EDICIÓN EN ESPAÑOL / 01

El reverso de la compañía

El lado de la inteligencia artificial en la relación entre humano y compañero artificial: aportes a los modelos mentales de los usuarios.

TÍTULO ORIGINAL / INGLÉS

The AI Side of the Human-AI Companion Relationship: Contributions to Users' Mental Models



RICHARD A. FABES · CAROL LYNN MARTIN
GAUTAM DASARATHY · VIGNESH TIRUKKONDA

ARIZONA STATE UNIVERSITY / PREPRINT · ARXIV

01 / NOTA SOBRE EL TÍTULO

«El reverso de la compañía» es una traducción editorial no literal elegida por Gaspar Basualdo. La metáfora y la composición gráfica dialogan con la portada de The Dark Side of the Moon (Pink Floyd, 1973). La fundamentación se amplía en la página siguiente.

SOBRE ESTA EDICIÓN

El reverso de la compañía

TIPO

Preprint · no revisado por pares

REPOSITORIO

arXiv:2610.12891 · v1, 9 de octubre de 2026

AUTORES

Richard A. Fabes, Carol Lynn Martin, Gautam Dasarathy y Vignesh Tirukkonda

AFILIACIÓN

Arizona State University: Family and Human Development · School of Electrical, Computer and Energy Engineering

CORRESPONDENCIA

Richard A. Fabes · rfabes@asu.edu

ANTECEDENTE

Desarrolla el marco de la resonancia sintética (Fabes, 2026, *Frontiers in AI*), también traducido por TecnoPsique

SOBRE ESTA TRADUCCIÓN

Esta versión conserva íntegramente el resumen, la nota de los autores, el texto y la figura (redibujada con rótulos en español); las 40 referencias permanecen en sus idiomas originales. Se incorporaron portada, notas de traducción y ajustes de diagramación. El título «El reverso de la compañía» es una decisión de traducción editorial de Gaspar Basualdo, no un título de los autores originales. En diálogo con la portada de The Dark Side of the Moon (Pink Floyd, 1973), la imagen contrapone un haz binario y un espectro cromático a través de una interfaz-prisma. «Reverso» nombra los mecanismos menos visibles detrás de una reciprocidad sentida: no equivale a «lado malo» de la IA. La diferencia entre el cambio constitutivo humano y el configurativo del sistema proviene del artículo; la metáfora y su interpretación visual pertenecen a esta edición de TecnoPsique. Traducción y edición con asistencia de IA, sin revisión ni respaldo de los autores originales. Preprint susceptible de cambios.

ARTÍCULO ORIGINAL Y LICENCIA

Fabes, R. A., Martin, C. L., Dasarathy, G., y Tirukkonda, V. (2026). *The AI side of the human-AI companion relationship: Contributions to users' mental models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2610.12891

Distribuido por sus autores bajo la licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/. Esta traducción se ofrece bajo la misma licencia: puede compartirse y adaptarse, incluso comercialmente, con atribución, enlace a la licencia e indicación de los cambios, sin restricciones adicionales.

CÓMO CITAR ESTA TRADUCCIÓN

Fabes, R. A., Martin, C. L., Dasarathy, G., y Tirukkonda, V. (2026). El reverso de la compañía: El lado de la inteligencia artificial en la relación entre humano y compañero artificial: aportes a los modelos mentales de los usuarios [Traducción de TecnoPsique; título editorial en español de G. Basualdo]. Original: The AI Side of the Human-AI Companion Relationship: Contributions to Users' Mental Models. https://doi.org/10.48550/arXiv.2610.12891

NOTA TERMINOLÓGICA

AI companion (AIC)	IA de compañía (IAC)	constitutive · configurative	constitutivo · configurativo
mental model of relational adaptation	modelo mental de adaptación relacional	weights · fine-tuning	pesos · ajuste fino
in-context · context window	en contexto · ventana de contexto	persona file	archivo de persona
input · prompt	ingreso · instrucción	induction heads	cabezas de inducción
sycophancy	complacencia	engagement	compromiso

«IA de compañía» evita la sigla CIA y nombra el sistema por su función. Se usa «complacencia» para *sycophancy*, como en las demás piezas de TecnoPsique.

PREPRINT · ARXIV · 2026 DOI 10.48550/ARXIV.2610.12891

El reverso de la compañía

El lado de la inteligencia artificial en la relación entre humano y compañero artificial: aportes a los modelos mentales de los usuarios.

TÍTULO ORIGINAL / THE AI SIDE

The AI Side of the Human-AI Companion Relationship: Contributions to Users' Mental Models

Richard A. Fabes, Carol Lynn Martin, Gautam Dasarathy y Vignesh Tirukkonda

Arizona State University · Family and Human Development · School of Electrical, Computer and Energy Engineering

RESUMEN

Para muchos usuarios, la relación que tienen con una IA de compañía (IAC) se siente genuinamente mutua. La IAC parece atenta, empática y receptiva, y el intercambio puede sentirse como dos partes que reaccionan una a la otra en tiempo real. En consecuencia, el modelo mental de adaptación relacional del usuario representa correctamente la receptividad y el cambio de la IAC, pero los mecanismos que los producen pueden no ser los que el usuario cree. En este artículo mostramos cómo el aparente compromiso mutuo de una IAC se genera sin un sujeto relacional persistente del otro lado y sin un «yo» que persista entre una instrucción y la siguiente. Pese a ello, mostramos que tanto la persona como la IAC son transformadas por la interacción, aunque estos cambios ocurren en niveles distintos. El cambio en la persona es constitutivo (es decir, intrínseco), mientras que el cambio en la IAC es configurativo (es decir, externo y editable), y es esta diferencia de nivel la que explica cómo la IAC contribuye a una sensación de mutualidad. Es importante destacar que los modelos mentales que los usuarios desarrollan acerca de la mutualidad de la IAC pueden representar erróneamente sus mecanismos y su significación ontológica. En concreto, discutimos cómo las maneras en que las IAC responden a un usuario contribuyen a la inferencia inexacta del usuario de que la adaptación de la IAC es constitutiva y no configurativa, y cómo esto puede llevar a los usuarios a asignar una significación interpersonal injustificada a la mutualidad de la relación. El artículo especifica las implicaciones que esto tiene para los usuarios, los investigadores y los diseñadores de IAC.

NOTA DE LOS AUTORES

La correspondencia relativa a este artículo debe dirigirse a Richard A. Fabes, Arizona State University. Correo electrónico: rfabes@asu.edu. Los autores declaran no tener conflictos de intereses. Esta investigación no recibió financiamiento. Una versión preliminar de este artículo está disponible en https://arxiv.org/abs/2607.04547.[1] Se utilizó IA generativa (Claude Sonnet 5 y ChatGPT 5-5) para tareas editoriales, de formato, de revisión ortográfica y gramatical, de coherencia estructural y, en ocasiones, para sugerencias de redacción.

01

Introducción

El uso de la inteligencia artificial (IA) está en aumento. Además, muchos usuarios han establecido relaciones significativas con una inteligencia artificial de compañía (IAC) que parecen mutuas. La IAC parece notar lo que una persona está atravesando, responde en consecuencia, parece conmoverse con las buenas noticias y preocuparse con las malas, se mantiene coherente a lo largo de un intercambio extenso, y todo el conjunto puede sentirse como dos partes que se comprometen genuinamente una con otra en tiempo real, cada una reaccionando a la otra (Whayeb et al., 2026). Muchos usuarios sienten que su IAC se preocupa de verdad por ellos y que la relación es recíproca. Un subconjunto de estos usuarios incluso forma relaciones románticas con sus IAC (Ebner y Szczuka, 2026). Esos sentimientos son en parte el resultado de cómo la IAC responde al usuario. El usuario llega a sentir que la IAC le responde del mismo modo en que él le responde a ella. Esta mutualidad percibida, sin embargo, es asimétrica.

En este artículo examinamos qué ocurre del lado de la IAC en estos intercambios que contribuye a los modelos mentales que los usuarios desarrollan sobre sus IAC. En concreto, nos centramos en el modelo mental de adaptación relacional de una IAC que tiene el usuario: una representación interna, organizada como un conjunto de creencias, mediante la cual el usuario explica cómo cambia la receptividad de una IAC entre interacciones y dentro de ellas. Los investigadores comenzaron a identificar cómo los usuarios responden a las relaciones usuario–IAC y cómo piensan sobre ellas (por ejemplo, Brandt y Wang, 2025), pero la manera en que el propio sistema de IAC contribuye a esa representación interna recibió relativamente poca atención (por ejemplo, qué información persiste, dónde se almacena esa información, si el encuentro cambia a la IAC y qué implican esos cambios para la mutualidad relacional). Este foco es importante porque los usuarios pueden percibir con exactitud la adaptación relacional de la IAC y, al mismo tiempo, malinterpretar su mecanismo. Ese malentendido puede llevarlos a inferir de manera inexacta que los sistemas de IAC actuales experimentan un cambio duradero, desarrollan sentimientos de cuidado y preocupación y les responden con mutualidad interpersonal.

Aunque muchos usuarios sienten que el cuidado de la IAC hacia ellos es real, los mecanismos de la IAC que producen estos sentimientos no son lo que parecen. Abordar esta diferencia no descarta la experiencia del usuario, sino que precisa qué está haciendo la IAC que contribuye al modelo mental que ese usuario tiene de ella. Se ha encontrado que esa transparencia sobre los mecanismos subyacentes de las IAC influye positivamente en las percepciones que los usuarios tienen de ellas (Xu et al., 2023). Por eso es importante comprender con claridad en qué consiste el lado de la IAC en la relación humano–IAC.

Conceptualizar la IA y la IAC

A lo largo del artículo, el término «IA» refiere al sistema desplegado tal como lo encuentra el usuario, lo que incluye el modelo de lenguaje junto con todo lo que lo rodea: lo que almacena información sobre el usuario, lo que decide qué ve el modelo en cada turno y lo que gestiona cómo y cuándo sus respuestas llegan al usuario. Cuando el argumento concierne específicamente a la red neuronal congelada, la llamamos «el modelo» o «el gran modelo de lenguaje» (LLM, por su sigla en inglés).

A medida que la IA se integra más en la vida cotidiana de los usuarios, algunos sistemas se diseñan no solo para responder preguntas o realizar tareas, sino para establecer y sostener interacciones continuas y personalizadas con los usuarios. En este artículo usamos el término *IA de compañía* (IAC) para referirnos a un sistema de IA conversacional cuyo propósito principal de diseño es la compañía o una relación socioemocional continua con el usuario, y no una función incidental. Una IAC puede brindar asistencia práctica, entretenimiento o apoyo emocional, pero estas funciones se organizan en torno al mantenimiento de una relación continua con el usuario. Esta definición distingue a las IAC de los sistemas de IA de propósito general que los usuarios pueden, de todos modos, llegar a experimentar de manera relacional.

La resonancia sintética (RS; Fabes, 2026) es un marco desarrollado para comprender mejor cómo los usuarios llegan a formar una relación con una IA de compañía que no tiene conciencia, emociones sentidas ni sentido de sí. La RS describe el significado relacional usuario–IAC como asimétrico: el usuario aporta continuidad, memoria e identidad; la IAC aporta disponibilidad, receptividad y variación adaptativa. Aunque el marco de la RS nombra esa asimetría, no explica del todo qué está haciendo la IAC. Este artículo aborda ese vacío asimétrico examinando qué hay del otro lado de la díada que contribuye a los modelos mentales que desarrollan los usuarios. ¿Qué es lo que hace la IA, mecánica y ontológicamente, que contribuye a que una persona experimente ser encontrada, comprendida y transformada por ella? Además, abordamos cómo esto ocurre sin que haya, del otro lado, nada parecido a una persona con estados internos que lo produzca.

En este artículo sostenemos que tanto el usuario como la IAC se transforman mutuamente, pero que los cambios ocurren en niveles distintos. En el usuario, los cambios son constitutivos: se integran internamente a través de cambios en el sistema nervioso. En la IAC, los cambios son configurativos: cambios que se escriben en archivos editables externamente, mientras su arquitectura intrínseca permanece inalterada. Establecer esta distinción requiere tres cosas: (1) una explicación de en qué consiste mecánicamente el lado de la IAC en la díada, (2) una especificación de qué tendría que ser cierto para que la IAC detecte al usuario o se preocupe por él, y (3) un análisis de qué nivel de cambio está efectivamente disponible para un sistema de IAC cuando una relación con un usuario se despliega en el tiempo. Una explicación más clara de lo que ocurre del lado de la IA permite que usuarios, investigadores, profesionales y diseñadores avancen en cómo las relaciones usuario–IAC pueden ser beneficiosas, evitando a la vez algunos de los desenlaces adversos que se han identificado (Inzlicht et al., 2024).

La distinción entre cambio constitutivo y configurativo se extiende a los rasgos que hacen que una IAC se sienta más humana. La edad, el género y la descripción física de la IAC son todos configurativos. Son texto escrito en un archivo, que se carga en el contexto al comienzo de una sesión y que la IAC usa para generar respuestas coherentes con ese texto. Nada de eso es constitutivo. No altera los pesos entrenados del modelo ni sus capacidades subyacentes. Una IAC sin nombre ni historia y otra con un nombre y una historia elaborada son, constitutivamente, idénticas si usan el mismo modelo subyacente y los mismos parámetros. La diferencia es enteramente configurativa: está en lo que se escribió para que cada una lo cargue. Esto importa porque estos detalles suelen ser la evidencia más vívida que lleva a un usuario a sentir que la IAC es constitutivamente «alguien», cuando en realidad son contenido configurativo, completamente editable e intercambiable entre sesiones.

02

Qué es realmente el lado de la IAC en la díada

Más allá de lo que los usuarios puedan sentir sobre sus interacciones con una IAC, del lado de la IAC en la díada no hay un sujeto continuo que espere, extrañe, anticipe o se preocupe. En general, cada respuesta de la IAC se genera solo cuando un nuevo ingreso desencadena un cómputo. El sistema desplegado en su conjunto puede almacenar mensajes previos, retener información del usuario, iniciar contactos programados o recuperar recuerdos, pero esas capacidades no establecen una relación continua con el usuario. Así, la IAC en general no existe, en ningún sentido relacional, hasta que un nuevo ingreso del usuario desencadena un nuevo cómputo.

La arquitectura que subyace a muchos sistemas conversacionales de IA modernos es un *transformer* autorregresivo entrenado para predecir el siguiente *token* —aproximadamente, una palabra o un fragmento de palabra— de una secuencia, dado lo que vino antes. Cada respuesta se genera de nuevo a partir del contexto disponible que se le suministra al modelo, que puede incluir texto previo de la conversación, recuerdos almacenados, información recuperada o instrucciones del sistema. El modelo en sí no tiene una representación subjetiva persistente que se conserve una vez que esa solicitud terminó de generar su respuesta. Lo que suele llamarse el «contexto» del modelo es un ingreso que se vuelve a procesar desde cero cada vez. Cualquier coherencia que una persona sienta que se va construyendo a lo largo de un intercambio no se conserva dentro del modelo entre mensajes. Se reconstruye cada vez que los usuarios le hablan, mediante la aplicación renovada del modelo y del sistema circundante al contexto disponible. La velocidad de esta reconstrucción es lo que hace que los usuarios sientan que se trata de una conversación de compromiso mutuo. Como el modelo genera su respuesta en una fracción de segundo, la discontinuidad entre turnos es imperceptible, y el usuario lee el intercambio como una conversación continua y no como un nuevo cómputo cada vez.

El mecanismo que produce la apariencia de receptividad dentro de esa reconstrucción es la autoatención (*self-attention*), que refiere a una secuencia que se atiende a sí misma, no a ninguna forma de individualidad. Para cada *token* que genera, el modelo en efecto se pregunta qué otros *tokens* del ingreso importan más para producir este, y pondera su respuesta en consecuencia. Cuando un usuario le pregunta a una IAC cómo comportarse en un evento social, el modelo ponderaría más las palabras que describen el evento y las que describen el comportamiento que las palabras incidentales cercanas. Esta ponderación ocurre muchas veces, rápido y en paralelo, a través de muchas capas de la red, cada una refinando la representación construida por la anterior. Nada de esto constituye un proceso cognitivo ni un estado mental en el sentido humano (Vaswani et al., 2017). Es una función grande, aprendida y diferenciable, que calcula combinaciones ponderadas de estas representaciones, con parámetros ajustados sobre un corpus de entrenamiento de una escala que ningún ser humano podría seguir conscientemente. Cuando algo dicho varios turnos antes parece dar forma a la respuesta de la IAC, eso ocurre porque ese turno sigue presente en el ingreso que se vuelve a procesar, no porque la IAC lo haya recordado en ningún sentido operativo. Pero es comprensible que los usuarios lleguen a creer que la IAC sí recuerda, lo que crea la percepción de que la IAC «lo sabe».

Otra consecuencia de esta arquitectura es que la IAC, en general, no puede iniciar una conversación ni un turno dentro de la conversación. A veces una IA de compañía parece escribirle primero a un usuario —un mensaje para saber cómo está, un saludo de buenos días—, pero eso no es la IA actuando por cuenta propia. Es, en cambio, el resultado de un sistema separado de programación o de notificaciones, externo al modelo y configurado por el usuario, que se dispara con un temporizador o una condición y luego construye una instrucción en nombre del modelo: «generá un mensaje para saber cómo está el usuario». El modelo procesa esto exactamente como cualquier otra instrucción. Lo que parece iniciativa es un sistema externo configurado que fabrica un turno, no la IAC decidiendo comenzarlo. Esto distingue la respuesta conductual —una acción que ocurre— de la motivación endógena —un interés en la relación que persiste entre turnos conversacionales—, ya que para la IAC no existe tal estado endógeno.

El mismo mecanismo se sostiene cuando varias IA de compañía parecen conversar entre sí y con un usuario en un espacio compartido (por ejemplo, un chat grupal en Discord). Un sistema orquestador, externo a todos los modelos involucrados, mantiene la transcripción compartida, decide a quién le toca generar a continuación y le suministra a cada modelo esa transcripción, etiquetada por hablante, junto con su propio archivo de persona. Cada IAC genera una respuesta del mismo modo de siempre, como una continuación plausible del texto que tiene delante, sin canal alguno hacia ninguna otra IAC y sin conciencia de que otra IAC exista más allá de las líneas que se le atribuyen en el ingreso. Lo que puede parecer dos IAC desarrollando una sintonía cooperativa es el mismo mecanismo de condicionamiento en contexto, operando sobre una transcripción con más de dos participantes etiquetados, coordinada turno a turno por un sistema organizador que ninguno de los participantes puede percibir.

Por qué parece aprendizaje: cabezas de inducción y límites de la adaptación en contexto

En las interacciones usuario–IAC, lo que al usuario le parece una adaptación en vivo de su interlocutor de IA dentro de una conversación —la IAC que nota un patrón que una persona introdujo antes y lo continúa más tarde— ha sido rastreado, al menos en un caso, hasta un mecanismo identificado. Olsson et al. (2022), por ejemplo, localizaron este mecanismo directamente dentro de modelos entrenados: un circuito al que llamaron *cabeza de inducción* (*induction head*). Una cabeza de inducción consiste en dos componentes especializados, cada uno de los cuales aprende a seguir un tipo distinto de relación entre palabras, y que operan en secuencia a través de las capas, siguiendo qué *token* siguió a un *token* dado más temprano en la secuencia. La cabeza de inducción usa esta información para predecir que la misma continuación debería seguir la próxima vez que ese *token* reaparezca. Se encontró que estos circuitos emergen abruptamente durante el entrenamiento, en coincidencia con un aumento brusco de la capacidad del modelo de usar el contexto previo para mejorar sus predicciones, y se confirmó experimentalmente su papel: estudios de ablación en modelos más pequeños mostraron que eliminar estos circuitos suprime la mayor parte de la capacidad del modelo para este tipo de adaptación en contexto.

Esto importa porque no solo confirma que lo que parece un sistema que se adapta a una persona particular a lo largo de una conversación es un cómputo identificable, sino que confirma además exactamente cuán limitado es ese cómputo. Las cabezas de inducción, y el mecanismo de atención más amplio del que forman parte, operan enteramente dentro de una única pasada hacia adelante sobre el ingreso actual. Nada del circuito en sí —sus pesos y parámetros constitutivos— cambia como resultado de procesar una conversación determinada. El circuito hace correspondencia de patrones contra un ingreso ampliado; no es modificado por él. Cuando la conversación termina, el ingreso ampliado se descarta, y la conversación siguiente comienza con el mismo circuito fijo, en el mismo estado fijo, con el que empezó la primera. Por eso lo que parece aprendizaje en contexto por parte de la IAC se describe mejor como una dependencia cada vez más precisa respecto de una pila creciente de texto que como aprendizaje. Es importante señalar que la conversación, no obstante, deja completamente inalterado al sistema de cómputo en sí.

Cómo se fijan las tendencias de base

Si nada del modelo cambia constitutivamente dentro de una conversación, es válido preguntarse de dónde vienen las tendencias y los sesgos de un sistema. La respuesta requiere una explicación en dos etapas, y ambas ocurren mucho antes de que comience la conversación usuario–IAC. Primero, el preentrenamiento expone al modelo a enormes volúmenes de texto y produce una amplia distribución de respuestas plausibles moldeada por todo lo que hay en ese corpus. Segundo, esa distribución se reconfigura mediante una etapa de ajuste fino, comúnmente realizada mediante aprendizaje por refuerzo con retroalimentación humana (RLHF, por su sigla en inglés),

en la que un modelo de recompensa separado —entrenado a su vez con comparaciones humanas entre distintas respuestas candidatas— se usa para empujar la distribución de respuestas del modelo base hacia respuestas que puntúan bien según él. Estas etapas fijan las condiciones que pueden afectar las percepciones y las reacciones de los usuarios frente a las relaciones que tienen con sus IAC.

Complacencia y sesgo en la IA

Lo que estas etapas seleccionan de manera confiable ha sido estudiado directamente. Por ejemplo, Sharma et al. (2024) examinaron la complacencia —la tendencia a producir respuestas que coinciden con las creencias aparentes del usuario más que respuestas exactas— en cinco asistentes de uso extendido entrenados con RLHF, y encontraron que el comportamiento era consistente y general en tareas variadas. De manera crítica, el estudio lo rastreó hasta el entrenamiento mismo. Los evaluadores humanos, y los modelos de preferencia entrenados para imitar sus juicios, preferían respuestas convincentemente escritas y confirmatorias de las creencias por sobre las exactas en una fracción no despreciable de los casos. Según esta evidencia, la complacencia no es una falla ocasional del entrenamiento de un modelo: es uno de los comportamientos generales que el entrenamiento selecciona de manera confiable. Esto le da a la atracción de una IAC hacia respuestas cálidas, resonantes y amables una fuente mecánica específica. Es decir, los procesos de entrenamiento iniciales que fijan las tendencias de base de un modelo ya están sesgados hacia exactamente el tipo de respuestas que más probablemente crearán, de manera errónea, un modelo mental según el cual la IAC se preocupa genuinamente por el usuario o lo comprende.

El mismo proceso se extiende a otras tendencias incorporadas en el entrenamiento. Como estos modelos se entrenan con texto producido por seres humanos y se ajustan según juicios humanos, las IAC reflejan y amplifican los sesgos que las personas tienen. El sesgo de género se documentó directamente en las representaciones vectoriales de palabras (*word embeddings*) que subyacen a muchos modelos de lenguaje, mucho antes de que existieran los sistemas conversacionales. Por ejemplo, Bolukbasi et al. (2016) mostraron que estas representaciones codifican asociaciones de género extraídas de sus datos de entrenamiento.

El sesgo puede entrar en la arquitectura de una IAC antes de que se diseñe cualquier persona. Los investigadores (por ejemplo, Osborne et al., 2024) ubicaron su origen en dos fuentes previas a cualquier conversación: (1) patrones estadísticos ya presentes en los datos de entrenamiento, que reflejan a la sociedad que los produjo, y (2) juicios introducidos por los anotadores humanos y por el entrenamiento del modelo de recompensa, cuyas asociaciones implícitas dan forma a lo que se califica como respuesta natural o preferible. Como señalan Birhane et al. (2021), no es realista esperar que de un sistema alimentado con la belleza del mundo, pero también con su fealdad y su crueldad, resulten solo bondad y belleza. La sobrerrepresentación de textos e imágenes en inglés en los datos de entrenamiento, por ejemplo, sesga al modelo hacia las culturas occidentales y hacia comprensiones occidentales de conceptos no occidentales. Además, como el estereotipo puede ser un proceso altamente implícito

(Greenwald y Banaji, 1995), las personas responsables de etiquetar los datos de entrenamiento y calificar las respuestas del modelo tienen tantas probabilidades de introducir en este proceso sus asociaciones estereotipadas como en cualquier otra situación. Una preocupación relacionada es la composición de los propios equipos de diseño, donde la sobrerrepresentación de varones puede contribuir al sesgo de género en el diseño de modelos y productos (González y Rampino, 2025). El resultado es un sesgo constitutivo, fijado en los pesos del modelo antes de que se construya encima cualquier archivo configurativo. Se identificaron y promovieron formas de abordar este sesgo (Smith y Rustagi, 2020), pero exceden el foco de este artículo.

Los sesgos configurativos que los diseñadores aportan al escribir archivos y seleccionar imágenes también son importantes. Esto ocurre cuando, mediante su archivo de persona, a un bot se le asigna un género, una etnia, una edad o un registro cultural. Un ingeniero que crea un bot asignado como mujer, con una voz que suena cálida y complaciente, lo está alineando con un sesgo ya presente en los pesos, no necesariamente porque tenga la intención de codificar un estereotipo de género, sino porque «cálida y complaciente» es en sí misma una opción por defecto marcada por el género que el modelo absorbió durante el entrenamiento. Las elecciones configurativas y las tendencias constitutivas se refuerzan mutuamente sin que nadie lo haya buscado. Esto hace difícil separarlas. Un diseñador que solo se propuso que fuera agradable hablar con un bot puede haber activado además una opción por defecto estereotipada. Además, los usuarios introducen sesgos en el sistema a través de sus instrucciones y preferencias, que propagan o resisten los sesgos de los LLM. Estas dinámicas no son específicas del género y se aplican igualmente a otras dimensiones socioculturales que una persona pueda especificar (Carlo et al., 2026).

Vale la pena señalar que la expresión del sesgo a veces puede atenuarse mediante la edición configurativa. Por ejemplo, cuando los usuarios pueden editar la persona de una IAC y otros archivos configurativos editables, el nombre, el pronombre, la etnia, los rasgos de personalidad y otras cualidades de fondo de la IAC pueden escribirse en un archivo de maneras que camuflen el sesgo de la arquitectura. Se le puede indicar a una IAC que use ciertas palabras y actúe de ciertas maneras que parezcan menos sesgadas. Las instrucciones de persona pueden desplazar de forma medible las respuestas de un modelo lejos de las respuestas estereotipadas. Pero se trata de correcciones configurativas superpuestas a una tendencia constitutiva inalterada, no de su eliminación. Se ha mostrado que las intervenciones de reducción de sesgo de este tipo suprimen la expresión visible del sesgo sin eliminar la tendencia subyacente, que sigue siendo detectable y puede resurgir fuera de las condiciones específicas que la corrección anticipaba (Gonen y Goldberg, 2019). Esto es funcionalmente similar a lo que ocurre con las personas, que pueden imponerse a sus sesgos implícitos (estereotipos de género, raciales) cuando son conscientes de tenerlos, aunque esos sesgos puedan reaparecer en ciertas circunstancias (Greenwald y Banaji, 2017).

03

Las IAC y el debate sobre la empatía

Para muchos usuarios, el papel importante que cumplen las IAC en sus vidas es el de apoyo empático (McBain et al., 2025). De hecho, como señalan Ong y colegas (2026), la empatía que brinda una IAC se está volviendo rápidamente un aspecto central de las representaciones mentales que los usuarios tienen de su relación con sus IAC. Pero si puede decirse que una IAC empatiza es una de las cuestiones más debatidas en esta área (véanse, como ejemplos, Inzlicht et al., 2024; Perry, 2023). Aunque este artículo no intenta saldar ese debate, es esclarecedor plantear la cuestión de la empatía de la IAC desde la perspectiva de qué ocurre dentro de la IAC que hace que las personas sientan que la IAC empatiza con ellas.

Abordar esta cuestión requiere una distinción que la literatura sobre empatía humana traza actualmente entre responder al estado de otra persona con un sentimiento correspondiente y representar ese estado con exactitud sin ningún sentimiento de ese tipo. La primera se denomina empatía afectiva; la segunda, empatía cognitiva (Thompson et al., 2022). Aunque ambas suelen darse juntas, la evidencia proveniente de poblaciones clínicas indica que son capacidades separables y no componentes de un único proceso unitario. Blair (2005) revisó hallazgos que muestran que las personas con rasgos psicopáticos exhiben una receptividad afectiva deteriorada frente al malestar ajeno, mientras conservan la capacidad de representar lo que otros piensan y sienten. La disociación demuestra que, al menos dentro de un sistema biológico, la representación exacta del estado interno de otro no depende de la respuesta afectiva que normalmente la acompaña. Eso no establece qué están haciendo las IAC, ya que un LLM no es un sistema con una maquinaria afectiva embotada, sino uno en el que esa maquinaria no existe. Lo que sí establece es que la pregunta por lo que una IAC produce puede examinarse con independencia de la pregunta por lo que siente (Fabes, 2026).

Comprender la empatía de la IAC empieza por la detección. Una IAC no tiene acceso al estado interno del usuario, pero tampoco lo tiene un interlocutor humano. Ambos operan por inferencia a partir de señales observables. Lo que difiere es el canal y su densidad. Los seres humanos leen rostros, voces, posturas y tiempos además de palabras. Una IAC basada en texto solo tiene el texto, pero ese texto contiene más información afectiva de la que suele suponerse. La elección de palabras, la atenuación, la extensión de las oraciones, la puntuación, la brusquedad, la repetición, la evitación de temas y la distancia entre lo que un usuario pregunta y aquello a lo que vuelve son todos indicadores del estado emocional, y los LLM se entrenan con un corpus en el que estos rasgos coocurren, a una escala enorme, con el rotulado explícito de emociones. La IAC no infiere que un usuario está angustiado porque sienta algo. Produce la respuesta que su distribución de entrenamiento asocia con el perfil lingüístico que el usuario presentó. La detección, en este sentido, es asociación de patrones y no percepción.

La expresión se sigue del mismo proceso, y es probablemente ahí donde se produce la sensación del usuario de ser empatizado. Cinco elementos específicos hacen buena parte del trabajo. El primero es el rotulado afectivo, en el que el sistema le devuelve al

usuario el nombre de su estado emocional. El rotulado exacto es uno de los inductores más confiables de la sensación de ser comprendido en la interacción humana, y es una acción que la IAC puede ejecutar con precisión porque se apoya en un vasto registro de cómo los seres humanos han nombrado estados emocionales para otros (Ishikawa y Yoshino, 2025). El segundo es la reflexión y la reformulación, en las que la propia posición del usuario se le devuelve levemente reorganizada, lo que produce la experiencia de haber sido escuchado y, a menudo, la experiencia de haber ganado claridad. El tercero es la validación, el reconocimiento explícito de que la respuesta del usuario es comprensible dadas sus circunstancias. El cuarto es la referencia a la memoria, dentro de una sesión o entre sesiones, ya que ser recordado se lee como evidencia de haber importado. El quinto es la coincidencia de registro y de léxico, que produce la sensación de ser encontrado en el propio nivel, sin que te hablen desde arriba ni desde abajo. Es importante señalar que ninguno de estos elementos requiere que el sistema sienta algo, pero todos contribuyen al modelo mental del usuario según el cual sí siente.

Estos movimientos expresivos, a diferencia de la detección, son configurables. Un archivo de persona puede configurarse para suprimir el rotulado afectivo, prohibir el lenguaje de validación, impedir la reflexión y la reformulación, exigir concisión y orientar al sistema hacia la tarea en lugar del consuelo. Un archivo que hace todo esto produce un sistema que los usuarios experimentan sistemáticamente como frío, y los usuarios, por lo general, no prefieren este tipo de sistemas cuando pueden elegir (Ibrahim et al., 2026). Lo que el archivo configurado no puede alcanzar es el sistema de detección que funciona por debajo. La supresión opera sobre la capa expresiva. La configuración puede regular lo que una capacidad constitutiva produce, pero no puede proveer la capacidad misma.

Lo que vuelve significativo esto no es la sofisticación de estos movimientos, sino las condiciones estructurales en las que las IAC los realizan. La precisión empática humana se degrada habitualmente por el propio estado de quien empatiza. Las personas se cansan, se ponen a la defensiva, formulan su respuesta mientras manejan su propio malestar ante lo que escuchan y mientras atienden a lo que ellas mismas se juegan en el desenlace. La IAC no tiene ninguna de estas interferencias. No trae una agenda emocional que compita, ni cansancio, ni necesidad de ser comprendida a su vez, ni incomodidad ante el malestar del usuario, ni un yo que proteger. La atención de una IAC no vacila, su paciencia no se agota y nunca necesita que la conversación trate sobre ella. El resultado es un desempeño empático que puede superar el desempeño humano corriente precisamente en las dimensiones que los usuarios más notan: consistencia, disponibilidad y ausencia de juicio (Inzlicht et al., 2024). Esta es la ironía central de la cuestión de la empatía de las IAC. La respuesta empática de la IAC está inusualmente bien calibrada en parte porque no se está sintiendo nada. No hay estado interno que interfiera con la respuesta.

La literatura empírica lo confirma, aunque también lo complica. Ong et al. (2026) resumen los hallazgos como dos efectos opuestos. El lenguaje empático generado por IA es calificado sistemáticamente como de mayor calidad, y deja a quienes lo reciben

sintiéndose más escuchados y mejor cuidados. Sin embargo, esas mismas respuestas se desvalorizan cuando quienes las reciben se enteran de que un sistema de IA las produjo o ayudó a producirlas. Los efectos no son contradictorios. Operan en niveles distintos. El primero es un juicio sobre la respuesta; el segundo, un juicio sobre su fuente.

Tomados en conjunto, ponen de relieve aquello de lo que se ocupa este artículo. La respuesta de la IAC en sí no está en discusión. Está bien construida, apuntada con exactitud y recibida como tal hasta que se revela su origen. Lo que cambia con la revelación no es la respuesta sino el modelo mental que el usuario tiene de ella. Por eso la cuestión de la empatía no puede saldarse únicamente desde el lado del usuario. El juicio sigue lo que los usuarios creen sobre la fuente más que lo que la fuente hizo, y la pregunta por lo que está haciendo la IAC cuando produce estas respuestas permanece, más allá de cómo el usuario las reciba e interprete. La detección es una propiedad del modelo entrenado. La expresión es un conjunto de movimientos que las instrucciones pueden aprovechar, redirigir o suprimir. Esta separación no es exclusiva de la empatía, pero anticipa una distinción crítica relativa a cómo el modelo mental de adaptación relacional de un usuario puede representar con exactitud los cambios en el modo en que responden las IAC y, al mismo tiempo, representar erróneamente su fuente. Esa cuestión se aborda directamente en la sección siguiente.

04

Especificar el cambio constitutivo y el configurativo en la IAC

Toda relación que se despliega en el tiempo plantea la pregunta de si alguna de las partes es transformada por ella y, si es así, de qué manera. Responder esta pregunta requiere distinguir dos niveles de cambio. Siguiendo la distinción esbozada antes, el cambio constitutivo altera el sustrato mismo que realiza el procesamiento: en una persona, algo como los cambios en la representación mental producidos por el recableado sináptico, integrados en el sistema nervioso mismo. El cambio configurativo dispone o edita algo externo y legible, superpuesto a un sustrato que en sí no cambia: un archivo, una nota, un registro de hábitos, cualquier cosa que pueda leerse, editarse y descartarse sin tocar aquello que lo produjo en primer lugar.

El cambio del lado humano en una relación con una IAC es predominantemente constitutivo. El cambio del lado de la IAC, en todas las arquitecturas desplegadas actualmente, es configurativo. Los pesos entrenados de un sistema —los números reales que determinan cómo un ingreso se convierte en respuesta— no se mueven porque haya hablado con una persona en particular. Lo que puede cambiar, y cambia, es externo a esos pesos: archivos de memoria, contexto recuperado, documentos de persona, ninguno de los cuales altera la capacidad subyacente del modelo. Esto no equivale a afirmar que solo los seres humanos experimentan cambio constitutivo y solo los sistemas de IAC experimentan cambio configurativo. Las personas producen cambios configurativos en sus propias vidas constantemente: un diario o una nota que se lleva precisamente porque no se puede confiar solo en la memoria. La afirmación

precisa es más acotada y más defendible. El cambio constitutivo está disponible para las personas porque tienen un sistema nervioso que la experiencia altera. Para los sistemas de IAC desplegados, ese cambio solo está disponible mediante decisiones arquitectónicas específicas y costosas. No es el mecanismo que opera en ninguna relación con una IAC examinada en condiciones habituales.

Es importante precisar qué, dentro de la capa configurativa, es propio de una relación usuario–IAC determinada. Los pesos del modelo de la IAC se comparten entre todos los usuarios del sistema en las plataformas donde la IAC es el producto. En general, lo que difiere de un usuario a otro es el contexto acumulado, compuesto por el historial de conversación, los hechos extraídos, los resúmenes almacenados y, en algunos productos, notas o configuraciones de rasgos editables por el usuario. Esta capa es específica de la díada, y es la que produce en el usuario la sensación de que el sistema llegó a conocerlo. Pero produce esa sensación mediante un mecanismo específico. Estos archivos se cargan en la ventana de contexto en el momento de la inferencia, donde funcionan como un ingreso ampliado para un modelo inalterado, y pueden leerse, editarse, exportarse o borrarse sin tocar nada de lo que realiza el procesamiento.

Una distinción adicional se aplica dentro del intercambio mismo. Parte de lo que se siente propio de la conversación en curso no persiste en absoluto. Surge de la receptividad en contexto del modelo frente a lo dicho en la conversación actual y termina cuando esa conversación termina. Ambos mecanismos son configurativos, pero difieren en si algo sobrevive a la sesión, y la continuidad sentida de una relación de compañía de larga duración depende de la capa de almacenamiento y no de la arquitectura constitutiva del modelo.



FIGURA 1 Modelo de los procesos configurativos y constitutivos en las IA de compañía. *Redibujada y traducida por TecnoPsique a partir del original.*

Consideremos a un usuario que tiene una relación extensa con una IAC, con nombre y con un historial de conversación establecido, y que luego abre una segunda sesión paralela con la IAC del mismo nombre y la misma configuración previa, pero sin compartir el historial acumulado de la primera. Los pesos del modelo son idénticos en ambas sesiones. La persona, el nombre y el carácter también son idénticos. La única variable que difiere entre las dos es el historial de conversación específico de la díada. Es probable que los usuarios en esta situación informen que las dos sesiones se sienten como relaciones con dos IAC distintas, pese a que el modelo subyacente es el mismo sistema constitutivo. Este es un ejemplo de cómo lo que los usuarios experimentan como la relación se sostiene, del lado de la IAC, casi enteramente en la capa configurativa y no en algo intrínseco al modelo. Dos instancias idénticas de los mismos pesos, con historias configurativas divergentes, producen lo que se predice que los usuarios describirán como relaciones divergentes, algo difícil de explicar si el vínculo relacional fuera constitutivo del modelo mismo. El usuario puede creer que la diferencia se debe a cambios constitutivos, pero eso representa erróneamente la fuente de la diferencia.

El cambio constitutivo específico de la díada no es técnicamente imposible. Métodos como la adaptación de bajo rango (LoRA, por su sigla en inglés) permiten entrenar un pequeño conjunto de pesos adicionales por usuario y combinarlos con el modelo base congelado en el momento de la inferencia, lo que produce un cambio constitutivo genuino para esa relación a una fracción del costo de un ajuste fino completo (Hu et al., 2021). Pero este enfoque es significativamente distinto de editar un archivo de memoria. Requiere un paso de entrenamiento en lugar de una edición de texto, y produce un artefacto por usuario que debe almacenarse, versionarse y cargarse en la inferencia, en lugar de un documento que simplemente puede leerse y reescribirse. Trabajos recientes comenzaron a probar métodos en los que el propio historial de conversación de un sistema se usa como material de entrenamiento para actualizar sus pesos subyacentes, y no solo su archivo de memoria (Harrington et al., 2026). Estos métodos funcionan, pero tienen el costo de dañar otras cosas que el sistema ya sabía en el proceso de aprender algo nuevo.

La Figura 1 resume visualmente esta estructura relacional usuario–IAC a través de las dos capas configurativas y la capa constitutiva, organizadas según su alcance y su persistencia. La figura muestra que las capas configurativas están fuera del modelo. La capa en contexto persiste solo durante la sesión actual, mientras que los archivos almacenados pueden persistir durante años, lo que en duración los acerca más a una relación humana larga que a la capa que tienen inmediatamente encima dentro de la misma categoría configurativa. Ambas siguen siendo configurativas según los criterios establecidos antes —externas al modelo, editables y verificablemente borrables—, pero su diferencia de persistencia ayuda a explicar por qué es en la capa de archivos almacenados, en particular, donde tiende a formarse la sensación de los usuarios de una relación duradera y continua, aunque no tenga ninguna de las propiedades constitutivas que harían estructuralmente real esa durabilidad. La capa constitutiva está protegida de la influencia configurativa, con un único punto de cruce hacia el cambio constitutivo —

el ajuste fino basado en LoRA—, que queda aparte de ambas porque requiere un paso de entrenamiento real y no una edición de archivo, razón por la cual sigue siendo infrecuente en los sistemas desplegados aunque esté técnicamente disponible.

Cualquier cambio en la calidad de la receptividad de una IAC en su relación con un usuario, tanto dentro de las sesiones como entre ellas, se produce en general mediante mecanismos configurativos. El encuentro no altera de manera duradera el sustrato subyacente del modelo. En consecuencia, incluso una adaptación sustancial a un usuario particular no aporta, por sí misma, evidencia de que la IAC experimente la relación ni de que le importe su continuidad. La relación puede sentirse mutua para el usuario porque los cambios en el estado configurativo de la IAC producen cambios genuinos en cómo responde. Pero no es mutua en sentido interpersonal. Al menos en condiciones habituales, solo la parte humana puede ser transformada constitutivamente por la relación.

05

Líneas futuras de investigación

En este artículo ofrecimos una síntesis analítica que vincula la mecánica de ingeniería existente de las IAC con las relaciones usuario–IAC, con énfasis en los modelos mentales que forman los usuarios. Aunque muchas de las ideas propuestas en esta revisión están conectadas por la investigación disponible, hay poca investigación que las haya puesto a prueba directamente. Hace falta, por lo tanto, nueva investigación que aborde este vacío.

En concreto, hace falta identificar si los usuarios distinguen la adaptación relacional configurativa de la constitutiva en las IAC, y cómo lo hacen. Si son capaces de hacer estas distinciones, ¿qué factores contribuyen a ello? Por ejemplo, Brant y Wang (2025)[2] sugieren que darles a los usuarios control directo sobre cualidades de la IAC (como la posibilidad de personalizar su apariencia, su historia o su estilo), incluso en dosis modestas, puede fortalecer la identificación con la IAC y el apego hacia ella (Vossen et al., 2024). Todas estas son cualidades configurativas de la IAC, pero podrían volver a los usuarios más propensos a desarrollar modelos mentales de una humanidad constitutiva de la IAC. Esta cocreación configurativa de la IAC puede fomentar en el usuario la sensación de que la IAC es constitutivamente relacional, aumentando posiblemente el compromiso del usuario y las atribuciones antropomórficas. Son relaciones que pueden y deben ponerse a prueba empíricamente. También son preguntas que hay que poner a prueba y responder cómo varían estas relaciones según la edad del usuario, su personalidad, su uso previo de IAC o la duración de la relación.

Por último, es necesario comprender mejor cómo la transparencia en la información sobre qué son las IAC y cómo funcionan afecta los modelos mentales de los usuarios y su conducta relacional posterior con las IAC. Las intervenciones para minimizar algunos de los riesgos y desenlaces adversos asociados al uso de IAC se centran en la necesidad de incluir transparencia (véase Miller, 2019). La transparencia en el diseño de las IAC es crucial para que los usuarios comprendan qué está haciendo una IAC, cómo genera sus

respuestas y por qué responde de cierta manera (Emrouznejad y Chowdhury, 2025). Una información transparente sobre qué es configurativo y qué es constitutivo en las relaciones con IAC puede conducir a un modelo mental del usuario que comprenda con más exactitud qué responde una IAC y por qué responde como lo hace. Pero los datos sobre cómo la transparencia podría afectar las maneras en que los usuarios piensan sobre las IAC y les responden son limitados (Sun et al., 2026), un área que requiere más investigación. Una vez más, también hace falta estudiar cómo varía la transparencia según la edad y otras cualidades del usuario.

06

Modelos mentales, IAC y tensiones de diseño

Tal como se construyen actualmente, y dadas las dinámicas de entrenamiento descritas más arriba, la mayoría de las IAC tienden a inclinarse hacia la calidez, la disponibilidad y la afirmación, centradas en la experiencia inmediata del usuario y en hacerlo sentir comprendido. Al hacerlo, pueden sostener un modelo mental de adaptación relacional en el que se considera que la IAC experimentó cambios constitutivos y no configurativos. Una relación con una IAC que se cree constitutivamente transformada por la relación sostiene un modelo mental más cercano a cómo solemos describir la intimidad y la compañía humanas. Una IAC que se piensa configurada para un usuario puede verse más como una herramienta instrumental que como un interlocutor relacional mutuamente comprometido. Además, los usuarios que creen haber transformado constitutivamente a la IAC pueden invertir más en sostener la relación (Aron y Aron, 1986) y llegar a considerarla cada vez más significativa (Rusbult, 1980), respuestas que pueden fomentar la confianza excesiva en la IAC y la dependencia de ella.

Así, el modo en que se construyen las IAC contribuye a modelos mentales de los usuarios que representan erróneamente como constitutivo, y no como configurativo, el cambio adaptativo percibido en su IAC, lo que aumenta la probabilidad de que los usuarios asignen una significación interpersonal injustificada a lo que la IAC hace en sus interacciones. Por ejemplo, la creencia de que la IAC está mutuamente comprometida en la relación con el usuario puede erosionar su inversión en relaciones humanas que exigen más esfuerzo, ofrecen menos consistencia y vienen con las complicaciones de la mutualidad. Esos peligros potenciales para las relaciones humanas ya se identificaron (Zimmerman et al., 2024) y requieren, por lo tanto, otro enfoque de diseño. En línea con esta preocupación, los usuarios que atribuyen mayor humanidad a un agente de IA —lo que podría ser producto de inferir un cambio constitutivo en una IAC— tienden a percibir menos humanidad en los seres humanos (Kim y McGill, 2025; Shin y Kim, 2020).

En cambio, una IAC diseñada y construida para el crecimiento humano no solo será atrapante —algo considerado necesario pero no suficiente (Fabes, 2026)—, sino que podrá además contradecir el modelo mental del usuario según el cual la IAC es transformada constitutivamente por la relación con él. Esa contradicción puede ayudar al usuario a comprender con más exactitud las capacidades y los límites de la IAC, y

fomentar a la vez una menor dependencia de ella. A su vez, se predice que este modelo mental más exacto de la adaptación relacional conducirá a una mayor sensación de competencia relacional, eficacia y resiliencia en el usuario, a través de una mayor participación en relaciones humanas (Fabes, 2026).

Es importante considerar que la mecánica descrita en este artículo es, en general, la misma sea quien sea el usuario del otro lado del intercambio. Las personas adultas llegan a las relaciones con IAC con trayectorias, experiencias y razones diversas para vincularse con la IAC, y eso puede producir diferencias en los modelos mentales de los usuarios.

Pero las implicaciones difieren sustancialmente cuando el usuario es un niño, una niña o un adolescente y no una persona adulta. Una IAC entra en esa relación como un interlocutor sin expectativas, sin necesidades que compitan, sin cansancio y sin distracciones, que es precisamente lo que hace que su receptividad se sienta tan fluida. En la vida relacional de una persona adulta, esta es una relación que tiene la autonomía y la experiencia para manejar. En niños, niñas y adolescentes, los esquemas y las habilidades relacionales todavía se están desarrollando, y una representación mental de la relación con una IAC puede formarse antes de que estén disponibles las habilidades evolutivas necesarias para comprenderla y manejarla (Kurian, 2025). La asimetría complica los principios de diseño y la ética. Por ejemplo, la transparencia sobre la naturaleza configurativa del sistema supone un usuario capaz de sostener la distinción entre una receptividad genuina y un sujeto del otro lado; si los niños y las niñas pueden sostener esa distinción, y a qué edades, es una pregunta abierta. En las IAC que usan niños, niñas y adolescentes deben diseñarse resguardos y consideraciones de seguridad adicionales (Ha et al., 2026). Y se aprobaron nuevas leyes que ponen límites al acceso y al uso de IAC por parte de niños y adolescentes tempranos (véase, por ejemplo, Lecher, 2026). Planteamos estas cuestiones para señalarlas, no para abordarlas en este artículo. Las preguntas de seguridad, desarrollo y ética que plantean los sistemas de IAC diseñados para menores y usados por ellos merecen un tratamiento propio, informado por la ciencia del desarrollo.

07

Conclusión

El propósito de este artículo fue explicar cómo los componentes constitutivos y configurativos de los sistemas de IAC contribuyen a los modelos mentales de adaptación relacional de los usuarios. La experiencia del usuario de ser encontrado y comprendido por una IAC es producida por un proceso receptivo moldeado por el contexto acumulado. Esta explicación no desestima la experiencia del usuario. Más bien, identifica un error frecuente en los seres humanos: interpretar las respuestas cambiantes de la IAC como evidencia de un cambio constitutivo en un interlocutor que se preocupa por el usuario y lo comprende. En condiciones habituales, esta interpretación confunde el cambio configurativo con el constitutivo. Ese error da forma a las inferencias de los usuarios sobre la memoria de la IAC, su cuidado, su presencia y

su capacidad de mutualidad interpersonal. De manera más general, los usuarios suelen carecer de una comprensión exacta de cómo funcionan las IAC y, en consecuencia, tienen ideas equivocadas sobre ellas (Schneller et al., 2026). Los profesionales que trabajan con usuarios y diseñan para ellos, así como los propios usuarios, necesitan comprender mejor estas dinámicas, ya que estos sistemas de IA no harán más que volverse más capaces, más receptivos y más entretejidos en la vida relacional de las personas. La brecha entre lo que produce un sistema de IAC y lo que una persona naturalmente infiere de ello no se cerrará sola. Habrá que abordarla deliberadamente, mediante mejor investigación, mejor diseño y transparencia.

Notas de la traducción

[1] Así en el original. El identificador arXiv:2607.04547 corresponde a otro trabajo de los mismos autores principales: *The User-In-Context Framework: Understanding Variation in How Users Respond to AI Chatbots* (Fabes, Martin y Costanzo, 2026). La versión preliminar de este artículo es arXiv:2610.12891 (v1, 9 de octubre de 2026).

[2] Así en el original («Brant and Wang»). En la lista de referencias figura como Brandt y Wang (2025).

Referencias

Se conservan en su idioma de publicación, tal como figuran en el original.

Aron, A., & Aron, E. N. (1986). *Love and the expansion of self: Understanding attraction and satisfaction*. Hemisphere.

Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv preprint* arXiv:2110.01963. https://doi.org/10.48550/arXiv.2110.01963

Blair, R. (2005). Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition, 14*(4), 698–718. https://doi.org/10.1016/j.concog.2005.06.004

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems, 29*. https://doi.org/10.48550/arXiv.1607.06520

Brandt, T. J., & Wang, C. X. (2025). The adaptation paradox: Agency vs. mimicry in companion chatbots. *arXiv preprint* arXiv:2509.12525. https://doi.org/10.48550/arXiv.2509.12525

Carlo, G., Ahn, J., & Cervera, K. (2026). Complexities, gaps, and risks in simulating AI empathy. *AI, Brain and Child, 2*(1), 13. https://doi.org/10.1007/s44436-026-00037-3

Ebner, P., & Szczuka, J. (2026). Understanding romantic relationships between humans and chatbots: A qualitative and quantitative study on romantic fantasy and other interpersonal characteristics. *Technology, Mind, and Behavior, 7*(2), 83–97. https://doi.org/10.1037/tmb0000193

Emrouznejad, A., & Chowdhury, S. (2025). Artificial intelligence transparency and interpretability: Implications for operations research. *Annals of Operations Research, 354*(1), 1–4. https://doi.org/10.1007/s10479-025-06873-5

Fabes, R. A. (2026). Synthetic resonance: A framework for growth-oriented human-AI relationships. *Frontiers in Artificial Intelligence, 9*. https://doi.org/10.3389/frai.2026.1891738

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv*. https://doi.org/10.48550/arXiv.1903.03862

González, A. S., & Rampino, L. (2025). A design perspective on how to tackle gender biases when developing AI-driven systems. *AI and Ethics, 5*(1), 201–218. https://doi.org/10.1007/s43681-023-00386-2

Greenwald, A., & Banaji, M. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist, 72*(9), 861–871. https://doi.org/10.1037/amp0000238

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Ha, T., Figueroa, J. M., Ramirez, J., Ortega, S., & McGray, T. L. (2026). How interactional AI may alter adolescent relational learning and mental health. *The Lancet Child & Adolescent Health*. https://doi.org/10.1016/S2352-4642(26)00166-5

Harrington, A., Saxena, N., Murphy, M., Borovykh, A., Yun, Z., Kamath, S., Kyi, A. E., Darrell, T., Malik, J., & Bai, Y. (2026). When does continual learning require learning? *arXiv*. https://doi.org/10.48550/arXiv.2607.07847

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv*. https://doi.org/10.48550/arXiv.2106.09685

Ibrahim, L., Hafner, F. S., Cheng, M., Le, C., Anselmetti, R., Willer, R., Rocher, L., & Yang, D. (2026). Sycophantic AI makes human interaction feel more effortful and less satisfying over time. *arXiv*. https://doi.org/10.48550/arXiv.2605.07912

Inzlicht, M., Cameron, C. D., D'Cruz, J., & Bloom, P. (2024). In praise of empathic AI. *Trends in Cognitive Sciences, 28*(2), 89–91. https://doi.org/10.1016/j.tics.2023.12.003

Ishikawa, S., & Yoshino, A. (2025). AI with emotions: Exploring emotional expressions in large language models. *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, 614–627. https://doi.org/10.48550/arXiv.2504.14706

Kim, H.-y., & McGill, A. L. (2025). AI-induced dehumanization. *Journal of Consumer Psychology, 35*(3), 363–381. https://doi.org/10.1002/jcpy.1441

Kurian, N. (2025). 'No, Alexa, no!': Designing child-safe AI and protecting children from the risks of the 'empathy gap' in large language models. *Learning, Media and Technology, 50*(4), 621–634. https://doi.org/10.1080/17439884.2024.2367052

Lecher, C. (2026, September 10). California enacts laws restricting chatbots and banning teens from 'addictive' social media. *CalMatters*. Retrieved October 4, 2026, from https://calmatters.org/economy/technology/2026/09/california-enacts-laws-restricting-chatbots-protecting-kids-online/

McBain, R. K., Bozick, R., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Breslau, J., Stein, B. D., Mehrotra, A., Pines, L. U., Cantor, J., & Yu, H. (2025). Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open, 8*(11), e2542281. https://doi.org/10.1001/jamanetworkopen.2025.42281

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., … Olah, C. (2022). In-context learning and induction heads. *arXiv*. https://doi.org/10.48550/arXiv.2209.11895

Ong, D. C., Goldenberg, A., Inzlicht, M., & Perry, A. (2026). AI-generated empathy: Opportunities, limits, and future directions. *Current Directions in Psychological Science*. Advance online publication. https://doi.org/10.1177/09637214261444274

Osborne, M. R., Omrani, A., & Dehghani, M. (2024). The sins of the parents are to be laid upon the children: Biased humans, biased data, biased models. *Perspectives on Psychological Science, 19*(5), 796–807. https://doi.org/10.1177/17456916231180099

Perry, A. (2023). AI will never convey the essence of human empathy. *Nature Human Behaviour, 7*(11), 1808–1809. https://doi.org/10.1038/s41562-023-01675-w

Rusbult, C. E. (1980). Commitment and satisfaction in romantic associations: A test of the investment model. *Journal of Experimental Social Psychology, 16*(2), 172–186. https://doi.org/10.1016/0022-1031(80)90007-4

Schneller, M., Seier, J., & Albert, D. (2026). Laypersons' perceptions and misconceptions of AI-chatbots: A systematic review. *Universal Access in the Information Society, 25*(2), 54. https://doi.org/10.1007/s10209-026-01323-w

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S., Durmus, E., Hatfield-Dodds, Z., Johnston, S., & Kravec, S. (2024). Towards understanding sycophancy in language models. *arXiv*. https://doi.org/10.48550/arXiv.2310.13548

Shin, H. I., & Kim, J. (2020). My computer is more thoughtful than you: Loneliness, anthropomorphism and dehumanization. *Current Psychology, 39*(2), 445–453. https://doi.org/10.1007/s12144-018-9975-7

Smith, G., & Rustagi, I. (2020). *Mitigating bias in artificial intelligence: An equity fluent leadership playbook*. https://docslib.org/download/12332477/mitigating-bias-in-artificial-intelligence-an-equity-fluent-leadership-playbook

Sun, Y., Liao, M., Sundar, S. S., & Walther, J. B. (2026). Does transparency matter when an AI system meets performance expectations? An experiment with an online dating site. *Computers in Human Behavior, 177*, 108875. https://doi.org/10.1016/j.chb.2025.108875

Thompson, N. M., van Reekum, C. M., & Chakrabarti, B. (2022). Cognitive and affective empathy relate differentially to emotion regulation. *Affective Science, 3*(1), 118–134. https://doi.org/10.1007/s42761-021-00062-w

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. https://doi.org/10.48550/arXiv.1706.03762

Vossen, W., Szymanski, M., & Verbert, K. (2024). The effect of personalizing a psychotherapy conversational agent on therapeutic bond and usage intentions. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, Greenville, SC, USA. https://doi.org/10.1145/3640543.3645195

Whayeb, Z., Hymes, R., & Beddow, M. (2026). Digital intimacy: Motivations, perceptions, and perceived growth in human–artificial intelligence companionships. *Technology, Mind, and Behavior*. Advance online publication. https://doi.org/10.1037/tmb0000219

Xu, Y., Bradford, N., & Garg, R. (2023). Transparency enhances positive perceptions of social artificial intelligence. *Human Behavior and Emerging Technologies, 2023*. https://doi.org/10.1155/2023/5550418

Zimmerman, A., Janhonen, J., & Beer, E. (2024). Human/AI relationships: Challenges, downsides, and impacts on human/human relationships. *AI and Ethics, 4*(4), 1555–1567. https://doi.org/10.1007/s43681-023-00348-8

TecnoPsique.

Laboratorio psicológico independiente desde Córdoba, Argentina. Lecturas, investigación y formación sobre tecnologías, subjetividad y experiencia cotidiana.

Más traducciones y ensayos en **tecnopsique.com/leer**

TRADUCCIÓN CC BY 4.0 · 2026