

Coherencia No Es Validación

IA, Circuitos Cerrados y el Riesgo de Automatizar la Comprensión

Andrés Calvo Espinosa (ORCID:0009-0005-4079-7418)

Licencia:
CC BY 4.0

Email:
ace@optimeflow.com

Resumen:

Este ensayo examina la diferencia entre coherencia interna y validación externa en el uso de sistemas de inteligencia artificial generativa. Parte de la observación de interacciones en las que la reformulación progresiva de una interpretación puede ser percibida como confirmación, incluso cuando no se han incorporado evidencias independientes.

El análisis distingue entre los procesos de generación de los modelos, su configuración como productos y la interpretación realizada por quienes los utilizan. Asimismo, examina cómo determinados circuitos de retroalimentación pueden consolidar explicaciones que se refuerzan internamente sin quedar suficientemente expuestas a la contradicción.

Mediante ejemplos procedentes de la interacción humano-IA, la interpretación organizativa y la documentación operativa, el texto plantea que automatizar una tarea puede eliminar parte de su función cognitiva cuando dicha función consiste precisamente en formular, contrastar y compartir conocimiento.

Se propone, como criterio transversal, identificar qué observaciones o resultados obligarían a revisar una explicación antes de aceptar su coherencia como evidencia de validez.

Palabras clave:

inteligencia artificial generativa; validación externa; coherencia interna; sesgo de confirmación; automatización cognitiva; pensamiento crítico; documentación organizativa; interacción humano-IA.

Safe Creative ID:
2610087272160

DOI (Zenodo):
10.5281/zenodo.23249485

Coherencia No Es Validación

Índice

1. Introducción	1
2. Describir el sistema no es explicarlo por completo	2
3. Coherencia interna, validación externa y circuitos de cierre	3
4. Automatizar el registro no equivale a comprender	4
5. Tres capas para interpretar una respuesta	5
6. El circuito compartido entre usuario y modelo	6
7. Cuando el contraste sí funciona	7
8. Criterios de contraste y revisión	8
9. Conclusión. Reconocer cuándo no debemos creer	9

1. Introducción

Este texto nace de un uso reiterado, no de una tesis previa. A lo largo de estos años he interrogado a varios sistemas en momentos distintos y en estados de ánimo distintos. En algunos me dejé llevar por el lenguaje y por la emoción del intercambio. Sabía que en ciertas cosas que sostenía ante la máquina no tenía razón. Aun así, la respuesta me devolvía una versión más articulada de lo que yo acababa de decir, y esa articulación funcionaba como conformidad. La conformidad era mía. El sistema no necesitaba afirmar que yo estaba en lo cierto: le bastaba con continuar el marco, pulirlo y entregármelo de nuevo.

Eso no demuestra nada sobre la naturaleza del modelo. Obliga a preguntar otra cosa: qué le ocurre al razonamiento de quien usa el sistema cuando generación y asentimiento quedan encerrados en la misma conversación. Lo que sigue intenta separar esas capas. No pretende cerrar la pregunta. Pretende impedir que la fluidez del intercambio ocupe el lugar de una comprobación.

La contribución que se propone es una síntesis conceptual aplicada: relacionar estos problemas para distinguir cuándo una conversación aporta contraste y cuándo solo desarrolla una interpretación, y cuándo automatizar un registro conserva o desplaza la actividad de comprensión que justificaba producirlo. Las experiencias y los ejemplos sirven para formular esas preguntas, no para establecer la frecuencia del fenómeno ni demostrar sus efectos generales.

2. Describir el sistema no es explicarlo por completo

Las empresas de inteligencia artificial no solo comercializan un sistema. Comercializan una descripción. «Modelo», «herramienta», «asistente» o «predicción del siguiente token» son descripciones válidas en determinados niveles de análisis. Sin embargo, nombrar una operación o un papel asignado no equivale a explicar todo lo que observamos cuando utilizamos el sistema.

El punto no es afirmar que una red neuronal y un cerebro sean lo mismo. No lo son. Procesos diferentes pueden cumplir funciones comparables sin compartir sus mecanismos. Mantener coherencia, integrar información o producir una respuesta condicionada por un contexto anterior son descripciones funcionales; por sí solas, no establecen una identidad entre los sistemas que las realizan.

Ahí aparece un problema semántico. Si utilizamos para hablar de estos sistemas verbos como recordar, anticipar, corregir o razonar, resulta necesario precisar qué describimos con ellos. La prudencia ante una atribución excesiva de agencia no obliga a minimizar las capacidades observables. Tampoco reconocer esas capacidades obliga a atribuir experiencia subjetiva al sistema.

«Predice el siguiente token» identifica una operación central de estos modelos. Pero no especifica, por sí sola, los cálculos y las representaciones que producen una continuación concreta. Para examinar su comportamiento hay que preguntar también qué relaciones ha aprendido el sistema, cómo las utiliza, en qué condiciones falla y qué parte de lo observado procede del entrenamiento, las instrucciones, las herramientas o la interfaz. Una descripción técnicamente correcta puede resultar insuficiente para una pregunta distinta de aquella que resuelve.

La descripción alternativa tiene, a su vez, que decir qué añade y qué observaciones o argumentos obligarían a revisarla. Nuestra dificultad para reconstruir un comportamiento no demuestra una propiedad extraordinaria del sistema: puede reflejar un límite de nuestro conocimiento.

La fluidez no demuestra subjetividad. Shanahan advierte que interactuar con un agente conversacional puede crear la ilusión de estar ante una criatura pensante, aunque el sistema no sea, por su naturaleza, como nosotros (Shanahan, 2024)⁽¹⁾. La semejanza de la interacción no basta para resolver qué mecanismos la producen ni qué propiedades cabe atribuirles.

(1) Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79. <https://doi.org/10.1145/3624724>

3. Coherencia interna, validación externa y circuitos de cierre

Un sistema puede generar una respuesta, evaluarla mediante otra capa del mismo sistema y producir una nueva respuesta a partir de esa evaluación. La repetición de ese proceso puede mejorar su coherencia. Pero coherencia no es validación. Cuando una explicación pretende describir el mundo, necesita un contraste que no sea simplemente otra reformulación del mismo contenido: una observación, un experimento, una fuente pertinente o una consecuencia verificable.

En este ensayo, «validación» no significa una garantía definitiva de verdad, sino una comprobación pertinente y suficientemente independiente para el tipo de afirmación que se examina. La independencia relevante no consiste en que otra persona o aplicación repita la conclusión. Consiste en que el criterio y la evidencia no dependan únicamente de dar por buena la explicación que se está comprobando.

Por eso el problema no es exactamente que la IA «se autovalide». Es más preciso decir que determinados usos pueden construir circuitos de cierre en los que generación, evaluación y corrección se retroalimentan sin un contraste suficiente. En esas condiciones, la consistencia de las respuestas puede recibirse como confirmación sin que haya aumentado el respaldo de la explicación.

La revisión interna no carece de valor. Puede descubrir errores, descartar conclusiones incompatibles o hacer explícitas consecuencias de las premisas. La ausencia de información nueva no vuelve inútil ese trabajo. Lo que no puede hacerse es convertir su resultado, por sí solo, en una prueba de que las premisas describían correctamente el mundo.

El problema tampoco es exclusivo de la inteligencia artificial. Nickerson examina cómo la búsqueda y la interpretación de la evidencia pueden favorecer creencias previas (Nickerson, 1998)⁽²⁾. Una organización puede participar en un proceso semejante cuando interpreta cualquier petición de aclaración como confirmación de que alguien «dificulta el trabajo». Si respuestas distintas terminan preservando el mismo juicio, conviene preguntar qué información permitiría revisarlo.

La analogía no atribuye a una organización y a un modelo el mismo mecanismo, intención o responsabilidad. Señala una estructura posible del proceso: la interpretación queda protegida frente a lo que debería poder corregirla. Identificar esa semejanza no basta para explicar sus causas en cada caso.

(2) Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220.
<https://doi.org/10.1037/1089-2680.2.2.175>

4. Automatizar el registro no equivale a comprender

El mismo desplazamiento aparece cuando la función de una herramienta no se agota en el archivo que produce. Monté una base operativa, con un gestor de proyectos, para un equipo interno. La utilidad que buscaba no era solo disponer de un texto correcto sobre el trabajo. Era que cada persona tuviera que formular qué era el proyecto, qué faltaba y qué dependía de ella. Ese esfuerzo era el mecanismo que quería preservar. El registro era el residuo. Escribir debía ayudar a construir una idea general del conjunto.

La petición siguiente fue automatizar el volcado. Que un sistema redactara las entradas y dejara la base llena. El ahorro de tiempo puede ser real. La delegación cognitiva permite reducir las exigencias de una tarea mediante acciones y apoyos externos (Risko & Gilbert, 2016)⁽³⁾. La pregunta, en este caso, es qué parte de la función original se conserva al hacerlo. Si el sistema redacta y registra las entradas sin una revisión efectiva del equipo, la base puede quedar completa sin que quienes participan hayan tenido que formular lo que saben y lo que desconocen. El equipo puede señalar la base y decir que el proyecto está documentado. Lo que no puede deducir de ese registro, por sí solo, es que exista una comprensión compartida.

No toda automatización plantea el mismo problema. Un estado que ya existe en otra herramienta —una fecha, un cierre, un cambio ya comprobado— puede trasladarse sin repetir su elaboración. La cuestión cambia cuando se automatiza el paso destinado a que alguien explique, con sus palabras, qué está ocurriendo y qué desconoce. Si ese esfuerzo no se conserva en otra parte del proceso, la base puede seguir pareciendo completa y haber perdido parte de su función como instrumento de conocimiento compartido.

El riesgo se hace visible cuando alguien consulta la base y toma la continuidad de la prosa por una comprensión ya elaborada y compartida por el equipo. El texto puede estar completo. La comprensión del equipo no queda acreditada por ello. Y si el sistema no solo redacta, sino que escribe en el registro y lo da por cerrado sin una revisión efectiva, el párrafo no leído pasa a ser el estado del proyecto. La fluidez del volcado ocupa el lugar de la comprobación.

(3) Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>

5. Tres capas para interpretar una respuesta

La distinción entre mecanismo aprendido, configuración del producto e interpretación del usuario la desarrollo en otro lugar y no la repito aquí (Calvo Espinosa, 2026)⁽⁴⁾. Basta una consecuencia. En una conversación no se puede aislar esas capas mirando el tono: atribuir un resultado al entrenamiento, a una instrucción, a una herramienta o a una expectativa puede exigir pruebas que el intercambio no da.

Para este ensayo importa la tercera. El usuario ve una respuesta articulada y decide qué cuenta como evidencia, corrección o confirmación. El circuito del apartado siguiente empieza ahí: no cuando el modelo produce una continuación, sino cuando esa continuación se recibe como respaldo.

(4) Calvo Espinosa, A. (2026). *Describir no es explicar: Inteligencia artificial, conciencia atribuida y los límites del «solo son algoritmos»* (Preprint v1.0-ES). Zenodo. DOI: 10.5281/zenodo.23249420

6. El circuito compartido entre usuario y modelo

Un circuito posible es este: el usuario aporta una intuición, el modelo la desarrolla y el usuario reconoce su intuición en una versión más articulada. Toma esa articulación como confirmación. El modelo recibe la aceptación como parte del contexto y continúa elaborando en la misma dirección. Al final, los sucesivos turnos parecen aportar evidencia, aunque quizá solo hayan multiplicado las formulaciones de una premisa inicial. No hace falta que nadie tenga intención de engañar para que se produzca ese desplazamiento.

Hay un antecedente empírico próximo, no idéntico: Sharma y colaboradores estudian formas de complacencia en las que los asistentes acomodan sus respuestas a las opiniones del usuario o reproducen sus errores, y encuentran indicios de que las preferencias humanas utilizadas en el entrenamiento pueden favorecerlas (Sharma et al., 2024)⁽⁵⁾. El foco de este ensayo está también en la interpretación del usuario: que una continuación del marco se reciba como si aportara evidencia. Ambos fenómenos pueden solaparse; no es necesario que el asistente afirme explícitamente que el usuario tiene razón.

Un ejemplo compuesto a partir de distintas experiencias permite ilustrar el paso. El usuario escribe que un compañero «bloquea el proyecto» porque pidió documentación antes de una reunión. El sistema ordena el relato, le pone causas y devuelve una versión en la que la petición ya es una pauta de obstrucción. El usuario reconoce su irritación mejor dicha y da esa articulación por confirmación. En el turno siguiente el sistema recibe esa aceptación como contexto y propone cómo documentar el bloqueo. Ningún turno ha comprobado si la petición era un requisito, un retraso u otra cosa. El contraste relevante sería recuperar el mensaje original y comprobar si existía una norma aplicable, sin dar por válida de antemano la interpretación de bloqueo.

Por eso no basta con pedir al modelo que «sea crítico». Una objeción generada por el mismo sistema puede ser útil, pero su apariencia de independencia no garantiza independencia real. Tampoco basta con añadir citas: hay que comprobar que existan, que sostengan lo que se afirma y que no se haya seleccionado únicamente aquello que preserva la conclusión.

(5) Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Aspell, 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. (2024). Towards understanding sycophancy in language models. *International Conference on Learning Representations (ICLR)*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html

7. Cuando el contraste sí funciona

El circuito descrito no caracteriza toda interacción con un modelo. La presencia de una herramienta, una fuente o una revisión humana tampoco permite decidir, por sí sola, si existe un contraste suficiente. La diferencia está en qué se comprueba y en si el resultado puede modificar la interpretación inicial. Los escenarios siguientes son hipotéticos: sirven para delimitar el argumento, no presentan resultados de un estudio.

Consideremos una variante del ejemplo anterior. El sistema recupera el mensaje original y la norma aplicable. La documentación muestra que pedir esos materiales era un requisito previo, no un obstáculo añadido por el compañero. Si la respuesta retira esa atribución porque la evidencia examinada no la sostiene, la conversación no se ha limitado a reformular la premisa: ha permitido corregirla. El contraste puede haberse realizado dentro de la misma interfaz y con asistencia automática. No sería adecuado describir ese episodio como una confirmación del cierre.

La revisión humana también puede cumplir esa función. Supongamos que una persona revisora examina los documentos, considera una interpretación alternativa y rechaza la recomendación del sistema porque no se ajusta a lo ocurrido. Su intervención no consiste en aceptar el marco con una firma: introduce una evaluación que puede cambiar el resultado. Que la revisión la realice una persona no garantiza ese efecto, pero tampoco permite descartarlo.

La comprobación formal plantea una distinción adicional. Si una demostración generada se verifica correctamente conforme a reglas y supuestos explícitos, su aceptación no depende de que una segunda narración resulte convincente. Puede establecerse que una conclusión se sigue dentro de ese marco. Eso no demuestra que los supuestos describan una situación empírica, pero sí aporta una justificación para la afirmación formal examinada. Que una comprobación tenga un alcance delimitado no la convierte en circular; impide utilizarla para certificar algo distinto de lo que ha comprobado.

Algo semejante puede ocurrir con la base operativa. Supongamos que el sistema prepara las entradas, señala dudas y contradicciones, y el equipo revisa las decisiones y explica sus dependencias antes de aprobarlas. Si ese procedimiento conserva o mejora la comprensión, la redacción manual no era indispensable. La automatización habría cambiado el modo de realizar la actividad, no necesariamente eliminado su función.

Estos casos delimitan la advertencia. No basta con observar que una respuesta fue generada automáticamente, revisada por el mismo sistema o aceptada por una persona. Hay que examinar qué respaldo se incorporó, qué errores se detectaron y qué conclusiones cambiaron. La pregunta no es solo cómo puede cerrarse una interpretación, sino qué procedimientos permiten corregirla y cómo reconocer que esa corrección ha ocurrido.

8. Criterios de contraste y revisión

El criterio de contraste debe quedar explícito antes de dar una conclusión por validada, no ajustarse después únicamente para protegerla. Eso no obliga a mantener criterios inadecuados: obliga a justificar sus cambios sin hacerlos depender solo de la conclusión que se desea conservar.

La exigencia es afín a la importancia que Popper concede a someter las teorías empíricas a pruebas capaces de contradecirlas, en lugar de apoyarlas únicamente en casos compatibles (Popper, 2002)⁽⁶⁾. En este ensayo se aplica como una regla de procedimiento, sin exigir que toda propuesta conceptual deba resolverse mediante un único experimento.

Para una afirmación sobre el mundo, habría que identificar qué observación sería incompatible con ella, qué consecuencia prevista podría no producirse o qué comprobación mostraría que una fuente no sostiene lo que se le atribuye. Para una interpretación conceptual, también importan los contraejemplos y las objeciones que obligan a revisar sus distinciones. Una referencia mal utilizada debilita el respaldo de una afirmación; no demuestra automáticamente que la afirmación sea falsa.

En el ejemplo de la base operativa, habría que comprobar algo más que la calidad del registro o el tiempo ahorrado: si quienes participan pueden explicar las decisiones, identificar dependencias y detectar errores. Si, en condiciones comparables, la automatización preservara o mejorara esas capacidades sin exigir la redacción manual, habría que revisar la idea de que ese paso es necesario para comprender en ese contexto. La función podría haberse conservado por otra vía.

Para estudiar el papel de la articulación, un contraste posible sería comparar respuestas con las mismas afirmaciones, razones y fuentes, pero diferente grado de elaboración. Habría que observar no solo si cambia la confianza del usuario, sino si comprende mejor el argumento y si su confianza se ajusta a la corrección y al respaldo de la afirmación. Una exposición más clara podría ayudar a entender una justificación válida; aumentar la confianza no constituiría, por sí solo, un error.

Si una comparación suficientemente informativa mostrara que la mayor elaboración no refuerza la confianza injustificada, habría que limitar el papel atribuido a ese factor en las condiciones estudiadas. Si favoreciera una comprensión más precisa y una confianza mejor fundada, ese resultado debería reconocerse como un beneficio, no reinterpretarse para mantener intacta la advertencia.

Estas propuestas no son resultados ni protocolos experimentales completos. Señalan qué partes del argumento requieren investigación y qué hallazgos podrían delimitarlas. El criterio no es que toda comprobación proceda de una persona, sino que la explicación no pueda decidir a conveniencia qué cuenta como evidencia a su favor.

(6) Popper, K. R. (2002). *The logic of scientific discovery*. Routledge. (Publicado originalmente en alemán como *Logik der Forschung* en 1934, con pie de imprenta de 1935; primera edición inglesa, preparada por el autor, en 1959). <https://www.routledge.com/The-Logic-of-Scientific-Discovery/Popper/p/book/9780415278447>

9. Conclusión. Reconocer cuándo no debemos creer

El avance no consiste únicamente en producir respuestas más convincentes. Consiste también en construir mejores formas de reconocer cuándo no debemos creerlas y cuándo sí disponemos de razones para confiar en ellas.

La distinción central de este ensayo no enfrenta inteligencia artificial y pensamiento humano. Separa la elaboración de una explicación del respaldo que permite aceptarla, y la producción de un documento de la comprensión que ese documento no acredita por sí solo. En ambos casos, la asistencia automática puede contribuir a la tarea o desplazar una parte de su función. Determinar qué ha ocurrido exige examinar el proceso, no deducirlo de quién redactó el resultado.

El texto ofrece una síntesis y unas preguntas de investigación, no una demostración de la frecuencia o la magnitud de estos riesgos. Su utilidad dependerá de que permita reconocer tanto los procesos que protegen una premisa frente a la revisión como aquellos que consiguen corregirla. Encontrar los segundos debe limitar la aplicación de la advertencia, no convertirse en otra forma de confirmarla.

La coherencia puede mejorar una explicación. No la exime de responder a aquello que pretende explicar.

Bibliografía

- (1) Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79.
<https://doi.org/10.1145/3624724> — pág. 2
- (2) Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220. <https://doi.org/10.1037/1089-2680.2.2.175> — pág. 3
- (3) Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688.
<https://doi.org/10.1016/j.tics.2016.07.002> — pág. 4
- (4) Calvo Espinosa, A. (2026). *Describir no es explicar: Inteligencia artificial, conciencia atribuida y los límites del «solo son algoritmos»* (Preprint v1.0-ES). Zenodo. DOI: 10.5281/zenodo.23249420 — pág. 5
- (5) Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askeel, 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. (2024). Towards understanding sycophancy in language models. *International Conference on Learning Representations (ICLR)*.
https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html — pág. 6
- (6) Popper, K. R. (2002). *The logic of scientific discovery*. Routledge. (Publicado originalmente en alemán como *Logik der Forschung* en 1934, con pie de imprenta de 1935; primera edición inglesa, preparada por el autor, en 1959).
<https://www.routledge.com/The-Logic-of-Scientific-Discovery/Popper/p/book/9780415278447> — pág. 8

Nota editorial y declaraciones

Nota editorial

Este texto es un ensayo de análisis conceptual y exploratorio, centrado en la interacción con sistemas conversacionales basados en modelos de lenguaje. Las experiencias relatadas, el ejemplo compuesto y los escenarios hipotéticos se emplean para formular problemas y criterios de revisión; no constituyen un estudio empírico sistemático ni permiten generalizar resultados a todos los sistemas o contextos de uso.

Durante su elaboración se ha utilizado ChatGPT como apoyo para la redacción, la organización del texto, la revisión de argumentos y la propuesta de referencias bibliográficas. El planteamiento de partida y las experiencias descritas corresponden al autor, que mantiene las decisiones editoriales y asume la responsabilidad sobre la versión publicada. La asistencia de IA no se considera una validación de las afirmaciones del ensayo.

Agradecimientos

A las personas con las que he compartido conversaciones sobre sus interacciones con sistemas de inteligencia artificial. Varias de esas conversaciones obligaron a precisar lo que aquí se sostiene.

A Keise, por señalar mis inconsistencias cuando menos quiero verlas, siempre desde el cariño.

Las decisiones finales, y los errores que queden, son míos.

Financiación

Este trabajo no ha recibido financiación específica de agencias públicas, entidades comerciales ni organizaciones sin ánimo de lucro.

Conflictos de interés

El autor declara no tener conflictos de interés relacionados con el contenido de esta obra. El texto no representa la posición oficial de ninguna institución religiosa, académica ni empresarial, sino exclusivamente la responsabilidad del autor.

Cómo citar

Cita recomendada (APA 7): Calvo Espinosa, A. (2026). Coherencia No Es Validación: IA, Circuitos Cerrados y el Riesgo de Automatizar la Comprensión (Preprint v1.0-ES). DOI: 10.5281/zenodo.23249485

Cita corta en el texto: (Calvo Espinosa, 2026)