

¿Pueden sentir las máquinas?

Indistinguibilidad, experiencia subjetiva y autonomía en la delegación cognitiva

Ricardo Bravo de Medina Hernández

Doctor en Psicología Clínica y de la Personalidad. Especialista en Psicología Clínica

11 de septiembre de 2026

RESUMEN

La capacidad de los sistemas conversacionales para producir respuestas emocionalmente convincentes plantea un problema que excede la identificación de su origen: qué autoriza a inferir la semejanza conductual sobre la experiencia subjetiva y sobre los efectos de la interacción. Este ensayo examina ese problema junto con una segunda cuestión: las consecuencias de delegar operaciones cognitivas en tecnologías que también proporcionan explicaciones y juicios. Mediante una selección crítica de literatura filosófica, psicológica y educativa, se defiende que la equivalencia del resultado es insuficiente para establecer tanto la presencia de experiencia en el sistema como la conservación del aprendizaje en el usuario. Se distinguen desempeño funcional, conciencia fenomenal, experiencia relacional y autonomía cognitiva. La evidencia sobre empatía percibida, ayudas convencionales e inteligencia artificial generativa permite reconocer beneficios situados y riesgos de sustitución de la práctica, sin justificar una tesis general de deterioro mental. Se propone una evaluación en dos planos complementarios: examinar la evidencia que fundamenta la atribución de estados mentales y valorar si la interacción preserva capacidades de comprensión, revisión y decisión. Esta propuesta es conceptual y requiere contrastación empírica; no constituye una prueba de conciencia ni un instrumento clínico validado.

Palabras clave: conciencia fenomenal; antropomorfización; descarga cognitiva; metacognición; agencia epistémica.

Introducción

Una respuesta que reconoce el sufrimiento puede producir alivio sin aclarar si quien la genera experimenta preocupación. De forma semejante, disponer de una explicación correcta puede resolver una tarea sin demostrar que su destinatario haya aprendido a resolverla. Ambas situaciones describen una dificultad de evaluación: el resultado visible admite interpretaciones diferentes sobre el proceso que lo produce y sobre las capacidades que deja tras de sí.

La inteligencia artificial conversacional vuelve relevante esta dificultad en contextos de apoyo emocional, aprendizaje y trabajo intelectual. Aquí no interesa determinar si cualquier conversación con una máquina constituye una relación ni si toda automatización disminuye capacidades. La pregunta es más delimitada: ¿qué podemos inferir de una respuesta funcionalmente adecuada cuando evaluamos, por un lado, la posible experiencia del sistema y, por otro, la participación cognitiva de la persona?

Este ensayo defiende que la indistinguibilidad de ciertos resultados ofrece evidencia situada de desempeño, pero no basta para establecer equivalencia subjetiva, reciprocidad interpersonal o conservación del aprendizaje. La tesis contiene una exigencia doble: evaluar las atribuciones mentales con evidencia pertinente y examinar la relación tecnológica por las posibilidades de comprensión, revisión y decisión que sostiene. Esta segunda dimensión se denomina aquí autonomía cognitiva, entendida como capacidad situada de orientar y examinar el propio juicio, también mediante apoyos externos.

El argumento se construye mediante una selección narrativa de literatura filosófica y empírica, sin pretensión de revisión sistemática. El punto de partida histórico se relaciona con

investigaciones sobre antropomorfización, empatía percibida y descarga cognitiva. Su articulación constituye una propuesta interpretativa: no se presupone un mecanismo psicológico único que explique ambos problemas ni se presentan resultados originales.

De la pregunta de Turing a la distinción entre función y experiencia

En *Computing Machinery and Intelligence*, Turing (1950) propuso reemplazar la pregunta sobre si las máquinas piensan por un problema definido mediante el juego de imitación. Su formulación inicial incluía un hombre, una mujer y un interrogador; posteriormente introducía una máquina como participante. El desplazamiento permitía examinar respuestas y juicios de identificación bajo condiciones específicas. No afirmaba que los mecanismos internos humanos y computacionales fueran idénticos ni resolvía la conciencia mediante una observación directa. Su discusión de la objeción de la conciencia confirma que conocía ese problema. Presentar su posición como indiferencia ante la existencia de pensamiento real sería una simplificación.

Conviene distinguir tres objetos de análisis. El desempeño funcional comprende operaciones como reconocer patrones, producir respuestas pertinentes o adaptar una explicación. La experiencia subjetiva designa el carácter vivido de un estado. La conciencia fenomenal remite a esa dimensión experiencial, que Nagel (1974) situó en el centro de su análisis de la perspectiva del organismo. En este ensayo, sentir dolor requiere una experiencia con carácter desagradable; describir dolor es una conducta cuya relación con esa experiencia necesita justificación.

Esta distinción tampoco convierte automáticamente toda función artificial en apariencia vacía. Una explicación puede ser correcta y una intervención comunicativa puede ser útil. El término *simulación* solo resulta informativo si se especifica qué se simula: una forma expresiva, una organización funcional o una experiencia atribuida. Confundir esos sentidos puede llevar tanto a sobreinterpretar capacidades como a negarlas sin examinarlas.

La objeción de Searle (1980) a identificar ejecución de programas y comprensión muestra una alternativa filosófica relevante, no una refutación experimental de toda conciencia artificial. Su argumento discute la suficiencia del programa y concede importancia a los poderes causales del sistema. La cuestión permanece abierta a posiciones rivales: distinguir función y experiencia no exige decidir de antemano qué soporte material podría realizar esta última.

La indistinguibilidad como evidencia limitada

La indistinguibilidad no es una propiedad absoluta que un sistema posea con independencia del observador. Es una relación entre alternativas, procedimientos y capacidades de discriminación. Dos respuestas pueden resultar indistinguibles para un lector y diferenciables para otro; también pueden dejar de parecer equivalentes al ampliar la conversación o cambiar la tarea. Además, no detectar una diferencia estadística no demuestra equivalencia: esa conclusión exige un diseño y criterios adecuados.

El valor de una prueba debe corresponder al alcance de la conclusión. Un procedimiento basado en identificación conversacional informa sobre identificación conversacional. Atribuir experiencia a partir de ese resultado requiere premisas adicionales sobre la relación entre conducta, organización y conciencia. Del mismo modo, la palabra «comprensión» puede nombrar rendimiento semántico, vivencia de significado o sensación de haber sido comprendido. Que esas acepciones se relacionen en la vida cotidiana no autoriza a intercambiarlas en la evaluación científica.

Butlin et al. (2026) proponen derivar indicadores de teorías de la conciencia y examinar su presencia en sistemas concretos. Su enfoque desplaza la evaluación hacia propiedades relevantes para esas teorías y reconoce incertidumbre tanto en los indicadores como en sus fundamentos. No ofrece un detector infalible. Su utilidad para este argumento reside en mostrar por qué una declaración verbal debe situarse dentro de un conjunto más amplio de evidencias, especialmente cuando puede generarse sin que concurren otras condiciones postuladas.

La consecuencia es una cautela epistémica, no una prohibición de atribuir conciencia. Exigir evidencia adicional deja abierta la posibilidad de encontrarla. También evita que el desconocimiento opere selectivamente: la ausencia de una demostración consensuada en chatbots no prueba imposibilidad futura, pero tampoco transforma cada respuesta afectiva en indicio suficiente de experiencia.

Esta cautela tiene una vertiente relacional. Aunque no se establezca experiencia en el sistema, sí puede estudiarse qué provoca su conducta en personas. La incertidumbre ontológica no suspende la investigación de consecuencias psicológicas observables.

La experiencia relacional no exige una reciprocidad demostrada

La atribución de estados mentales está influida por condiciones del observador y de la interacción. Epley et al. (2007) proponen un marco de antropomorfización que relaciona conocimiento accesible sobre agentes humanos, motivación de comprensión y necesidad de vinculación. Se trata de una explicación teórica de variaciones en la atribución, no de una descripción diagnóstica de quienes conversan con una IA.

En los sistemas conversacionales, esa atribución tampoco depende exclusivamente del usuario. El diseño puede facilitar interpretaciones de presencia mediante continuidad, personalización y expresividad. Peter et al. (2025) destacan el papel de las capacidades comunicativas y sus posibles beneficios y riesgos. Su perspectiva permite cuestionar una explicación centrada solo en la credulidad individual: las expectativas se forman también ante propiedades deliberadamente configuradas en la interfaz.

El estudio de Yin et al. (2024) diferencia el efecto del contenido y el de su origen atribuido. Los mensajes de IA favorecieron la sensación de escucha en las condiciones examinadas, mientras que identificarlos como artificiales redujo ese efecto. El resultado indica que la experiencia del receptor depende de algo más que de las palabras. No demuestra mejoría clínica sostenida ni reciprocidad afectiva.

Por su parte, Ayers et al. (2023) compararon respuestas médicas de un foro con respuestas generadas por un chatbot. Profesionales sanitarios valoraron más favorablemente la empatía de estas últimas, que también eran más extensas. La medida corresponde a evaluaciones de textos, no a sentimientos del sistema ni a resultados de una psicoterapia. Las diferencias de diseño impiden tratar ambos trabajos como pruebas intercambiables de una misma capacidad.

La inferencia propuesta es que el valor psicológico de una interacción y la atribución de experiencia a su emisor deben evaluarse por separado. Un alivio no se vuelve ficticio porque proceda de una mediación tecnológica. Sin embargo, valorar ese alivio tampoco obliga a atribuir a la herramienta compromiso, comprensión biográfica o responsabilidad asistencial.

Desde esta perspectiva, el criterio clínico relevante incluye la trayectoria del uso: si facilita elaborar una experiencia y participar en otros vínculos, o si va sustituyendo oportunidades de contraste y acción. Esta distinción es una propuesta de evaluación, no una afirmación de que los chatbots produzcan necesariamente aislamiento. Requiere atender a alternativas disponibles, preferencias y condiciones previas, sin patologizar el afecto hacia una tecnología.

Delegar operaciones: entre ampliación funcional y aprendizaje

La pregunta por si las máquinas acabarán pensando en lugar de nuestras mentes presupone una separación demasiado simple entre pensamiento interno y recurso externo. Clark y Chalmers (1998) sostienen que ciertos acoplamientos con el entorno pueden formar parte del proceso cognitivo. Su tesis de la mente extendida no identifica automáticamente extensión cognitiva y conciencia compartida. Tampoco implica que cualquier servicio consultado constituya una parte estable de la mente.

La investigación sobre descarga cognitiva ofrece un nivel empírico complementario. Risko y Gilbert (2016) describen cómo las acciones y recursos externos modifican las demandas de

procesamiento y examinan el papel de los juicios metacognitivos en su uso. La descarga puede mejorar la ejecución; sus efectos deben relacionarse con la tarea y con el criterio de éxito. Menor esfuerzo interno no equivale, por definición, a menor capacidad.

Las calculadoras permiten concretar esa distinción. El metaanálisis de [Ellington \(2006\)](#), sobre 42 estudios con calculadoras gráficas no-CAS, encontró beneficios en resultados educativos que variaban según las condiciones de enseñanza y evaluación. Esta evidencia cuestiona un supuesto general de incompatibilidad entre asistencia y aprendizaje, aunque no responde sobre toda forma de cálculo automatizado ni sobre exposición durante décadas.

La navegación plantea otro problema. [Dahmani y Bohbot \(2020\)](#) encontraron asociaciones entre mayor uso del GPS y peor memoria espacial en navegación autónoma, con 50 participantes y seguimiento de 13. El diseño observacional y ese seguimiento reducido limitan la atribución causal. Llegar correctamente y construir conocimiento del entorno son resultados distintos; los datos no demuestran atrofia cerebral general.

En apoyo ortográfico, [Rimbar \(2017\)](#) observó, en un cuasiexperimento con 30 estudiantes, correcciones del dictado asistido sin eliminación posterior de los errores. El resultado diferencia mejora del producto y adquisición, pero no demuestra pérdida de una habilidad previamente consolidada. Además, señalar errores, proponer alternativas y corregir automáticamente no son intervenciones equivalentes.

Los experimentos de [Sparrow et al. \(2011\)](#) relacionaron la expectativa de acceso posterior con diferencias en recuerdo del contenido y de su localización. El denominado efecto Google requiere cautela: [Hesselmann \(2020\)](#) no encontró evidencia concluyente para replicar su componente de activación informática mediante una tarea Stroop. Esa dificultad no refuta todos los experimentos originales, pero impide convertir un conjunto heterogéneo de resultados en una ley de amnesia digital.

Estos antecedentes no trazan una secuencia inevitable de decadencia. Muestran que el beneficio inmediato y la adquisición o conservación de una destreza pueden divergir. Distinguir delegación, falta de adquisición y pérdida de habilidad previa es indispensable para que *deskilling* no se convierta en una etiqueta aplicada a cualquier uso de ayuda.

La delegación del juicio y la evaluación de la comprensión

La IA generativa añade la posibilidad de recibir no solo una operación resuelta, sino una justificación formulada en lenguaje natural. En este ensayo, delegar el juicio significa aceptar una conclusión cuya evaluación sustantiva se transfiere al sistema. Esta definición no presupone que toda aceptación sea irracional: la división del trabajo intelectual exige confiar en capacidades ajenas. El problema aparece cuando esa confianza deja de guardar relación con la evidencia disponible o con las consecuencias del error.

[Bastani et al. \(2025\)](#) aportan evidencia experimental situada: en matemáticas de secundaria, una modalidad de tutor basado en GPT-4 mejoró el rendimiento asistido, pero produjo peores resultados al retirar la ayuda frente al grupo sin IA. Una modalidad con salvaguardas pedagógicas mitigó ese perjuicio. La diferencia se refiere a aprendizaje evaluado a corto plazo en un contexto concreto; no establece deterioro general o irreversible. El diseño de la ayuda importa tanto como su disponibilidad.

En una encuesta a 319 trabajadores, [Lee et al. \(2025\)](#) encontraron asociaciones entre mayor confianza en IA y menor pensamiento crítico autoinformado, además de cambios hacia verificación e integración de respuestas. Esos datos no permiten diagnosticar pérdida objetiva de capacidad. Sí sitúan la calibración de la confianza como problema relevante: juzgar cuándo revisar, contrastar o pedir asistencia experta también forma parte del trabajo cognitivo.

La conexión con el eje emocional es una analogía de evaluación, no una identidad causal. Una respuesta empática puede favorecer atribuciones sobre su emisor; una explicación fluida puede favorecer atribuciones sobre la propia comprensión. En ambos casos se propone examinar

si el resultado observado justifica la conclusión atribuida. Harían falta estudios específicos para demostrar que ambas inferencias comparten mecanismos o se potencian mutuamente.

De ello se deriva un criterio de autonomía compatible con la ayuda: poder reconocer incertidumbre, solicitar razones pertinentes, detectar una discrepancia relevante y recurrir a otra fuente cuando sea necesario. No exige reproducir cada operación. Exige conservar, dentro de posibilidades reales, una participación proporcionada en las decisiones que afectan a la propia vida.

Objeciones y límites de la propuesta

Una objeción funcionalista sostiene que una organización causal adecuada podría ser suficiente para la experiencia; separar conducta y subjetividad no debería imponer un privilegio biológico. La objeción es válida contra cualquier prohibición basada solo en el material del sistema. La tesis aquí defendida es más limitada: la equivalencia conversacional no establece por sí sola equivalencia de organización ni permite decidir qué teoría de la conciencia es correcta.

Una segunda objeción recuerda que tampoco observamos directamente la experiencia ajena. Exigir a una máquina una certeza inaccesible en humanos introduciría una asimetría injustificada. La respuesta consiste en exigir evidencia comparable en relevancia, no identidad de soporte ni certeza absoluta. La atribución puede apoyarse en convergencia de criterios y mantenerse revisable; un umbral razonable de confianza no necesita resolver definitivamente el problema de otras mentes.

Una tercera objeción cuestiona medir autonomía retirando la ayuda. Para quien depende de tecnologías de apoyo, esa retirada puede empeorar deliberadamente una capacidad distribuida. Por tanto, una prueba sin herramienta es pertinente cuando interesa aprendizaje independiente o contingencias de fallo; no es una definición universal de competencia. Evaluar el sistema persona-recurso puede ser igualmente necesario. La autonomía incluye disponer de apoyos fiables y poder elegirlos.

Finalmente, la literatura seleccionada es heterogénea y no permite estimar un efecto global. Una encuesta laboral, un ensayo escolar, una evaluación de textos y una propuesta filosófica responden a preguntas distintas. Faltan seguimientos prolongados, comparaciones entre diseños y medidas de transferencia. Las relaciones entre confianza, participación y beneficio también pueden depender del conocimiento previo y del contexto. La propuesta organiza preguntas; no proporciona una escala validada ni predice efectos individuales.

Implicaciones clínicas, educativas y de investigación

En clínica, conviene explorar la función de la interacción tecnológica sin descalificarla ni asumir eficacia terapéutica. Interesa qué necesidades atiende, cómo se integra con otros apoyos y si facilita decisiones examinadas. Distinguir reconocimiento emocional y confirmación de interpretaciones ayuda a evitar que una respuesta agradable se convierta automáticamente en criterio de verdad. Estas implicaciones generales no sustituyen evaluación individual.

En educación y trabajo, el objetivo de uso debería orientar la evaluación. Si se busca producir, importa calidad y verificación del resultado; si se busca aprender, deben examinarse adquisición y transferencia. Si la tarea exige conservar una habilidad crítica, procede evaluar mantenimiento y respuesta ante fallos. No todos esos objetivos requieren la misma proporción de asistencia ni las mismas pruebas.

Tabla 1
Preguntas complementarias para evaluar una interacción con IA

Plano	Pregunta pertinente	Conclusión que debe evitarse
Desempeño	¿Qué resuelve y bajo qué condiciones?	Una respuesta adecuada demuestra competencia general.
Experiencia atribuida	¿Qué evidencia fundamenta atribuir estados subjetivos?	Una expresión afectiva demuestra una vivencia.
Efecto relacional	¿Qué experimenta la persona y cómo evoluciona el uso?	El alivio inmediato demuestra eficacia clínica duradera.
Autonomía cognitiva	¿Qué puede comprender, revisar y decidir con sus apoyos?	Un buen resultado demuestra aprendizaje o independencia.

Nota. Elaboración conceptual propia a partir del argumento del ensayo. No constituye un instrumento psicométrico ni una prueba de conciencia.

Para investigar esta propuesta sería útil comparar ayudas orientadas a entregar soluciones con ayudas que soliciten una elaboración previa y permitan contraste. Deberían medirse conjuntamente ejecución asistida, comprensión, transferencia, confianza y uso posterior, evitando reducir todos esos resultados a satisfacción. En el eje relacional, separar origen real, identidad comunicada y contenido permitiría examinar efectos sin confundirlos con conciencia del sistema.

La responsabilidad tampoco recae únicamente en quien utiliza la tecnología. Una interfaz transparente, la posibilidad de revisar fuentes y unas condiciones institucionales que concedan tiempo para verificar hacen practicable la participación crítica. Pedir supervisión humana sin recursos para ejercerla puede convertirse en una asignación nominal de responsabilidad. Esta es una implicación normativa del argumento, cuya eficacia debe evaluarse en contextos concretos.

Conclusiones

La conducta conversacional disponible no resuelve por sí sola si las máquinas pueden sentir. La pregunta requiere distinguir capacidad funcional y experiencia subjetiva, y justificar qué indicadores permiten pasar de una a otra. Mientras esa cuestión permanece abierta, los efectos psicológicos en el interlocutor pueden investigarse y reconocerse sin atribuir automáticamente reciprocidad o eficacia clínica.

Del mismo modo, que una máquina ejecute operaciones cognitivas no implica que la mente humana deje de pensar. La evidencia examinada permite identificar beneficios de asistencia y perjuicios situados para el aprendizaje, pero no sustenta una tesis universal de atrofia. El problema exige especificar qué se delega, con qué propósito y qué capacidades se evalúan después.

El aporte del ensayo consiste en vincular ambas cuestiones mediante una exigencia de inferencia proporcionada: no deducir de la adecuación del resultado aquello que necesita evidencia adicional. Esto permite valorar tecnologías que amplían capacidades y, a la vez, examinar usos que dificultan la comprensión o el contraste. Una relación tecnológica favorable debería sostener posibilidades reales de orientación, revisión y decisión; determinar cómo hacerlo exige investigación y no puede resolverse mediante la semejanza de una respuesta.

Referencias

- Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6), 589-596. <https://doi.org/10.1001/jamainternmed.2023.1838>
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), Article e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6), 488-501. <https://doi.org/10.1016/j.tics.2025.10.011>
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7-19. <https://doi.org/10.1093/analys/58.1.7>
- Dahmani, L., & Bohbot, V. D. (2020). Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Scientific Reports*, 10, Article 6310. <https://doi.org/10.1038/s41598-020-62877-0>
- Ellington, A. J. (2006). The effects of non-CAS graphing calculators on student achievement and attitude levels in mathematics: A meta-analysis. *School Science and Mathematics*, 106(1), 16-26. <https://doi.org/10.1111/j.1949-8594.2006.tb18067.x>
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864-886. <https://doi.org/10.1037/0033-295X.114.4.864>
- Hesselmann, G. (2020). No conclusive evidence that difficult general knowledge questions cause a “Google Stroop effect”. A replication study. *PeerJ*, 8, Article e10325. <https://doi.org/10.7717/peerj.10325>
- Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Article 1121, pp. 1-22). Association for Computing Machinery. <https://doi.org/10.1145/3706598.3713778>
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435-450. <https://doi.org/10.2307/2183914>
- Peter, S., Riemer, K., & West, J. D. (2025). The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, 122(22), Article e2415898122. <https://doi.org/10.1073/pnas.2415898122>
- Rimbar, H. (2017). The influence of spell-checkers on students' ability to generate repairs of spelling errors. *Journal of Nusantara Studies*, 2(1), 1-12. <https://doi.org/10.24200/jonus.vol2iss1pp1-12>
- 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>
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424. <https://doi.org/10.1017/S0140525X00005756>
- Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776-778. <https://doi.org/10.1126/science.1207745>
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460. <https://doi.org/10.1093/mind/LIX.236.433>
- Yin, Y., Jia, N., & Waksalak, C. J. (2024). AI can help people feel heard, but an AI label diminishes this impact. *Proceedings of the National Academy of Sciences*, 121(14), Article e2319112121. <https://doi.org/10.1073/pnas.2319112121>