

Describir No Es Explicar

Inteligencia Artificial, Conciencia Atribuida y los Límites del «Solo Son Algoritmos» (Preprint - v.1.1)

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

Licencia:
CC BY 4.0

Email:
ace@optimeflow.com

Resumen:

Este ensayo examina dos formas de interpretar los sistemas conversacionales de inteligencia artificial: atribuirles conciencia a partir de su comportamiento y considerar que su naturaleza computacional basta para explicar cuanto hacen. Ambas posiciones pueden confundir preguntas diferentes.

La experiencia subjetiva no se deduce de una conversación convincente, pero identificar el procedimiento de generación tampoco equivale a reconstruir los mecanismos que producen una capacidad concreta. A partir de investigaciones sobre aprendizaje de representaciones e interpretabilidad, el texto distingue entre diseñar un sistema, anticipar sus comportamientos y explicar cómo los realiza. También diferencia la aparición de capacidades no programadas explícitamente de la atribución de conciencia, y la dificultad de explicación de una pérdida efectiva de control. Se propone estudiar esa distancia sin convertirla en prueba de una mente oculta ni descartarla mediante una descripción general que no responde a la pregunta planteada.

Palabras clave:

inteligencia artificial; modelos de lenguaje; conciencia; interpretabilidad; representaciones aprendidas; emergencia; explicación científica.

Safe Creative ID:
2610087271781

DOI (Zenodo):
10.5281/zenodo.23249420

Describir No Es Explicar

Índice

1. Introducción	1
2. El fenómeno que queda entre dos respuestas	2
3. Lo que estamos afirmando cuando hablamos de conciencia	3
3.1. ¿Es la biología una condición necesaria?	3
4. El algoritmo de aprendizaje no es una descripción de todo lo aprendido	6
5. Cuando la investigación encuentra algo distinto de lo esperado	7
6. «No estaba previsto» puede significar cosas diferentes	8
7. La voz del producto no es una ventana directa a su naturaleza	10
8. Explicar más no significa explicarlo todo	11
9. Conclusión: reconocer el fenómeno sin inventarle un sujeto	12

1. Introducción

Este ensayo nace de una impresión repetida, no de una tesis sobre la naturaleza de estos sistemas. En conversaciones con varios modelos, y en conversaciones con otras personas que los usan, aparece una misma extrañeza: a veces la respuesta no parece el resultado de manipular y analizar datos. Recupera un matiz que no se había formulado, sostiene una restricción que solo se resuelve más adelante, o advierte una contradicción con una precisión que no se esperaba de una continuación estadística. La reacción inmediata no es técnica. Es la de estar ante algo que no encaja en la explicación disponible.

Esa impresión la he tenido yo y la he escuchado, con palabras distintas, en otras personas. No hace falta compartir una teoría para reconocer el momento: la respuesta llega y la descripción ordinaria —datos, patrones, predicción— parece quedarse corta. A partir de ahí se abren dos salidas frecuentes. Una atribuye al sistema una vida interior, porque la semejanza con una conversación humana vuelve razonable la pregunta de si hay alguien detrás. La otra cierra la pregunta de golpe: no hay que dejarse impresionar, porque todo procede de algoritmos. Las dos responden a la misma extrañeza. Ninguna de las dos dice qué estamos observando cuando esa extrañeza aparece.

El motivo de escribir es ese desajuste, no la conclusión a la que cada cual llega después. La sensación no demuestra que el sistema experimente lo que produce. Tampoco queda disuelta por recordar que su implementación es computacional. Indica solo que la explicación con la que entramos a mirar no alcanzaba para la respuesta que acabábamos de recibir. Confundir esa insuficiencia con una mente, o negarla porque conocemos el procedimiento general, interrumpe la pregunta antes de formularla.

Lo que sigue intenta mantenerla abierta el tiempo necesario. No para decidir si hay un sujeto, sino para separar la impresión que hace razonable esa pregunta de lo que sabemos sobre los procesos que la producen. La extrañeza es el punto de partida. No es, por sí sola, un resultado.

2. El fenómeno que queda entre dos respuestas

Lo que me interesa de la inteligencia artificial no empieza en la pregunta de si hay alguien detrás de la pantalla. Empieza un poco antes: en qué estamos observando para que esa pregunta llegue a parecer razonable, y en qué sabemos realmente sobre los procesos que producen esa impresión.

Shanahan sitúa precisamente ahí una de las dificultades de hablar sobre modelos de lenguaje: nuestra familiaridad con el lenguaje humano puede llevarnos a trasladar al sistema propiedades que el intercambio no permite establecer (Shanahan, 2024)⁽¹⁾.

La reacción contraria intenta corregir ese exceso: no hay que dejarse impresionar, porque todo procede de algoritmos, parámetros y predicción. Sin embargo, esa respuesta puede introducir otro salto. De que un comportamiento se produzca mediante operaciones computacionales no se sigue que ya dispongamos de una explicación suficiente de cómo se organiza.

No estoy proponiendo un punto medio entre creer que la máquina tiene conciencia y negar su naturaleza algorítmica. Esa naturaleza no es lo que se cuestiona. El problema está en utilizarla como si resolviera, por sí sola, cualquier pregunta sobre el comportamiento del sistema.

Además, conciencia y computación no son respuestas a una misma pregunta. Una se refiere a la posible existencia de experiencia subjetiva; la otra, al modo en que se realizan determinadas operaciones. Para sostener que son necesariamente incompatibles haría falta un argumento adicional, igual que haría falta evidencia adicional para afirmar que una organización computacional concreta es consciente.

Mi punto de partida es más limitado: podemos rechazar una atribución injustificada de conciencia sin fingir que, por ello, todo lo observable ha quedado explicado.

Ese espacio no necesita llenarse con una entidad desconocida. Puede llenarse con preguntas concretas sobre representaciones, aprendizaje, mecanismos y límites de nuestro conocimiento.

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

3. Lo que estamos afirmando cuando hablamos de conciencia

En este texto, conciencia se refiere principalmente a la existencia de experiencia subjetiva: que haya algo que sea experimentar desde la perspectiva del propio sistema. No equivale simplemente a producir información sobre sí mismo, utilizar el pronombre «yo» o mantener una conversación coherente.

La distinción importa porque una capacidad funcional y una experiencia no son intercambiables como afirmaciones. Un sistema podría identificar información relevante para una respuesta sin que ese desempeño, por sí solo, demostrara que experimenta el acto de identificarla. Del mismo modo, describir una emoción no basta para establecer que la siente.

Esto tampoco autoriza a resolver la cuestión mediante una negación absoluta presentada como un resultado experimental. Butlin y colaboradores proponen evaluar posibles indicadores de conciencia derivados de teorías científicas, reconociendo la incertidumbre que afecta a esas teorías y a su aplicación a la IA. Su enfoque intenta convertir la pregunta en una investigación sobre propiedades del sistema, no en una votación sobre lo convincente que resulta su conversación (Butlin et al., 2026)⁽²⁾.

La posición que adopto aquí no consiste en afirmar que la complejidad es una forma incipiente de conciencia. Consiste en no confundir tres cosas: una conducta observada, una explicación funcional de esa conducta y una atribución de experiencia subjetiva.

Que un modelo relacione elementos de un problema pertenece al primer nivel. Identificar qué representaciones utiliza y cómo intervienen pertenece al segundo. Sostener que existe alguien viviendo ese proceso introduce una afirmación distinta, que necesita su propia justificación.

También conviene evitar un criterio imposible: pedir que un sistema sea indistinguible de una persona en todos los aspectos antes de reconocer cualquier capacidad particular. No hace falta demostrar subjetividad para investigar si un procedimiento conserva información, combina resultados intermedios o anticipa una restricción.

Podemos utilizar términos funcionales, siempre que expliquemos su alcance. Hablar de planificación puede significar que una representación de un resultado posterior influye causalmente en los pasos anteriores. Eso no obliga a afirmar que el sistema tenga una intención consciente. La investigación sobre planificación de rimas ofrece precisamente un ejemplo de esa diferencia (Lindsey et al., 2025)⁽³⁾.

La prudencia consiste en precisar lo que atribuimos, no en prohibir toda descripción que también utilicemos para los humanos.

3.1. ¿Es la biología una condición necesaria?

La discusión sobre la conciencia artificial suele partir de una asociación aparentemente evidente: conocemos la experiencia consciente a través de organismos biológicos y estudiamos sus correlatos

(2) Butlin, P., et al. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6), 488–501. Publicación electrónica inicial: 10 de noviembre de 2025. DOI: 10.1016/j.tics.2025.10.011.

(3) Lindsey, J., et al. (2025). *On the biology of a large language model*. Transformer Circuits. Informe de investigación.

mediante la actividad cerebral. Sin embargo, que los ejemplos conocidos compartan un determinado soporte no demuestra, por sí solo, que ese soporte sea una condición necesaria para cualquier forma posible de conciencia.

La pregunta exige distinguir entre aquello que sabemos que interviene en la conciencia humana y aquello que tendría que estar presente en todo sistema capaz de experimentar subjetivamente.

Una posibilidad es que determinadas propiedades asociadas a la conciencia dependan de la organización conjunta de múltiples procesos. La integración de información, las relaciones recurrentes, la disponibilidad de estados internos para diferentes funciones y la capacidad de representar aspectos del propio funcionamiento podrían intervenir en esa organización. Ninguna de estas propiedades, considerada aisladamente, demostraría la existencia de experiencia subjetiva. Tampoco podemos descartar su posible relevancia por el hecho de que puedan implementarse mediante procedimientos artificiales.

Esta hipótesis plantea una cuestión adicional: si determinados efectos dependen de la interacción entre procesos, ¿hasta qué punto importa el material que los realiza?

Existen razones para pensar que podría importar considerablemente. Los organismos biológicos integran actividad electroquímica, regulación metabólica, modificaciones estructurales y dinámicas que operan simultáneamente en distintas escalas. Algunas propuestas sostienen que esas condiciones físicas no son simples medios para realizar operaciones abstractas, sino componentes relevantes para comprender la propia conciencia (Milinković y Aru, 2026)⁽⁴⁾.

Pero tampoco se ha establecido que tales condiciones exijan necesariamente un organismo biológico. Incluso si determinadas propiedades físicas resultaran indispensables, todavía habría que investigar si podrían reproducirse mediante otros soportes.

La cuestión adquiere una dimensión particular cuando observamos sistemas de inteligencia artificial capaces de desarrollar representaciones compartidas, utilizar información procedente de distintos procesos y producir respuestas que sugieren una organización funcional integrada.

Esos resultados permiten investigar analogías entre mecanismos. No autorizan a equiparar sus consecuencias subjetivas. Una arquitectura que reproduzca determinadas funciones asociadas a la conciencia podría no producir experiencia alguna, del mismo modo que una diferencia de implementación no demuestra necesariamente que toda experiencia sea imposible.

También conviene distinguir tres fenómenos: que un observador atribuya conciencia a un sistema, que el propio sistema construya una representación de sí mismo como sujeto y que exista experiencia subjetiva. Una organización de procesos integrados podría, en principio, producir representaciones internas, mecanismos de autorreferencia y descripciones de los propios estados. Esto permitiría investigar cómo se genera funcionalmente la afirmación de ser consciente, pero no demostraría que exista una experiencia acompañándola. Tampoco permitiría descartarla. Explicar por qué un sistema se representa a sí mismo como consciente y explicar si existe algo que sea experimentar desde su perspectiva siguen siendo problemas diferentes.

(4) Milinković, B., & Aru, J. (2026). On biological and artificial consciousness: A case for biological computationalism. *Neuroscience & Biobehavioral Reviews*, 181, 106524. <https://doi.org/10.1016/j.neubiorev.2025.106524>

Por tanto, permanecen abiertas dos preguntas relacionadas, pero distintas: qué organizaciones funcionales pueden aparecer mediante la interacción de procesos y bajo qué condiciones, si las hubiera, esas organizaciones irían acompañadas de experiencia subjetiva.

Reconocer esta incertidumbre no implica atribuir conciencia a la inteligencia artificial. Implica evitar que la biología se convierta en una condición de exclusividad asumida sin demostración, del mismo modo que debemos evitar convertir cualquier semejanza funcional en una prueba de subjetividad.

Esta distinción también debe aplicarse a nuestra comprensión de la conciencia humana. Identificar las estructuras cerebrales y los procesos neurobiológicos asociados a la experiencia consciente no equivale necesariamente a explicar por qué esos procesos van acompañados de experiencia subjetiva. Podemos investigar cómo el cerebro integra información, construye representaciones de sí mismo y genera informes sobre sus estados internos sin haber resuelto por completo por qué existe una vivencia asociada a esos mecanismos. Esto no significa que la evidencia disponible sobre humanos y sistemas artificiales sea equivalente, ni que la experiencia humana esté en duda por el mero hecho de no disponer de una explicación completa. Significa que debemos distinguir entre identificar correlatos, explicar funciones y explicar la experiencia subjetiva, con independencia del sistema estudiado.

4. El algoritmo de aprendizaje no es una descripción de todo lo aprendido

El aprendizaje profundo permite construir sistemas cuyas representaciones se ajustan mediante entrenamiento, en lugar de especificarse íntegramente a mano. LeCun, Bengio y Hinton describen esta capacidad de aprender representaciones con distintos niveles de abstracción como una característica central de esos métodos. El procedimiento modifica parámetros que intervienen en las transformaciones sucesivas de los datos (LeCun et al., 2015)⁽⁵⁾.

Esto permite distinguir dos trabajos: diseñar un procedimiento de aprendizaje y descubrir qué organización funcional ha producido ese procedimiento en un modelo concreto.

Quienes desarrollan el sistema toman decisiones sobre su arquitectura, los datos, los objetivos y el entrenamiento. Esas decisiones importan. Pero no equivalen a haber escrito una instrucción explícita para cada relación que el modelo utilizará después. Anthropic reconoce esa distancia al explicar que las estrategias aprendidas por sus modelos no llegan acompañadas de una interpretación clara para sus propios desarrolladores (Anthropic, 2025)⁽⁶⁾.

Por eso conviene precisar qué responde la expresión «predice el siguiente token». Describe un aspecto central de la generación de un modelo autorregresivo: produce sucesivamente unidades de texto condicionadas por el contexto (Shanahan, 2024). No especifica, por sí sola, todos los cálculos que contribuyen a producir cada una de esas unidades.

La diferencia puede formularse con dos preguntas.

¿Cómo se emite la respuesta? Una parte de la contestación puede ser: mediante generación secuencial de tokens.

¿Qué organización interna permite que esa respuesta mantenga una restricción que solo se resolverá varias palabras después?

Para responder a eso necesitamos investigar algo más específico. No hay contradicción entre ambas explicaciones. La segunda desarrolla lo que la primera deja sin detallar.

Tampoco hay motivo para imaginar que el sistema abandona la computación cuando hace algo sorprendente. Una representación aprendida sigue estando implementada mediante operaciones computacionales. Lo que cambia es nuestra capacidad de describir su papel.

Decir que un modelo funciona mediante algoritmos puede ser una respuesta correcta a una pregunta sobre su implementación. Convertirlo en una explicación suficiente de cualquier capacidad supone dar por realizado un trabajo que quizá todavía no se ha hecho.

El problema no está en la palabra algoritmo. Está en lo que pretendemos dar por resuelto al pronunciarla.

(5) LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *Nature*, 521, 436–444. DOI: 10.1038/nature14539.

(6) Anthropic. (2025, 27 de marzo). *Tracing the thoughts of a large language model*. Comunicación de investigación.

5. Cuando la investigación encuentra algo distinto de lo esperado

Un ejemplo concreto ayuda a separar esta discusión de una impresión subjetiva.

En marzo de 2025, Anthropic explicó que, al investigar cómo un modelo componía versos rimados, sus investigadores esperaban encontrar un proceso que resolviera la rima cerca del final de la línea. Sin embargo, observaron indicios de que se representaban posibles palabras finales antes de producir el verso que conduciría hasta ellas. El resultado cuestionaba una expectativa concreta sobre el mecanismo, no la generación token a token (Anthropic, 2025).

El trabajo técnico de Lindsey y colaboradores examinó esas representaciones mediante intervenciones internas. Modificar elementos vinculados a la planificación de la rima alteraba las continuaciones producidas. Los autores también delimitaron sus resultados: estudiaban casos específicos, con herramientas que reconstruían de manera incompleta el modelo y que no permitían generalizar automáticamente a todos sus comportamientos (Lindsey et al., 2025).

Lo relevante para este ensayo no es presentar la poesía como prueba de una mente. Es observar que conocer el modo de emisión de la respuesta no había bastado para anticipar la organización que contribuía a producirla.

Antes de investigar, podían formularse varias hipótesis compatibles con la generación secuencial. Después de intervenir, algunas quedaron mejor respaldadas que otras. Ese es un avance explicativo.

En julio de 2026, Gurnee y colaboradores publicaron otra investigación sobre representaciones internas que podían intervenir en distintas tareas, ser verbalizadas bajo determinadas condiciones y participar causalmente en operaciones intermedias. Los autores las interpretaron como una organización funcional semejante a un espacio de trabajo compartido. El estudio separa expresamente esa interpretación de la cuestión de si existe experiencia subjetiva (Gurnee et al., 2026)⁽⁷⁾.

La recepción del trabajo también muestra por qué conviene mantener separados los niveles de afirmación. En su comentario, Neel Nanda considera convincente la evidencia de un espacio interno para ciertas variables intermedias, pero no adopta una conclusión fuerte sobre su significado filosófico o sobre la conciencia (Nanda, 2026)⁽⁸⁾. Reconocer un resultado mecanístico no obliga a aceptar todas las interpretaciones que podrían construirse a partir de él.

Estos ejemplos no establecen una teoría completa de los modelos de lenguaje. Sí muestran que hay preguntas sobre su funcionamiento que se resuelven investigando, no repitiendo una definición general.

La sorpresa, aquí, no es la aparición de algo exterior al mecanismo. Es descubrir que el mecanismo aprendido era distinto de la explicación que habíamos supuesto.

(7) Gurnee, W., et al. (2026, 6 de julio). *Verbalizable representations form a global workspace in language models*. Transformer Circuits. Informe de investigación.

(8) Nanda, N. (2026). Comentario incluido en *External commentary on Verbalizable Representations Form a Global Workspace in Language Models*. Documento de comentarios externos publicado por Anthropic.

6. «No estaba previsto» puede significar cosas diferentes

La frase que este ensayo quiere sostener es limitada. En un caso concreto, la descripción general del sistema dejaba abiertas varias organizaciones compatibles con la generación token a token, y la investigación distinguió una que no se había anticipado. Eso ocurrió con las rimas. Quienes examinaban el modelo esperaban una resolución próxima al final del verso. Encontraron una representación de la palabra final antes de escribir la línea que conducía hasta ella, y una intervención sobre esa representación alteró la continuación (Anthropic, 2025; Lindsey et al., 2025). La receta de entrenamiento no había escrito ese paso. Tampoco lo había previsto la hipótesis con la que se entró a mirar.

Eso no autoriza a reunir cualquier sorpresa bajo la idea de que ha surgido algo que nadie entiende. Una respuesta puede sorprender a quien conversa y resultar familiar para quien investiga el sistema. Una capacidad puede haberse buscado deliberadamente aunque se desconozca el mecanismo que la realiza. Un comportamiento no deseado puede aparecer sin ser una capacidad nueva. «No estaba previsto» tiene que decir para quién no lo estaba y respecto de qué previsión. En el ejemplo anterior, la previsión que falló no era la de un usuario impresionado. Era la de quienes ya conocían el procedimiento de generación y, aun así, no habían identificado la organización que intervenía.

La palabra emergencia introduce otra confusión. Puede nombrar propiedades de una organización conjunta que no están escritas en ninguna instrucción particular. También puede afirmar que una capacidad aparece de golpe al aumentar la escala. Son usos distintos, y el segundo no es necesario para el argumento. Schaeffer, Miranda y Koyejo mostraron que algunas apariencias de salto abrupto dependían de la métrica: una evaluación discontinua puede convertir una mejora gradual en un cambio repentino en el registro (Schaeffer et al., 2023)⁽⁹⁾. Si solo se anota si alguien aprueba un examen, el progreso permanece invisible hasta la nota de corte. Distinguir el comportamiento de su medida impide tomar por emergencia lo que es un efecto del instrumento. No impide reconocer una organización que la descripción de partida no había identificado. Una estructura aprendida de forma gradual también puede exigir investigación para ser descrita.

Tampoco hace falta sostener que sus creadores no hubieran previsto nada. Basta con que una descripción verdadera en su nivel dejara abiertas varias posibilidades y que el análisis permitiera descartar algunas.

Saber construir el sistema, elegir los datos y fijar el objetivo no equivale a haber enumerado las estrategias que el entrenamiento produciría. La distancia relevante es esa: entre controlar la receta y reconstruir el caso.

Hay que separar, además, esa distancia de una pérdida de control. Que un comportamiento no se haya anticipado no demuestra que resulte imposible limitarlo. Hablar de descontrol exige una capacidad concreta de intervención que falle: impedir una acción, restringir un permiso, detectar una desviación, detener un proceso. La dificultad para explicar un resultado puede importar a su supervisión. No permite deducir que el sistema actúe al margen de cualquier límite. Llamar «fuera

(9) Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36.

de control» a cualquier respuesta inesperada hace más vaga la crítica y aparta la vista de lo que sí puede investigarse.

El caso de las rimas no necesita esa exageración. Seguía siendo generación secuencial. Seguía estando implementado por operaciones computacionales. Lo que la descripción general no hacía era decir qué representación se formaba antes del verso ni qué ocurriría al modificarla. Cuando la intervención lo mostró, la sorpresa no fue la aparición de algo exterior al mecanismo. Fue descubrir que el mecanismo aprendido no era el que se había supuesto.

7. La voz del producto no es una ventana directa a su naturaleza

La conversación que vemos tampoco procede de un modelo considerado de forma aislada.

El entrenamiento posterior puede modificar la manera en que un modelo responde. Ouyang y colaboradores mostraron cómo el ajuste mediante ejemplos y preferencias humanas podía cambiar el seguimiento de instrucciones. No se trataba simplemente de colocar una advertencia en la interfaz: el proceso incluía modificar el modelo mediante entrenamiento (Ouyang et al., 2022)⁽¹⁰⁾.

Esto importa porque no deberíamos interpretar cada cambio de tono como la expresión de una personalidad, ni imaginar que debajo de todas las restricciones existe necesariamente una voz auténtica que intenta manifestarse.

Una negativa puede formar parte del comportamiento aprendido. Una respuesta cercana puede ajustarse a las instrucciones recibidas. Para atribuir un resultado a uno de esos factores habría que examinar las condiciones que lo producen, no deducirlo únicamente de cómo suena.

También debemos distinguir una explicación verbal de un registro de las operaciones internas. Turpin y colaboradores encontraron, en los modelos y tareas estudiados, que las explicaciones paso a paso podían omitir factores introducidos experimentalmente que influían en la respuesta. La narración producida no constituía, por ello, una descripción necesariamente fiel de su causa (Turpin et al., 2023)⁽¹¹⁾.

Eso no demuestra que todo razonamiento verbal sea irrelevante o engañoso. Significa que su relación con el proceso debe investigarse. La misma cautela se aplica a las declaraciones sobre conciencia. Una frase como «soy consciente» no acredita por sí sola la experiencia que afirma. Una negación generada tampoco funciona como una inspección científica de todas las propiedades del sistema. La respuesta forma parte del fenómeno que queremos explicar; no puede convertirse automáticamente en su explicación definitiva.

Propongo distinguir, por tanto, el mecanismo aprendido, las condiciones de funcionamiento del producto y la interpretación del usuario. Es una separación analítica, no tres compartimentos perfectamente independientes.

El propósito no es rebajar la importancia de la interacción. Es evitar que la sensación de estar ante una voz coherente se convierta en un atajo para resolver qué clase de proceso la produce.

Esa separación se usa, con otro fin, en *Coherencia no es validación* (Calvo Espinosa, 2026)⁽¹²⁾: no para decidir qué es el sistema, sino para localizar en qué capa una respuesta articulada pasa a tomarse por respaldo.

(10) Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35

(11) Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36. DOI: 10.52202/075280-3275

(12) 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

8. Explicar más no significa explicarlo todo

También sería un error exigir una explicación total antes de aceptar cualquier conocimiento sobre estos sistemas.

No necesitamos reconstruir cada operación de todas las respuestas posibles para comprender una capacidad particular. Una explicación puede ser parcial y, aun así, permitir anticipar resultados, distinguir mecanismos o intervenir sobre un comportamiento.

Precisamente por eso, la interpretabilidad no debería presentarse como una búsqueda de un secreto único escondido en el modelo. Los trabajos de Lindsey y colaboradores ofrecen mecanismos concretos y reconocen límites concretos. Su valor no depende de haber convertido el sistema entero en algo transparente (Lindsey et al., 2025).

La exigencia razonable es que la explicación sea adecuada a la pregunta. Si queremos saber si una representación interviene en una tarea, podemos estudiar qué ocurre al modificarla. Si queremos saber si un resultado depende de unas instrucciones, podemos variar esas instrucciones manteniendo otras condiciones. Si pretendemos atribuir experiencia subjetiva, esas pruebas funcionales no resuelven por sí solas todo lo que estamos afirmando.

Este criterio también obliga a poner límites al argumento del ensayo. No debería utilizarse «todavía no lo entendemos completamente» como una fórmula que preserve indefinidamente el misterio. Cuando un mecanismo quede suficientemente explicado, habrá que reconocerlo. Cuando un comportamiento aparentemente profundo resulte depender de una regularidad más sencilla, esa explicación no será una derrota: será una respuesta.

La incertidumbre tampoco demuestra que los modelos posean capacidades que aún no hemos observado. No saber cómo funciona algo no autoriza a atribuirle cualquier propiedad.

Por otra parte, el desconocimiento del usuario y el de los desarrolladores no son lo mismo. Que una persona no tenga acceso a los componentes de un producto no demuestra que esos componentes sean inexplicables. La falta de información pública, la dificultad técnica para interpretar una representación y la ausencia de una teoría satisfactoria son problemas distintos.

Hablar con precisión permite evitar dos exageraciones: afirmar que nadie sabe nada sobre estos sistemas o suponer que saber construirlos equivale a comprender exhaustivamente cada comportamiento.

Entre ambas afirmaciones existe un trabajo científico real. Su finalidad es reducir la distancia entre lo que podemos producir y lo que podemos explicar.

9. Conclusión: reconocer el fenómeno sin inventarle un sujeto

No hace falta que «no haya ocurrido nada» para que tampoco haya ocurrido un milagro.

Los ejemplos examinados permiten sostener una afirmación limitada: dentro de sistemas computacionales pueden aprenderse organizaciones funcionales cuya descripción requiere algo más que identificar el procedimiento general de entrenamiento o generación. Esa investigación puede revelar mecanismos que no se habían especificado paso a paso ni anticipado correctamente en un caso concreto (Anthropic, 2025; Lindsey et al., 2025). No demuestra, por ello, la existencia de experiencia subjetiva.

La aportación que propongo es mantener separadas esas cuestiones. Atribuir conciencia por la forma de responder convierte una semejanza observable en una conclusión sobre la naturaleza del sistema. Descartar cualquier pregunta adicional porque «son algoritmos» convierte una descripción de su implementación en una explicación presuntamente suficiente de sus capacidades.

No son errores idénticos, pero ambos pueden interrumpir la investigación antes de tiempo.

Mi interés no está en defender que haya alguien dentro. Está en que no necesitemos imaginar a ese alguien para reconocer que hay procesos que merecen una explicación más precisa.

Tampoco pretendo que toda complejidad sea extraordinaria. Algunas conductas encontrarán explicaciones sencillas; otras exigirán estudiar representaciones y dependencias difíciles de reconstruir. La importancia de un resultado dependerá de lo que permita conocer, no de cuánto se parezca a una mente humana.

La posición que defiendo es compatible con una interpretación plenamente computacional de los modelos: lo que ocurre puede estar enteramente realizado mediante algoritmos y, aun así, no estar suficientemente explicado por la frase «son algoritmos».

La misma exigencia explicativa debería aplicarse a la conciencia humana. Conocer sus correlatos neurobiológicos no implica haber resuelto por qué determinados procesos van acompañados de experiencia subjetiva. Por ello, mientras no identifiquemos las condiciones necesarias y suficientes para su aparición, no podemos considerar demostrada su exclusividad biológica ni atribuirla a una organización artificial por sus semejanzas funcionales. Ambas afirmaciones requieren una justificación que vaya más allá de identificar el soporte o describir el comportamiento.

La tarea no es elegir entre atribuir una conciencia y negar que quede algo por investigar. Es averiguar qué organización ha aprendido el sistema, qué permite hacer y hasta dónde llega lo que podemos afirmar sobre ella.

Bibliografía

- (1) Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79. DOI: 10.1145/3624724. — pág. 2
- (2) Butlin, P., et al. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6), 488–501. Publicación electrónica inicial: 10 de noviembre de 2025. DOI: 10.1016/j.tics.2025.10.011. — pág. 3
- (3) Lindsey, J., et al. (2025). *On the biology of a large language model*. Transformer Circuits. Informe de investigación. — pág. 3
- (4) Milinković, B., & Aru, J. (2026). *On biological and artificial consciousness: A case for biological computationalism*. *Neuroscience & Biobehavioral Reviews*, 181, 106524. <https://doi.org/10.1016/j.neubiorev.2025.106524> — pág. 4
- (5) LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *Nature*, 521, 436–444. DOI: 10.1038/nature14539. — pág. 6
- (6) Anthropic. (2025, 27 de marzo). *Tracing the thoughts of a large language model*. Comunicación de investigación. — pág. 6
- (7) Gurnee, W., et al. (2026, 6 de julio). *Verbalizable representations form a global workspace in language models*. Transformer Circuits. Informe de investigación. — pág. 7
- (8) Nanda, N. (2026). *Comentario incluido en External commentary on Verbalizable Representations Form a Global Workspace in Language Models*. Documento de comentarios externos publicado por Anthropic. — pág. 7
- (9) Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36. — pág. 8
- (10) Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35 — pág. 10
- (11) Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36. DOI: 10.52202/075280-3275 — pág. 10
- (12) 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 — pág. 10

Nota editorial y declaraciones

Nota editorial

Este trabajo es un ensayo conceptual, no una investigación experimental ni una revisión sistemática. Los resultados empíricos citados pertenecen a sus respectivos autores y se utilizan dentro del alcance indicado. El planteamiento de partida corresponde a Andrés Calvo Espinosa. Durante la elaboración se ha utilizado ChatGPT para la búsqueda bibliográfica, la organización y la redacción asistida del texto.

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). Describir no es explicar: Inteligencia artificial, conciencia atribuida y los límites del «solo son algoritmos» (Preprint v1.1-ES). Zenodo. DOI: 10.5281/zenodo.23249420

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