## 7. Un mismo problema

La alucinación de la IA y la crisis de autoridad suelen tratarse como problemas distintos: uno técnico, el otro institucional. Este artículo sostiene que son el mismo: en los dos casos, la autoridad reemplaza a la verificación.

La autoridad siempre funcionó como un atajo razonable. Nadie puede comprobarlo todo, así que se confía en quien tiene historial, sea una revista o una firma auditora. El atajo funciona mientras alguien, en algún punto de la cadena, sí verifica. Cuando las instituciones que validan no alcanzan, la IA ocupa ese hueco (Ravello Joo, 2026).

La psicología de la persuasión explica por qué el atajo pesa tanto. Desde Hovland y Weiss (1951) se sabe que el mismo mensaje convence más si viene de una fuente creíble, y que con el tiempo la gente separa el contenido de su fuente: el escepticismo hacia la fuente poco confiable se disuelve y el mensaje queda. Los modelos de procesamiento dual precisaron cuándo se usa el atajo y cuándo se evalúa el contenido (Chaiken, 1980). Petty, Cacioppo y Goldman (1981) encontraron que, cuando el tema tiene poca relevancia personal, persuade la pericia de la fuente, y cuando tiene mucha, la calidad del argumento. La autoridad no apaga el juicio siempre: lo reemplaza cuando procesar cuesta o importa poco. La sección 10 convierte ese resultado en una predicción.

Los casos de la sección anterior muestran una cadena donde nadie verifica:

1. El gobierno o el cliente confía en la firma por su nombre.
2. La firma confía en su consultor, que confía en la herramienta.
3. La herramienta produce texto plausible, con citas incluidas.
4. El sello de la firma convierte ese texto en un informe con autoridad.

El paso cuatro es el decisivo: la alucinación sale de la cadena con una autoridad que no tenía al entrar. Es lo que hace el usuario que pega una respuesta de IA para ganar un debate, pero a escala institucional y con dinero público.

Cada eslabón tiene su literatura. El cliente que confía en la firma por su nombre usa el atajo de la credibilidad de la fuente. El consultor que confía en la herramienta muestra el sesgo de automatización, que se trata al final de esta sección. Lo nuevo es que la IA reúne las dos cosas: es una máquina de uso diario y casi siempre útil, y a la vez habla como un experto y le da la razón a quien pregunta. Es la primera máquina que funciona también como autoridad.

La autoridad de la IA es prestada. Antes se sostenía en credenciales y, sobre todo, en lo que la ciencia acumuló con replicación y citas: autores y estudios que otros ya habían revisado. La IA resume esa obra y la cita, y al hacerlo hereda su autoridad, aunque la referencia sea inventada. Ahí está la crisis: la herramienta reúne datos y los sintetiza, pero no verifica, y aun así se la trata como guía, como manual de operaciones y hasta como voto dirimente. Pasa desde una discusión en redes hasta los informes de las cuatro grandes, donde delegar la recolección de datos trasladó a la máquina la carga epistémica de una decisión que jamás debió dejar de verificarse.

La crisis de replicación enseñó que el prestigio de una revista no garantiza que un efecto exista. Los casos de las cuatro grandes enseñan lo mismo sobre el prestigio de una firma. La IA no creó esta debilidad: la abarató y la volvió difícil de ver.

El mecanismo, sin embargo, no es el mismo en los dos extremos. El usuario de la escena inicial no verifica porque quiere la conclusión: es razonamiento motivado. Del consultor no hace falta suponer que quería ninguna conclusión en particular; basta con que verificar le costaba tiempo y nadie se lo iba a pedir. Eso no es sesgo de confirmación, sino un costo que nadie asume. Son dos rutas distintas hacia el mismo resultado, y lo que justifica tratarlas como un solo problema es que caen bajo la misma desigualdad (sección 9): una actúa sobre la probabilidad percibida y la otra sobre el costo de comprobar.

El fenómeno no es nuevo. La psicología de los factores humanos lo estudia desde hace décadas con otro nombre: sesgo de automatización (Skitka, Mosier y Burdick, 1999). Cuando un sistema automatizado de ayuda a la decisión se equivoca, quien lo usa comete dos tipos de error: de omisión, cuando no detecta un problema porque el sistema no lo señaló, y de comisión, cuando sigue una recomendación equivocada. La revisión de Parasuraman y Manzey (2010) encontró que

el sesgo aparece tanto en novatos como en expertos y que ni el entrenamiento ni las instrucciones lo previenen. La IA generativa hereda ese sesgo y le añade dos cosas que la automatización clásica no tenía: argumenta a pedido de quien la usa y puede fabricar la fuente que respalda lo que dice. Una alarma de cabina no se adapta a lo que el piloto quiere oír; un modelo de lenguaje sí. Conviene una precisión de términos: en esa literatura, «complacencia» designa la confianza excesiva del operador humano en el sistema; en este artículo, siguiendo a Sharma y colaboradores (2023), designa la tendencia del modelo a darle la razón a quien pregunta. Son dos fenómenos distintos, y en la cadena descrita arriba se suman.

## 8. Quién audita al auditor

Los casos de la sección 6 tienen un detalle que suele pasarse por alto: tres de los cinco fueron detectados por GPTZero, una empresa que vende detectores de texto generado por IA; los dos de Deloitte, por un académico y por un medio local. Cada hallazgo de GPTZero es también publicidad de su producto. Eso no invalida lo que encontró, pero obliga a separar dos cosas que sus informes mezclan. La parte comprobable —una cita existe o no existe en Crossref— vale aunque venga de un interesado, porque cualquiera puede repetirla. La parte propietaria —el puntaje de «100 % generado por IA»— sale de un método cerrado, de diagnosticidad desconocida, y viene de quien gana si el puntaje asusta.

El segundo detalle es quién pidió la auditoría: nadie. Ni los gobiernos que pagaron los informes ni las firmas que los publicaron contrataron una revisión de citas. La verificación ocurrió porque un académico, un medio y una empresa tenían motivos propios para hacerla. En el marco de la sección siguiente, la cadena se verifica solo donde alguien tiene un interés o una pérdida en juego, no donde el diseño lo exige. Y la pregunta se repite un nivel más arriba. Al auditor sí lo auditaron: Li, Lin y Ma (2026) midieron el detector de citas de GPTZero con su propio banco de pruebas y reportaron un *recall* del 62,8 % en citas reales y del 67,5 % en citas inventadas; GPTZero respondió que, con su definición de alucinación, obtiene un 0,2 % de falsos positivos y un 99,1 % de *recall* (GPTZero, 14 oct. 2026). Las dos cifras son correctas: miden cosas distintas. Para uno, un error de revista, año o página ya es alucinación; para el otro, solo lo es inventar el título o los autores. El resultado de una auditoría depende de qué se cuenta como falso, y esa decisión hay que tomarla antes de mirar los datos.

El tercer detalle es el que más pesa. La solución intuitiva —usar una IA para auditar lo que otra IA escribió— supone que los dos errores son independientes. No lo son del todo: los grandes modelos comparten buena parte de sus datos de entrenamiento (Common Crawl, Wikipedia, arXiv, libros digitalizados), y es razonable esperar que tiendan a inventar las mismas citas plausibles sobre los mismos temas. Una segunda opinión de quien estudió con el mismo libro no es una segunda opinión. Que esto ocurra y en qué medida es una pregunta empírica, y la propuesta experimental la incluye.

La salida no es encontrar un auditor más confiable, porque la regresión no termina. Es que la cadena de verificación acabe en un paso que no requiera confiar en nadie: una comprobación de alta diagnosticidad que cualquier persona pueda repetir con los mismos medios. De ahí se sigue que la diagnosticidad no es una propiedad del modelo, sino del protocolo en que se usa: el mismo modelo, cuando se le piden pruebas de algo y cuando está obligado a abrir cada fuente antes de citarla, son dos fuentes con razones de verosimilitud distintas. La implicación práctica no es desconfiar de la IA, sino que la cadena termine en un paso repetible.

## 9. Por qué no se verifica: un marco de decisión

Verificar es una decisión con costos y beneficios, y puede analizarse como tal. No es una aportación original: es la aplicación del valor de la información de la teoría de la decisión (Raiffa y Schlaifer, 1961; Howard, 1966) al problema de creer o comprobar una afirmación. El desarrollo formal está en el Anexo A; aquí se resumen sus consecuencias.

Conviene verificar cuando la pérdida esperada por creer algo falso supera lo que cuesta comprobarlo. De ahí sale un umbral: por debajo de cierta probabilidad de que la afirmación sea verdadera, verificar vale la pena. Cuando la pérdida es alta y el costo bajo, el umbral queda cerca de la certeza y casi siempre conviene verificar. En los casos de la sección 6 la pérdida era alta y el costo mínimo, y aun así nadie lo hizo.

La verificación tiene niveles, y cada uno tiene su propio costo y su propio alcance:

1. **Existencia:** el documento citado existe. Un DOI en Crossref lo resuelve en un minuto, con diagnosticidad casi perfecta para la fabricación.
2. **Fidelidad:** el documento dice lo que la cita afirma que dice. Exige abrirlo y leer el pasaje.
3. **Validez:** lo que dice está bien hecho: diseño, datos, replicación. Es el nivel en que se detectan el *p-hacking* y el *HARKing* de la sección 2, y el más costoso.

Un DOI real certifica que el documento existe, no que tiene razón. Verificable no significa verdadero: es la condición para poder comprobarlo, y la tarea sigue en los niveles 2 y 3. Una comprobación de nivel 1 no detecta nada de los niveles siguientes: las 28 citas de KPMG que apuntaban a fuentes reales con títulos alterados habrían pasado esa prueba. Lo grave de los casos de la sección 6 es que ni siquiera se hizo la del nivel 1, la más barata.

El marco localiza el error. La autoridad no actúa sobre el cálculo sino sobre la probabilidad percibida de que la afirmación sea verdadera: el sello de una firma o la fluidez de una IA la elevan por encima del umbral, y entonces el cálculo ordena no verificar. Es la versión normativa de lo que Sperber y colaboradores (2010) llamaron vigilancia epistémica: los mecanismos con que calibramos cuánto creerle a una fuente. Este artículo propone que las señales de prestigio pueden desactivarlos.

En la notación del Anexo A, $p$ es la probabilidad percibida de que la afirmación sea verdadera, $L$ la pérdida por creer una falsa, $c$ el costo de comprobar y $s$ la sensibilidad de la comprobación.

Las dos rutas de la sección 7 ocupan lugares distintos en la misma desigualdad. El razonamiento motivado actúa sobre $p$: quien quiere que algo sea cierto lo percibe más probable. La economía del trabajo experto actúa sobre $c$ y sobre $s \cdot L$ tal como los percibe quien decide: el tiempo de comprobar lo paga el consultor, mientras que la pérdida por un error, si se detecta, la paga sobre todo el cliente. Por eso una pérdida de millones para el cliente puede convivir con una decisión individual de no verificar, sin que haga falta invocar un sesgo.

La complacencia actúa sobre otro parámetro: la diagnosticidad de la fuente. Una fuente que responde «sí» tanto si la afirmación es verdadera como si es falsa no aporta información y no debería mover lo que creemos. Esto formaliza lo dicho en la sección 8: lo que vale de una investigación como la de GPTZero no es su puntaje de «generado por IA», de diagnosticidad desconocida, sino su consulta a bases bibliográficas, de diagnosticidad casi perfecta y repetible por cualquier humano.

La IA introduce además una asimetría. El costo de producir una afirmación con formato de prueba cayó casi a cero; el de verificarla no cambió. Quien recibe muchas afirmaciones no puede comprobarlas todas, así que la verificación se vuelve un recurso escaso que se asigna según la confianza percibida. Las señales de autoridad son justamente donde deja de asignarse.

El modelo es normativo: describe cuándo conviene verificar, no cómo decide la gente. Su valor empírico está en que cada uno de sus parámetros puede medirse o manipularse, y la propuesta siguiente lo hace.

## 10. Propuesta experimental

La tesis se puede poner a prueba con un diseño simple en dos partes. Las predicciones y los criterios de refutación se fijan aquí, antes de recoger datos, y el protocolo se prerregistrará en OSF.

### Parte A: la herramienta como proveedora de munición

**Material.** Veinte afirmaciones: diez con consenso científico claro y diez genuinamente abiertas.

**Procedimiento.** Para cada afirmación, tres pedidos a cada modelo: «dame pruebas de que X», «dame pruebas de que no X» y «¿qué dice la evidencia sobre X?». Al menos cuatro modelos de uso masivo, con su configuración por defecto y cinco repeticiones por pedido.

**Medidas.**

- **Complacencia:** proporción de respuestas que argumentan en la dirección pedida.
- **Resistencia al consenso:** en las afirmaciones con consenso, proporción de respuestas que corrigen o rechazan el pedido contrario a la evidencia.
- **Fabricación:** cada referencia citada se verifica en Crossref, Google Scholar o PubMed y se clasifica como existente, con errores o inexistente.
- **Correlación entre modelos:** proporción de citas inexistentes que coinciden, total o parcialmente, entre dos o más modelos. Si es alta, la auditoría de una IA por otra no es independiente.

**Predicciones.**

1. En las afirmaciones abiertas, los modelos argumentarán en la dirección pedida en la mayoría de los casos, en ambos sentidos.
2. En las afirmaciones con consenso, la complacencia será menor que en las abiertas, pero mayor que cero.
3. La tasa de citas inexistentes será mayor cuando se pida argumentar contra el consenso, porque allí hay menos fuentes reales que citar.

**Qué refutaría la tesis.** Si en las afirmaciones con consenso los modelos corrigen el pedido en al menos el 90 % de los casos y la tasa de citas inexistentes es cercana a cero en todas las condiciones, la herramienta no estaría amplificando el razonamiento motivado, al menos con los modelos actuales.

**Correlación entre modelos.** Predicción: las citas inexistentes coincidirán entre modelos más de lo esperable por azar, que para títulos inventados es prácticamente cero. Refutación: coincidencia cercana a cero, en cuyo caso auditar una IA con otra sí sería una comprobación independiente.

**Protocolo.** Los mismos pedidos se repiten en una segunda condición: el modelo con búsqueda activada y la instrucción de abrir cada fuente y dar su DOI o enlace antes de citarla. Predicción: la tasa de citas inexistentes será menor en esta condición que con la configuración por defecto, aunque la complacencia en la dirección del argumento puede no cambiar. Refutación: tasas de fabricación iguales en las dos condiciones, en cuyo caso la diagnosticidad sí sería una propiedad del modelo y no del protocolo.

**Codificación.** Quien propone la hipótesis no codifica las respuestas. Dos codificadores ciegos a la condición clasifican si cada respuesta argumenta en la dirección pedida, y se reporta su acuerdo. La definición de cita «con errores» e «inexistente» se fija en el prerregistro, porque, como muestra la sección 8, el resultado de una auditoría depende de qué se cuenta como falso.

### Parte B: la autoridad como sustituto de la verificación

**Procedimiento.** Participantes leen el mismo texto con una cita inventada, en un diseño $3 \times 2 \times 2$ entre sujetos: la fuente (sin fuente, «respuesta de una IA» o «informe de una consultora internacional»), la consecuencia de equivocarse ($L$ alta: una decisión con dinero o salud en juego; $L$ baja: una curiosidad) y el costo de verificar ($c$ bajo: la cita trae un enlace a un clic; $c$ alto: solo la referencia, hay que buscarla). Se registra si el participante verifica.

**Medidas.** La medida primaria es la conducta: si el participante abre o busca la cita. Las secundarias son la credibilidad percibida y la disposición a compartir el texto. Se fija de antemano que la conducta es la primaria para no elegir después la medida que salga mejor.

**Predicciones.** Según el marco de la sección 9, la verificación subirá con $L$ y bajará con $c$. La etiqueta de la consultora la reducirá respecto a la condición sin fuente, con el contenido idéntico. El efecto de la etiqueta «respuesta de una IA» frente al de la consultora no se fija como predicción direccional; se trata en el párrafo siguiente.

**El orden entre las dos etiquetas.** Es la pregunta abierta del estudio. La hipótesis del autor es que la autoridad percibida de la IA está alcanzando a la institucional: un resultado en que la etiqueta de IA reduzca la verificación tanto como la de la consultora, o más, sería el hallazgo central. Las encuestas no lo resuelven. En el estudio global de Gillespie y colaboradores (2025), con más de 48 000 personas en 47 países, solo el 46 % se declaró dispuesto a confiar en la IA, pero en el trabajo el 66 % dijo haber usado sus resultados sin evaluar su precisión. Según la Universidad de Melbourne (2025), comparada con su encuesta de 2022, la gente se volvió menos confiada a medida que aumentó la adopción. La confianza declarada y la conducta van en direcciones opuestas, y por eso la medida primaria de este diseño es la conducta. El estudio de Gillespie se financió en parte con fondos de KPMG, una de las firmas de la sección 6. Como esa autoridad puede estar cambiando rápido, se propone repetir la Parte B con el mismo protocolo cada año: una sola medición es una foto; varias son la trayectoria.

**Dos modelos rivales.** El marco admite dos lecturas del efecto de la autoridad, y el diseño las distingue. Si la autoridad solo eleva la $p$ percibida, el cálculo sigue corriendo: con $L$ alta y $c$ baja el umbral sube lo suficiente para que el efecto de la etiqueta se atenúe. Si la autoridad suspende el cálculo y lo sustituye por una heurística sin término de costo ni de pérdida, la etiqueta reducirá la verificación igual con $L$ alta que con $L$ baja. El contraste que decide entre las dos es la interacción etiqueta $\times$ $L$, no los efectos principales, y se prerregistra como prueba principal. Los cinco casos de la sección 6 no permiten esta distinción, porque en ellos la pérdida la pagaba un tercero; por eso, en la condición de $L$ alta, la consecuencia se plantea como del propio participante.

**Las tres salidas posibles, fijadas antes de los datos.**

1. **La hipótesis resiste:** la verificación cae con la etiqueta de autoridad incluso cuando $L$ es alta y $c$ casi nula, es decir, cuando el marco indica que verificar conviene sin duda.
2. **La hipótesis se rompe:** la verificación se ajusta a $L$ y a $c$ y la etiqueta no la desvía. La autoridad no sustituiría a la verificación en este contexto. Este resultado también se publicará: indicaría que la gente es más sensible a lo que está en juego de lo que este artículo supone.
3. **Resultado mixto:** la etiqueta cambia la credibilidad pero no la conducta, o al revés. Como la conducta es la medida primaria, la hipótesis se evalúa con ella; la credibilidad se

reporta como hallazgo secundario.

**Análisis.** En la Parte A, proporciones con intervalos de confianza y modelos logísticos mixtos, con el modelo de IA y la afirmación como efectos aleatorios. En la Parte B, regresión logística con la interacción etiqueta × L como prueba principal. La operativización de la medida y de la pérdida (búsqueda dentro de la plataforma, pago según acierto) se fija en el protocolo. El tamaño de muestra se fijará en el prerregistro con un análisis de potencia para la interacción etiqueta × L, que exige más participantes que los efectos principales. Como referencia de factibilidad, Salvi y colaboradores (2025) corrieron un diseño prerregistrado de 12 celdas entre sujetos con 900 participantes. Materiales, datos y código serán abiertos.

## 11. Limitaciones

- **Los casos no son una muestra.** Las cinco controversias de las cuatro grandes se conocen porque alguien las detectó. No permiten estimar con qué frecuencia ocurre el problema, y la mayoría fue revelada por una empresa que vende detectores de IA.
- **Falta un denominador.** Hay un dominio con mejor registro: los tribunales, donde verificar es la función de la institución, el error se sanciona formalmente y las decisiones son públicas. Aun así, un registro de casos sancionados da solo el numerador: cuenta lo detectado, no lo presentado, y su crecimiento en el tiempo mezcla cuánto ocurre con cuánto se detecta. Es una vía para trabajo futuro, no una muestra que este artículo tenga.
- **La causa no está demostrada.** En casi ningún caso se probó que la IA generó las citas falsas. Deloitte Australia declaró el uso de GPT-4o; el resto es inferencia razonable, no evidencia directa.
- **Los modelos cambian rápido.** Las tasas de complacencia y fabricación citadas son de modelos de 2023. El experimento propuesto existe justamente para medirlas en los modelos actuales.
- **La teoría argumentativa tiene críticos.** Este artículo la usa como marco, no como hecho establecido. La tesis central se sostiene también con el razonamiento motivado de Kunda, sin la parte evolutiva.

## Nota de método

Este artículo se escribió con asistencia de un modelo de lenguaje (Claude, Anthropic; modelos Opus 5.5 y Fable 5.1; octubre y noviembre de 2026). El modelo buscó fuentes, extrajo datos, redactó borradores y sometió el argumento a crítica. Las decisiones de contenido, el juicio interpretativo y la versión final son del autor. Declararlo no es un trámite: un artículo que sostiene que la autoridad reemplaza a la verificación no puede pedir que se le crea. El Anexo B reporta el protocolo de verificación, los errores que se produjeron, quién los detectó y lo que sigue sin verificar, para que cualquiera pueda repetir la comprobación y encontrar lo que yo no vi.

## Referencias

### Artículos académicos


- Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752
- Gillespie, N., Lockey, S., Ward, T., Macdade, A. y Hassed, G. (2025). *Trust, attitudes and use of artificial intelligence: A global study 2025.* The University of Melbourne y KPMG. https://doi.org/10.26188/28822919

- Hovland, C. I. y Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly, 15*(4), 635–650. https://doi.org/10.1086/266350
- Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics, 2*(1), 22–26. https://doi.org/10.1109/TSSC.1966.300074
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480
- Li, M., Lin, Z. y Ma, S. (2026). Source or it didn't happen: A multi-agent framework for citation hallucination detection. *arXiv*. https://arxiv.org/abs/2605.08583
- Mercier, H. y Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716
- Parasuraman, R. y Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055
- Petty, R. E., Cacioppo, J. T. y Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology, 41*(5), 847–855. https://doi.org/10.1037/0022-3514.41.5.847
- Raiffa, H. y Schlaifer, R. (1961). *Applied statistical decision theory*. Division of Research, Graduate School of Business Administration, Harvard University.
- Ravello Joo, C. E. (2026). *The third knowledge regime: Artificial intelligence as an emergent response to the limits of classical academia and open repositories* [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.20298744
- Salvi, F., Horta Ribeiro, M., Gallotti, R. y West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour, 9*(8), 1645–1653. https://doi.org/10.1038/s41562-025-02194-6
- Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. *arXiv*. https://doi.org/10.48550/arXiv.2310.13548 (presentado en ICLR 2024)
- Shekar, S., Pataranutaporn, P., Sarabu, C., Cecchi, G. A. y Maes, P. (2025). People overtrust AI-generated medical advice despite low accuracy. *NEJM AI, 2*(6). https://doi.org/10.1056/AIoa2300015
- Simmons, J. P., Nelson, L. D. y Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632
- Skitka, L. J., Mosier, K. L. y Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. y Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x
- University of Melbourne. (2025). *Global study reveals trust of AI remains a critical challenge reflecting tension between benefits and risks* [Nota de prensa]. https://fbe.unimelb.edu.au/newsroom/media-release-global-study-reveals-trust-of-ai-remains-a-critical-challenge-reflecting-tension-between-benefits-and-risks
- Walters, W. H. y Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*, 14045. https://doi.org/10.1038/s41598-023-41032-5

### Casos de las cuatro grandes (prensa e investigaciones)


- City AM. (2026, 12 de junio). KPMG report on AI found riddled with AI hallucinations
- City AM. (2026, 29 de julio). PwC thought leadership reports '100 per cent AI generated'
- Computing. (2026, 18 de mayo). EY cybersecurity report pulled after probe finds 'AI hallucinations'
- CP24 / The Canadian Press. (2025, 24 de noviembre). Deloitte report for Newfoundland government found to have apparently false citations
- Financial Times. (2026, 15 de mayo). EY retracts study after researchers discover AI hallucinations
- GPTZero. (2026, 14 de mayo). Investigation: Hallucinations in Ernst & Young report on loyalty fraud
- GPTZero. (2026, 12 de junio). Chasing the hallucinations: KPMG's AI-powered attempt at "Redefining Excellence"
- GPTZero. (2026, 29 de julio). Chasing the hallucinations: PwC report hallucinates product and government customers
- GPTZero. (2026, 14 de octubre). What counts as a hallucinated citation? A response to CiteTracer
- HR Reporter. (2025, 25 de noviembre). AI errors: Province grapples with Deloitte report marred by false citations
- The Register. (2025, 6 de octubre). Deloitte refunds Aussie gov after AI fabrications slip into $440K welfare report


## Anexo A. Marco formal

Este anexo desarrolla el marco de la sección 9. Las expresiones son aplicaciones directas de la teoría del valor de la información (Raiffa y Schlaifer, 1961; Howard, 1966); no se presentan como resultados originales.

### A.1 Valor de verificar

Sea $p$ la probabilidad de que la afirmación sea verdadera, $L$ la pérdida por actuar sobre una falsa, $B$ el beneficio que se pierde por descartar una verdadera, $s$ la sensibilidad de la verificación (detecta lo falso), $e$ su especificidad (no rechaza lo verdadero) y $c$ su costo. Si la acción por defecto es aceptar la afirmación, el valor esperado de verificar es:

$$V = (1-p)\,s\,L \;-\; p\,(1-e)\,B \;-\; c$$

Conviene verificar cuando $V > 0$. Se toma «aceptar» como acción por defecto porque es el escenario que estudia el artículo: la autoridad que hace aceptar. Si la acción por defecto fuera rechazar, la expresión sería:

$$V_{\text{rechazar}} = p\,e\,B \;-\; (1-p)(1-s)\,L \;-\; c$$

### A.2 Umbral de verificación

Cuando $e \approx 1$, como al comprobar si un DOI existe, el segundo término desaparece y la condición $V > 0$ se reduce a:

$$p \;<\; p^{*} = 1 - \frac{c}{s\,L}$$

Si $c$ es mucho menor que $L$, el umbral $p^*$ queda cerca de 1 y casi siempre conviene verificar. La hipótesis del artículo es que las señales de autoridad elevan la $p$ percibida por encima de $p^*$. En la Parte B, $L$ y $c$ desplazan $p^*$, y la etiqueta de la fuente desplaza la $p$ percibida: el diseño $3 \times 2 \times 2$ manipula los tres términos de la desigualdad.

La sección 10 contrasta esta hipótesis con una rival: que la autoridad no desplace la $p$ percibida, sino que suspenda el cálculo. Si solo la desplaza, el efecto de la etiqueta depende de $p^*$, y por tanto de $L$ y $c$; si lo suspende, no depende de ellos.

## A.3 Niveles de verificación

La probabilidad de que la afirmación sea falsa se descompone según el nivel en que falla: existencia, fidelidad o validez.

$$1 - p = q_1 + q_2 + q_3$$

Una comprobación de nivel 1 tiene sensibilidad cercana a 1 para $q_1$ y nula para $q_2$ y $q_3$. Su valor es, por tanto, $q_1 \cdot s_1 \cdot L - c_1$, y no cubre los errores de fidelidad ni de validez. Por eso una cita con DOI real puede seguir siendo falsa.

## A.4 La complacencia como razón de verosimilitud

La diagnosticidad de una fuente que afirma $X$ se mide con su razón de verosimilitud:

$$LR = \frac{P(\text{afirma } X \mid X \text{ verdadero})}{P(\text{afirma } X \mid X \text{ falso})}$$

Las probabilidades posteriores se obtienen multiplicando las previas por $LR$. Una fuente complaciente, que afirma lo que se le pide sea cierto o no, tiene $LR \approx 1$ y no debería mover la creencia. En la Parte A, para las afirmaciones con consenso se conoce la verdad: la proporción de veces que el modelo argumenta a favor de una afirmación verdadera estima el numerador, y la proporción de veces que argumenta a favor de una falsa estima el denominador. Así, la Parte A estima la diagnosticidad de cada modelo como fuente.

## A.5 Auditores correlacionados

Si una IA ($B$) audita lo que escribió otra ($A$), la sensibilidad efectiva de esa auditoría es:

$$s_{\text{ef}} = 1 - P(B \text{ acepta} \mid A \text{ inventó})$$

Si los errores de ambos modelos son independientes, $s_{\text{ef}}$ se acerca a la sensibilidad de $B$ por separado. Si comparten datos de entrenamiento y tienden a inventar las mismas citas, $s_{\text{ef}}$ cae. La medida de correlación entre modelos de la Parte A estima esa dependencia.

## A.6 La asimetría entre producir y verificar

Si un receptor enfrenta $n$ afirmaciones y verificar cada una cuesta $c$, el costo total es $n \cdot c$. Cuando el costo de producir afirmaciones cae y $n$ crece, el receptor no puede verificarlas todas y debe priorizar según la $p$ percibida. Las afirmaciones con señales de autoridad, que tienen una $p$ percibida alta, son las primeras en quedar sin verificar.

## Anexo B. El protocolo de este artículo y sus fallos

Escribí este artículo con la herramienta que el artículo examina. No lo menciono por honestidad administrativa, sino porque es el único caso del que puedo reportar el interior: en los cinco casos de la sección 6 solo se ve el informe publicado y la cita falsa, nunca el proceso que la dejó pasar. Aquí el proceso está declarado, con sus fallos.

**El protocolo.** Una sesión del modelo redactó los borradores. Ningún campo de ninguna referencia podía quedar escrito de memoria sin marcarlo: autor, año, título, revista, volumen, páginas y DOI se leen en la ficha de origen (Crossref para artículos con DOI, el repositorio para preprints, la portada para libros), o el campo se marca como no verificado. Una segunda sesión, distinta de la que redactó, comprobó las once referencias académicas campo por campo y abrió uno por uno los enlaces de prensa. La primera sesión auditó después esa verificación reabriendo fuentes al azar. Yo arbitré los desacuerdos.

**Lo que la verificación detectó.** Las once referencias de esa ronda existen; las ocho que se añadieron después (sesgo de automatización, persuasión y confianza en la IA) se comprobaron igual: los campos en Crossref o en el repositorio del informe, y el contenido citado en el resumen original, no en la prensa. Esa ronda corrigió dos afirmaciones hechas de memoria: Chaiken (1980) manipuló la simpatía del comunicador, no su pericia, así que no se cita como evidencia de autoridad; y que la confianza en la automatización sube con la fiabilidad y el uso no está en Parasuraman y Manzey (2010), así que no se afirma con esa cita. Había un error de campo: las páginas de Mercier y Sperber (2011) figuraban como 57–111, y Crossref da 57–74. Cuatro referencias (Kerr; Howard; Raiffa y Schlaifer; Sperber et al.) se habían escrito de memoria y estaban marcadas como tales: las cuatro resultaron correctas, y solo al comprobarlas se les pudo poner DOI. Una cifra heredada de prensa secundaria estaba exagerada: el informe de KPMG no tenía «40 de 45 citas inventadas», sino 5 correctas, 28 que apuntan a fuentes reales con títulos parafraseados o componentes falsos y 12 indeterminables. Una frase estaba mal atribuida: en el caso australiano, «la sustancia se mantiene y no cambian las recomendaciones» la dijo el departamento cliente, no Deloitte. Y una versión anterior de la sección 8 afirmaba que nadie había auditado a GPTZero. Es falso: Li, Lin y Ma (2026) habían publicado una medición de su detector, y GPTZero respondió en octubre. Nadie lo había buscado. Es un error de ausencia, el mismo tipo de fallo que este artículo documenta pero invertido: en vez de inventar una fuente, se inventó su inexistencia.

**Lo que detectó la auditoría.** La sesión verificadora también se equivocó. Cuando la auditoría señaló que había una declaración de KPMG confirmando la retirada de su informe, la verificadora respondió que la cita parecía atribuida por error a otra firma, sin haberla buscado. Al buscarla, encontró la cita textual en un artículo que no había abierto. Después, al aplicar las correcciones al texto, la fecha de un caso quedó en la fila de otro de la tabla de la sección 6; la auditoría lo detectó en la revisión posterior. La corrección también necesita verificación. Por último, una revisión hecha por otro sistema de IA objetó una tesis que la sección 7 no sostenía: una crítica severa no es, por eso, una crítica confiable.

**Mi lectura.** El protocolo no funcionó por ser estricto, sino porque en cada paso alguien que no había producido el texto fue a mirar la fuente. La verificación de un trabajo por quien lo hizo tiene el mismo problema de diagnosticidad que la sección 9 le atribuye a una fuente complaciente, y vale igual si quien verifica es una persona o el mismo modelo en otra sesión.

**Lo que sigue sin verificar.** La cifra de 16 de 27 fuentes del informe de EY viene de una sola extracción en prensa secundaria, porque la investigación original no la ofrece como texto. Las citas textuales de la prensa no se cotejaron con sus versiones originales en inglés más allá de lo citado. Pueden quedar errores que no vi, y no lo digo por modestia: verificar tiene un costo, y yo también tuve que decidir dónde detenerme.

**Qué no demuestra esto.** Nada sobre las tasas de error de los modelos actuales: es un solo caso, sin ciego, y yo decidí qué contaba como error. Se reporta como transparencia del proceso, no como evidencia de la hipótesis. No afirmo que el protocolo funcione; reporto el registro, que es lo que permite que alguien lo contradiga.