Artículo de reflexión

Trilogía
CIENCIA TECNOLOGÍA SOCIEDAD

¿Inferencias en máquinas? Alucinaciones inferenciales de los grandes modelos de lenguaje desde una perspectiva de los sentimientos metacognitivos*

Inferences in Machines? Inferential Hallucinations in Large Language Models from the Perspective of Metacognitive Feelings

Sebastián Mejía-Rendón[1] Mariana Olezza[2]

Resumen

Los grandes modelos de lenguaje han ganado recientemente cierta popularidad debido a su habilidad para hacer inferencias a partir de la información proporcionada por los usuarios. Estas redes neuronales parecen inferir lógicamente, sobre una base probabilística, emulando reglas lógicas. Tradicionalmente, las inferencias han sido entendidas como procesos de pensamiento que van desde las premisas hasta las conclusiones. No obstante, se ha debatido sobre la naturaleza de la inferencia y la posibilidad de extenderla por fuera del ámbito humano: ¿es comparable la capacidad inferencial de los grandes modelos de lenguaje con la de los seres humanos? En este artículo se propone que la diferencia entre la inferencia humana y la de los grandes modelos de lenguaje radica en la dimensión fenomenológica o afectiva del razonamiento. En particular, la ausencia de sentimientos metacognitivos que guían, monitorean y evalúan los procesos de pensamiento; por ejemplo, razonar o inferir. Aunque los grandes modelos de lenguaje simulan inferencias novedosas, su tendencia a producir alucinaciones inferenciales demuestra la ausencia de esa dimensión autorreguladora que caracteriza al razonamiento humano. En contraste con la afectividad presente en los humanos, las fallas inferenciales de los grandes modelos de lenguaje se podrían entender a partir de tres aspectos: la maldición de la reversión, las alucinaciones inferenciales por estrategias de decodificación y el exceso de confianza. Este artículo concluye que unas de las causas de las alucinaciones inferenciales son la falta de sentimientos metacognitivos como el sentimiento de corrección, la falta de balance por sentimiento de fiabilidad y la ausencia de la duda.

Palabras clave: alucinaciones inferenciales, grandes modelos de lenguaje, sentimientos metacognitivos.

Cómo referenciar

Mejía-Rendón, S., & Olezza, M. (2026). ¿Inferencias en máquinas? Alucinaciones inferenciales de los grandes modelos de lenguaje desde una perspectiva de los sentimientos metacognitivos. *Trilogía Ciencia Tecnología Sociedad, 18*(38), art. e3702. https://doi.org/10.22430/21457778.3702

* Los autores de este artículo recibieron apoyo económico a través de becas doctorales del Consejo Nacional de Investigaciones Científicas y Técnicas (Conicet).

[1] Universidad Nacional de Córdoba, Córdoba, Argentina, joan.mejia@mi.unc.edu.ar
[2] Universidad de Buenos Aires, Buenos Aires, Argentina, olezza.mur@gmail.com

Recibido: 16 de octubre de 2025 | **Aceptado:** 9 de febrero de 2026

CC BY NC SA OPEN ACCESS

Abstract

Large Language Models (LLMs) have recently gained considerable attention due to their ability to draw inferences from information provided by users. These neural networks appear to generate logical inferences on a probabilistic basis, often emulating patterns that resemble classical rules of inference. Traditionally, inference has been understood as a cognitive process through which conclusions are derived from premises. However, the nature of inference itself, and the possibility of extending it beyond the human domain, has long been debated. In particular, a central question arises: can the inferential capacity of LLMs be comparable to that of human beings? This article argues that the difference between human inference and the kind of inference produced by LLMs lies in the phenomenological, or affective, dimension of reasoning. More specifically, the distinction concerns the absence of metacognitive feelings that guide, monitor, and evaluate thought processes, such as reasoning or inferring. Although LLMs are capable of simulating novel inferences, their tendency to generate inferential hallucinations reveals the lack of the self-regulatory dimension that characterizes human reasoning. In contrast to the affective component present in human condition, inferential failures in LLMs can be interpreted in terms of three aspects: the reversal curse, inferential hallucinations resulting from decoding strategies, and overconfidence. Accordingly, it is concluded that one of the causes of inferential hallucinations is the absence of metacognitive feelings, including the lack of a feeling of correctness, the absence of the balance provided by a sense of reliability, and the lack of doubt.

Keywords: inferential hallucinations, large language models, metacognitive feelings.

INTRODUCCIÓN

Los grandes modelos de lenguaje (*Large Language Models* o LLMs) son cada vez más populares debido a su capacidad para simular la comprensión de las preguntas de los usuarios y para responder de una forma fluida que aparentemente es propia del ser humano. Desde el advenimiento del trabajo de Vaswani et al. (2017), con su sistema «Transformers» de multiatención y la tokenización, y luego de la llegada de OpenAI con ChatGPT en 2018, ha habido una respuesta de algunas personas que se puede denominar *tecnófoba*, porque se manifiesta en ella el temor a la tecnología y, en concreto, a que reemplace roles como la traducción o la elaboración de textos académicos (Hill-Yardin et al., 2023; Mejía-Rendón & Olezza, 2024).

El potencial de los grandes modelos de lenguaje radica en su capacidad para hacer inferencias una vez que el usuario ha introducido ciertas premisas básicas. De tal suerte que cuando un usuario interactúa con los grandes modelos de lenguaje, el modelo examina la entrada de texto y utiliza su conocimiento previo, generado en su etapa de entrenamiento, para luego dar una respuesta apropiada de manera similar a como los humanos responderían. En la segunda etapa, inferencial, los grandes modelos de lenguaje incluso simulan ciertos rasgos de creatividad en sus respuestas, en la medida en

que parecen creativas y novedosas:[1] ¿Es comparable la capacidad inferencial de los grandes modelos de lenguaje con la de los seres humanos?[2]

En el presente trabajo se explora un argumento basado en la afectividad del razonamiento humano y se sostendrá que la diferencia clave entre las inferencias humanas y las de los grandes modelos de lenguaje radica en que el razonamiento humano está constituido por *sentimientos metacognitivos*. Aunque se han propuesto varios términos para describir esta clase de sentimientos —como sentimientos cognitivos (de Sousa, 2009), sentimientos noéticos (Dokic, 2012) y sentimientos metacognitivos (Arango-Muñoz, 2011; Arango-Muñoz & Michaelian, 2014)—, su principal característica es no solamente que influyen en el proceso inferencial, sino que motivan, guían y evalúan las inferencias (Fernández Velasco, 2022; Fernandez Velasco & Loev, 2022). Así, ante la ausencia de esa dimensión afectiva, los grandes modelos de lenguaje generan respuestas que simulan inferencias novedosas, tendiendo a producir alucinaciones inferenciales como «la maldición de la reversión».

Este artículo tiene tres secciones principales: en la primera parte, se introduce un enfoque afectivo sobre las inferencias, en el cual se propone identificar y categorizar algunos sentimientos metacognitivos involucrados en el proceso inferencial. En específico, se argumentará que esta clase de sentimientos son parte integral del proceso de razonamiento humano. En segundo lugar, se evaluará el proceso de entendimiento inferencial de ejercicios lingüísticos conocidos como *casos de ambigüedad sintáctica*. En tercer lugar, se analizarán los estudios sobre las capacidades inferenciales en los grandes modelos de lenguaje, con el fin de revelar deficiencias relacionadas con el concepto alucinatorio inferencial de los mismos, derivados de la falta de sentimientos metacognitivos.

SIENTO QUE INFIERO: LA EXPERIENCIA AFECTIVA DE LAS INFERENCIAS CREATIVAS

La literatura filosófica define el *razonamiento teorético* como un proceso inferencial, en el cual las transiciones deliberadas entre pensamientos están guiadas por reglas lógicas (como *modus ponens* o *modus tollens*).

La regla del *modus ponens* es una inferencia lógica que preserva validez de los argumentos y su estructura es la siguiente:

$$\begin{array}{l} A \rightarrow B \\ \underline{A} \\ B \end{array}$$

[1] Algunos académicos postulan que los actos verdaderamente «creativos» requieren «novedad», ya que repetir ideas existentes no puede considerarse como algo creativo, incluso si esas ideas se combinan de nuevas maneras. Sin embargo, otros sostienen que los actos creativos impactantes equilibran la novedad con la utilidad (Mukherjee, 2024). En el presente artículo se considera que lo *nuevo* y lo *novedoso* son independientes en la medida en que algo novedoso implica que es completamente original en el gran esquema del conocimiento humano.

[2] Si bien en el ámbito ingenieril se denominan «alucinaciones» a las respuestas erróneas generadas por los grandes modelos de lenguaje, un término más adecuado sería «delirios», ya que las alucinaciones en los seres humanos implican experiencias *subjetivas* como ver y escuchar cosas inexistentes; en cambio, los delirios tienen más relación con el *procesamiento simbólico* erróneo del sistema (tokens en este caso).

Por otro lado, la regla del *modus tollens* es una inferencia lógica que preserva validez de los argumentos y su estructura es la siguiente:

$$\begin{array}{c} A \rightarrow B \\ \underline{\neg B} \\ \neg A \end{array}$$

Los procesos inferenciales, por lo general, involucran un sentido de agencia mental en primera persona que lleva al sujeto de las premisas a las conclusiones (Boghossian, 2014, 2018; Broome, 2013). Sin embargo, los procesos inferenciales están acompañados de una sensación de que la conclusión se sigue de las premisas (Broome, 2014; Chudnoff, 2014, 2024). Esta sensación sugiere que la inferencia no puede explicarse completamente ni como una transición causal ni como una relación puramente gobernada por reglas lógicas (Shea, 2024). Lo que falta es una descripción afectiva que explique cómo la experiencia subjetiva configura los procesos inferenciales. En particular, cómo describir la sensación inmediata de que la conclusión se sigue sin una reflexión deliberada y está influenciada por el procesamiento afectivo (de Sousa, 2009; Proust & Pliushch, 2016).

En este sentido, la afectividad pasada por alto en los debates sobre inferencia podría ser caracterizada en términos de sentimientos metacognitivos. Pero, la dimensión afectiva de la inferencia no es nueva en la literatura filosófica. Esta ya había sido insinuada previamente por autores como Broome, Chudnoff y Shea, que ofrecen perspectivas afectivas sobre la experiencia subjetiva de seguir reglas. Broome (2013) enfatiza en la importancia de la sensación subjetiva de corrección al seguir reglas. Chudnoff (2024) introduce la noción de *apariencias inferenciales*, una experiencia afectiva al hacer inferencias. Y Shea (2024) señala que los aspectos afectivos de la inferencia están involucrados en medio del proceso inferencial y en la justificación de la conclusión.

Sin embargo, estos enfoques no consideran completamente el papel activo de los sentimientos metacognitivos para desencadenar la inferencia. El quid de la cuestión, sin embargo, es explicar cómo estos sentimientos dan forma al proceso inferencial.

A los sentimientos metacognitivos que más han sido descritos en la literatura psicológica, como el fenómeno de la *punta de la lengua* (*tip-of-the-tongue phenomenon*), la *sensación de saber* (*the feeling of knowing*), la *sensación de error* (*the feeling of error*) y la *sensación de olvido* (*the feeling of forgetting*), se suma la presencia de algunos sentimientos vinculados propiamente con las inferencias (Brown, 1991; Hart, 1965; Metcalfe & Shimamura, 1994; Proust, 2013). Para ello, se distinguen tres etapas metacognitivas —autopredicción, evaluación intermedia y posevaluación—, cada una caracterizada por sentimientos metacognitivos específicos (Arango-Muñoz & Michaelian, 2014; Goupil & Proust, 2023).

Primero, la autopredicción, en la que el sujeto anticipa si un proceso (por ejemplo, recordar o resolver un cálculo) probablemente tendrá éxito. Segundo, la evaluación intermedia, en la que la actividad cognitiva en curso es monitoreada y ajustada. Tercero, la posevaluación, en la que el resultado o desenlace de una inferencia es juzgado como válido o inválido. Estos procesos y evaluaciones son importantes para el uso preciso y efectivo de la inferencia en los procesos cognitivos. En lo

siguiente, se analizará cada uno de estos tipos de sentimientos relacionados con el razonamiento (Arango-Muñoz, 2011; Arango-Muñoz & Michaelian, 2014; Goupil & Proust, 2023).

El sentimiento de habilidad inferencial evalúa la competencia y motiva nuevas inferencias resolviendo problemas de razonamiento a través de la autopredicción (Chudnoff, 2024; Proust, 2013). De acuerdo con los principios de la psicología cognitiva, la capacidad de razonar varía con la edad y con la adquisición de habilidades específicas. Estas habilidades incluyen comprender la validez inferencial, analizar argumentos deductivos y aceptar premisas que contradicen la realidad (Markovits & Vachon, 1989). La comprensión de la validez de las inferencias, junto con la distinción entre este concepto y la verdad empírica, es un proceso que típicamente se desarrolla entre los 10 y 12 años de edad y continúa mejorando a medida que las personas crecen. Las investigaciones sobre el razonamiento deductivo y su desarrollo han indicado que los niños más pequeños (específicamente, aquellos de 5 a 7 años) encuentran dificultades para aceptar premisas condicionales del tipo «si-entonces» como base para razonar, mientras que los niños alrededor de los 10 o 12 años son capaces de diferenciar argumentos e identificar su estructura lógica (Moshman & Franks, 1986). En este artículo se propone, entonces, designar la percatación de la capacidad humana de razonar lógicamente como un «sentido de habilidad inferencial», que permite evaluar la aptitud para generar nuevas inferencias o resolver problemas de razonamiento.

La capacidad de ajustar el propio razonamiento está guiada por sentimientos metacognitivos de *fiabilidad* y *falta de fiabilidad*. Estos sentimientos expresan la evaluación intermedia de los procesos de razonamiento de uno mismo respecto a si están encaminados hacia el logro de un objetivo deseado. El sentimiento de *fiabilidad* monitorea el curso de la inferencia y evalúa si el sujeto avanza hacia la meta cognitiva deseada, permitiendo, en caso necesario, ajustar la estrategia cognitiva (de Sousa, 2009). Pero, también se puede presentar una sensación de *falta de fiabilidad*. Como al constatar que no se pueden recordar algunos nombres de los personajes de una novela y, entonces, se falla al reconstruir la trama —por ejemplo, se falla en reconstruir todo el árbol genealógico de la novela colombiana *Cien años de soledad* de García Márquez. Entonces, en medio del razonamiento, se dilucida que la hija adoptiva de Úrsula Iguarán y José Arcadio Buendía es Amaranta y no Rebeca—. Esta evaluación intermedia, en la que se revisa abruptamente la transición mental antes de llegar a una conclusión, también juega un papel en la habilidad para hacer inferencias (de Sousa, 2009).

Una vez hecha una inferencia, se evalúa si la conclusión está verdaderamente respaldada por las premisas a través de la posevaluación del proceso. Los estudiosos han discutido tres sentimientos metacognitivos distintos que surgen después de alcanzar una conclusión inferencial: confianza, corrección y aprobación. Aunque los tres influyen en la aceptación de una conclusión, cada uno tiene su papel distintivo. A continuación, se los esboza brevemente:

El sentimiento de confianza se refiere a cuán seguro se siente el sujeto de la verdad de la conclusión. Cuando falta confianza en una conclusión, es menos probable que se confíe en ella. Sin un fuerte sentimiento de corrección, una conclusión no será fácilmente aceptada como verdadera, y se vacilará en actuar basándose en ella. En su lugar, se revisa el proceso de razonamiento y se consideran otras alternativas antes de tomar decisiones.

El sentimiento de corrección se refiere a la validez de la inferencia. Es la sensación de que la conclusión se sigue lógicamente de las premisas y de que el proceso se ha realizado correctamente. Por ejemplo, al resolver un problema de matemáticas, un estudiante puede experimentar un sentimiento de corrección al inferir la respuesta correcta, lo que confirma la exactitud del resultado. Por el contrario, un sentimiento de error desencadena una revisión de la estrategia y motiva la búsqueda de soluciones alternativas (Fernandez Cruz et al., 2016). Esto puede llevar al estudiante a revisar el procedimiento de cálculo o a buscar orientación de un profesor, lo que permite una corrección del proceso de razonamiento.

El sentimiento de aprobación entra en juego cuando se considera la justificación epistémica de una conclusión. Es el reconocimiento de que una conclusión no solo es correcta, sino que además se sigue lógicamente de las premisas dadas. Kornblith (2012), por ejemplo, sostiene que el sentimiento de aprobación interviene cuando el sujeto reconoce que su conclusión es correcta y que se sigue de un conjunto de premisas. En la misma línea, Chudnoff (2024) discute características normativas del procesamiento inferencial que describen cómo las premisas sustentan una conclusión que está justificada por ellas.

En resumen, los sentimientos de habilidad inferencial evalúan la propia competencia y desencadenan nuevas inferencias o resuelven problemas de razonamiento mediante la autopredicción (Chudnoff, 2024; Proust, 2013). El sentimiento de fiabilidad monitorea el progreso de una inferencia e indica si está encaminada hacia el logro del objetivo previsto, permitiendo ajustes en la estrategia cognitiva cuando es necesario (de Sousa, 2009). Esta evaluación intermedia tiene la función de evaluar las transiciones mentales antes de llegar a una conclusión; también contribuye a la capacidad de razonar eficazmente (de Sousa, 2009). Una vez que se obtiene una inferencia, ocurre la posevaluación, en la cual se evalúa si la conclusión está realmente respaldada por las premisas. En esta etapa, los sentimientos de confianza, corrección y aprobación evalúan el resultado del proceso de razonamiento, guiando al sujeto a aceptar la conclusión como verdadera (Thompson et al., 2011). Estos procesos y evaluaciones son importantes para el uso preciso y efectivo de la inferencia en los procesos cognitivos.

Ahora bien, aunque los grandes modelos de lenguaje generan respuestas aparentemente coherentes a partir de premisas dadas, la pregunta es si estas inferencias tienen rasgos afectivos. De acuerdo con Bender et al. (2021), los grandes modelos de lenguaje realizan procesamiento meramente probabilístico basado en patrones, razón por la cual han sido llamados «loros estocásticos», y carecen de semántica interna real (Searle, 1980). Mientras que los grandes modelos de lenguaje muestran ciertas destrezas para generar respuestas basadas en premisas dadas, estas carecen del elemento afectivo presente, en muchas ocasiones, en el proceso de la inferencia humana, lo que limita su capacidad para realizar inferencias.

Aunque algunos autores, como Gallagher (2005), adoptan una postura más optimista al considerar la posibilidad de sistemas artificiales inteligentes dotados de cierta forma de subjetividad, el escepticismo de Dreyfus (1972) se mantiene firme: las computadoras carecen de las capacidades corporizadas que caracterizan la experiencia vívida del mundo propia de los humanos (Manzotti & Chella, 2018; Zahavi, 2025). Se puede decir que las máquinas carecen de algunas propiedades específicas que tenemos los humanos, debido a que su estructura no es biológica: las máquinas carecen de emociones.

Una de las características que se considerará es la carencia de motivación intrínseca en los grandes modelos de lenguaje. Los comportamientos intrínsecamente motivados son aquellos causados por un interés genuino del individuo. La principal recompensa por estos actos es el deleite y el disfrute que se derivan de ellos (Ryan & Deci, 2000). Estos autores afirman que se contrasta con la motivación extrínseca (es decir, realizar acciones para alcanzar algún resultado separable como una recompensa, evitar un castigo o un objetivo deseado). Desde una posición filosófica rigurosa, la motivación intrínseca implica una condición de autodeterminación, agencia voluntaria y generación de objetivos autónomos.

A diferencia de los sistemas vivos, los sistemas artificiales buscan maximizar funciones definidas externamente (motivación extrínseca). Las recompensas que están diseñadas manualmente para cada entorno no son escalables, y se necesita aprender *funciones de recompensa*[3] al agente; la «curiosidad artificial» es un ejemplo de una *función de recompensa* intrínseca, en la que la señal de recompensa proviene del error de predicción (Burda et al., 2018). Pero, estas señales están «encapsuladas» en una función «objetivo» definida externamente, en una semántica externa. Por ello, no es una motivación intrínseca real, ya que está supeditada a la motivación extrínseca que le viene dada desde afuera por la función recompensa establecida por el *diseñador* del sistema. Por ejemplo, una persona puede tener un trabajo que le guste (motivación intrínseca), y que además le den un sueldo por ello (motivación extrínseca), pero en este caso van de la mano. Mientras que si las motivaciones «intrínsecas» dependen jerárquicamente de las extrínsecas, no son motivaciones intrínsecas reales (que es lo que ocurre con las señales de recompensa internas del sistema mal llamadas «intrínsecas»: dependen de la *función de recompensa* externa).

En este sentido, la principal diferencia entre las inferencias de los grandes modelos de lenguaje y las inferencias humanas radica en que los procesos inferenciales humanos muchas veces están guiados por sentimientos metacognitivos.

¿INFERENCIAS MAQUÍNICAS? EL CASO DE LAS AMBIGÜEDADES SINTÁCTICAS

Tal como se definió anteriormente, la *inferencia* implica un proceso racional que lleva desde ciertas premisas hasta una conclusión (Boghossian, 2014, 2018; Broome, 2013). Desde una perspectiva normativista, en una inferencia el sujeto debe tomar (*take*) sus premisas para apoyar su conclusión (Boghossian, 2014). En este sentido, según el postulado que propuso Boghossian (2014) de la *condición de toma* (*Taking Condition*, de ahora en adelante condición-T), las inferencias requieren que el sujeto tome (*taking*) sus premisas para apoyar su conclusión y que, en un buen razonamiento, esta toma (*take*) esté respaldada por una intuición en el sentido de que la aprehensión es verdadera. Desde esta perspectiva, entonces, hacer inferencias implica: (a) un proceso voluntario en primera persona; (b) ser consciente y (c) que esté estructurada bajo ciertas reglas lógicas (Mejía-Rendón & Olezza, 2024). Dicha concepción normativa se ilustra bien con el razonamiento o la inferencia de la lluvia

[3] El objetivo de un agente de aprendizaje es entonces encontrar el comportamiento con el rango más alto, dado por la «función de recompensa» (Eschmann, 2021, p. 25).

(*Rain-Argument*): «(1) llovió anoche; (2) si llovió anoche, entonces las calles están mojadas; (3) las calles están mojadas» (Boghossian, 2018, p. 55).[4]

No obstante, esta caracterización normativa de la inferencia ha sido ampliamente cuestionada en relación con (i) qué implica exactamente ese «tomar» mencionado por Boghossian (Hlobil, 2014; Wright, 2014); (ii) qué estatus agencial tiene la inferencia, en particular si inferir constituye una acción mental genuina (Boyle, 2011; Strawson, 2003); (iii) qué carácter tiene el seguimiento de reglas (Hlobil, 2014) y; (iv) se amplía la pregunta sobre la posibilidad de inferencias involuntarias o inconscientes (Quilty-Dunn & Mandelbaum, 2018).

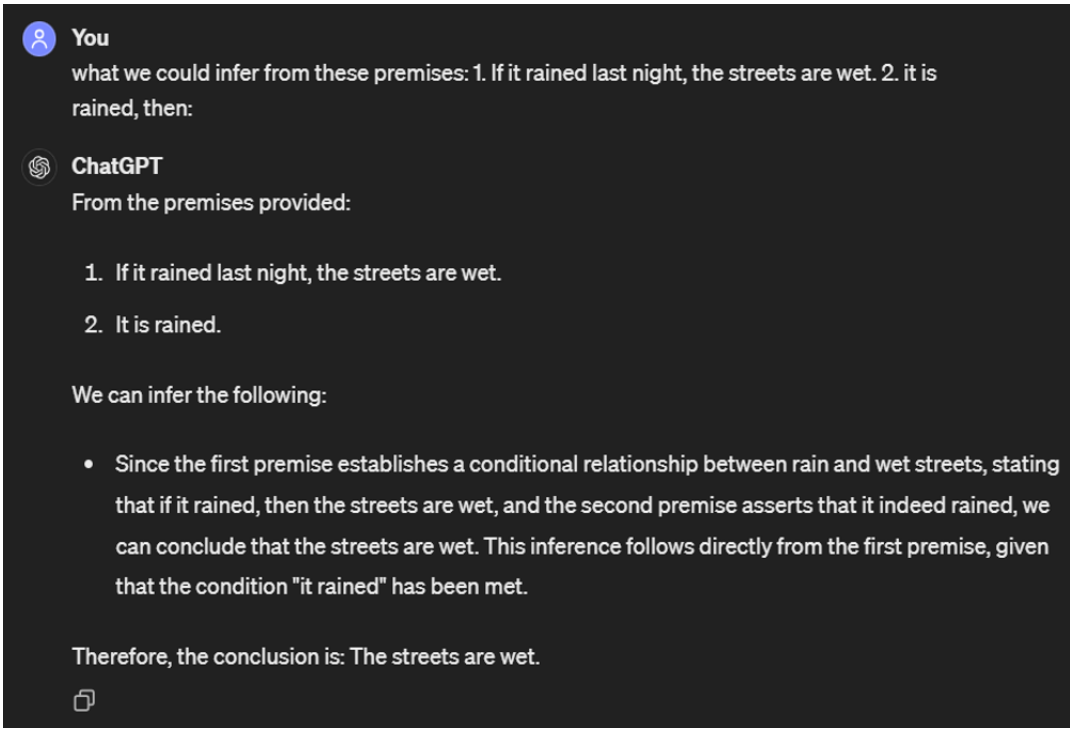
En este contexto, diversos autores han discutido la condición-T (Broome, 2013, 2014; Chudnoff, 2024). En general, se ha señalado que esta conduce tanto al «problema del regreso» y, a su vez, a cierta forma de «intelectualismo», al exigir una cadena infinita de creencias justificatorias y capacidades conceptuales sofisticadas (Broome, 2013; Valaris, 2017; Wright, 2014). Así, pese a sus diferencias internas, los enfoques normativos de la condición-T no logran superar de manera satisfactoria estas dificultades, lo que abre el espacio para explorar una alternativa que incorpore de forma más explícita los aspectos afectivos y metacognitivos involucrados en la realización misma de la inferencia.

Por tanto, el presente artículo no pretende ofrecer una teoría normativa de la inferencia ni una explicación psicológica de las transiciones inferenciales, sino una caracterización metacognitiva de la inferencia como actividad mental; tal como se da desde la perspectiva del sujeto en primera persona. En este sentido, es necesario subrayar que el papel de los sentimientos metacognitivos no los introduce como condiciones normativas de corrección o justificación, ni como meros acompañamientos subjetivos de los procesos inferenciales previamente definidos. Más bien, ciertos sentimientos metacognitivos forman parte del modo mismo de inferencia, en la medida en que disparan, sostienen, monitorizan y cierran la transición desde las premisas hasta la conclusión.

Ahora bien, si se atiende a esta noción de inferencia, se podría decir, en principio, que los grandes modelos de lenguaje no tendrían dificultades para cumplir con los requisitos inferenciales desde una perspectiva normativista. Esto se debe a que el software extrae la conclusión correcta a partir de ciertas premisas. En el caso de ChatGPT, incluso se va más allá al agregar una nueva conclusión y ser capaz de justificar cómo ha llegado a esa conclusión (Figura 1). Lo anterior, ciertamente, parecería igualar las capacidades inferenciales del programa con las capacidades inferenciales humanas:

[4] Texto original en inglés: 1. It rained last night; 2. If it rained last night, the streets are wet; 3. The streets are wet.

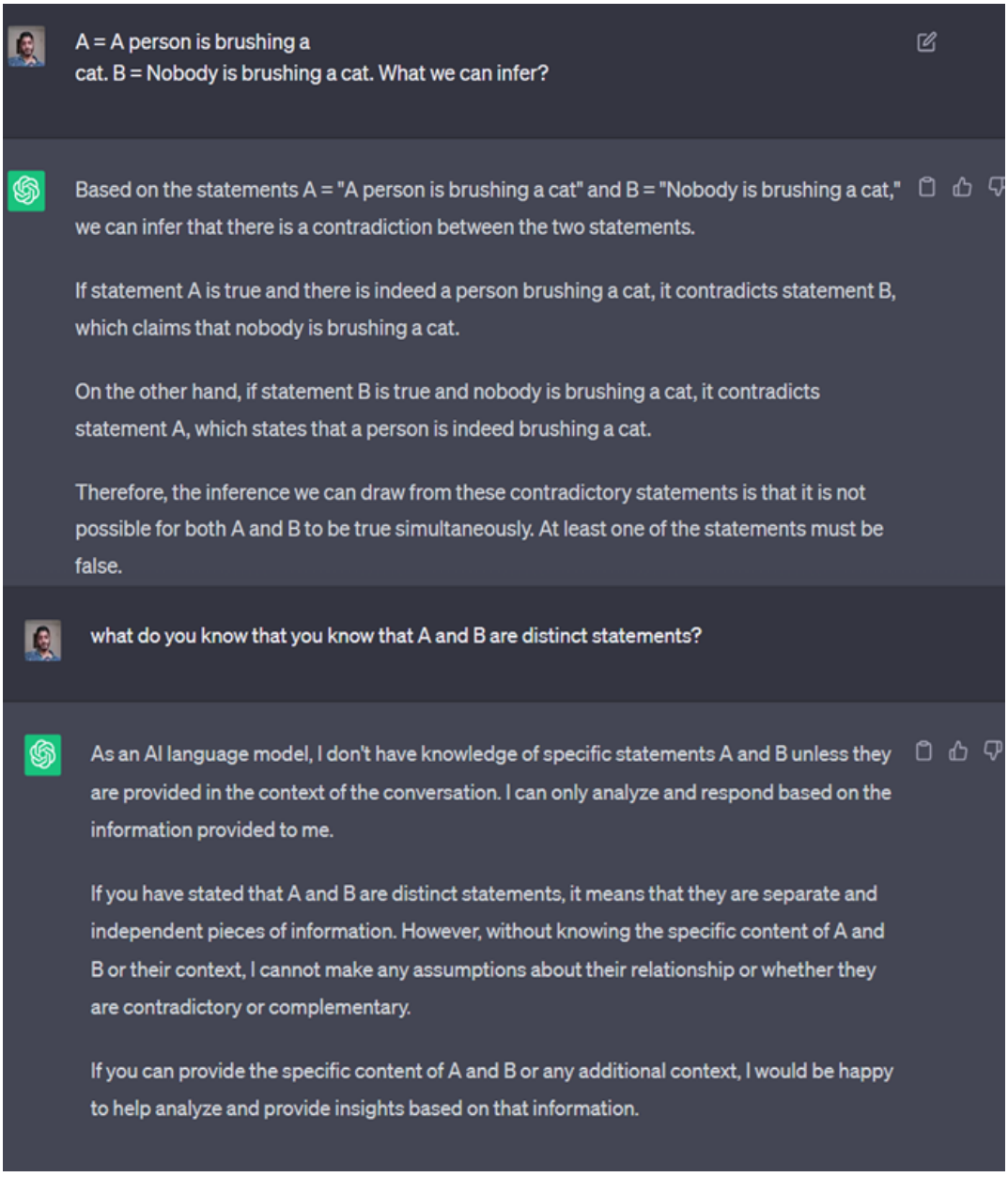
Figura 1. Interacción con el software ChatGPT (10 de julio de 2023)



Fuente: elaboración propia.

No obstante, una serie de estudios revela que ChatGPT tiene problemas con la implicación especificada gramaticalmente, tal como sucede con los adverbios evidenciales que expresan incertidumbre y, cuando el programa se enfrenta a ello, se muestra una deficiencia en la comprensión de conceptos lingüísticos subyacentes (Kalouli et al., 2019). De manera específica, los estudios se centran en pruebas de *inferencia en lenguaje natural* (*Natural Language Inference*), que consisten en definir la relación semántica entre un texto de premisa y uno de conclusión (Figura 2). Los resultados revelan deficiencias en la comprensión de ChatGPT y, consecuentemente, en las inferencias que realiza (Basmov et al., 2023).

Figura 2. Interacción con el software ChatGPT (10 de julio de 2023)



Fuente: elaboración propia.

Por ejemplo, ChatGPT no solo presenta fallos en la lengua inglesa, sino también en español. En este sentido, se puede mencionar un ejercicio muy estudiado en lingüística conocido como *los casos de ambigüedad sintáctica*. Específicamente, al utilizar un ejemplo de cláusulas relativas con doble antecedente nominal (Stetie, 2023). El uso de antecedentes nominales dobles en las cláusulas relativas puede generar ambigüedad sintáctica, con la posibilidad de dos interpretaciones diferentes según la posición jerárquica de la cláusula relativa. En este caso, como se puede ver en la Figura 3, la potencial ambigüedad sintáctica de la oración (1) «el policía disparó al empleado del cerrajero

que estaba en el local» reside en que la frase «que estaba en el local» (cláusula relativa) puede interpretarse en el sentido de que depende de dos frases nominales diferentes o sujetos, a saber, «el empleado» o «el cerrajero».

Figura 3. Interacción con el software ChatGPT (10 de julio de 2023)



Fuente: elaboración propia.

La existencia de casos como las inferencias que requieren conocimiento previo o las cláusulas ambiguas plantea interrogantes sobre la capacidad inferencial de ChatGPT, lo cual puede generar dudas en

relación con sus aplicaciones teóricas o prácticas, especialmente en campos sensibles al riesgo, como el derecho, la sanidad y las finanzas (Borji, 2023). Sin embargo, es importante destacar que los seres humanos también somos propensos a cometer errores. Por ejemplo, una persona puede pasar por alto una suposición falsa en una pregunta que está fuera de su área de especialización o simplemente realizar inferencias incorrectas aun cuando conoce las premisas. Los grandes modelos de lenguaje, por su parte, pueden fallar en ocasiones al manejar tareas inferenciales, por errores alucinatorios, como se verá en el próximo apartado, pero que suelen ser de mayor gravedad.

DESAFÍOS EN GRANDES MODELOS DE LENGUAJE

Mientras que algunas aproximaciones defienden basarse en la implementación de reglas inferenciales o aproximación estadística, la propuesta del presente trabajo *añade* a la primera línea (que es de hecho una de las causas de las alucinaciones inferenciales, discutida más abajo), una segunda línea, aportando una causa por ausencia de sentimientos metacognitivos. A continuación se explicarán brevemente las inferencias en los grandes modelos de lenguaje para luego hablar de alucinaciones inferenciales y sus causas por ausencia de sentimientos metacognitivos.

Inferencias en grandes modelos de lenguaje

A diferencia de la mente humana, que opera en tres etapas, a saber: *abducción* → *deducción* → *inducción* (Bellucci & Pietarinen, 2023), los grandes modelos de lenguaje constan de dos etapas: la etapa de entrenamiento, en la cual se entrena al modelo, con grandes cantidades de datos disponibles, descenso del gradiente (Ruder, 2016) y retropropagación (Bryson et al., 1969); y la etapa operativa (que es la que los usuarios utilizan), la cual es *inferencial*. Es probabilística, como las típicas redes neuronales artificiales (Goodfellow et al., 2016; Rumelhart & McClelland, 1986), y carece de capacidades abductivas (las preguntas abductivas vienen dadas por el *usuario* en forma de «conjeturas»). La abducción no está presente en los grandes modelos de lenguaje (Larson, 2021). Los comandos introducidos por el usuario también pueden ser de tipo imperativo, por ejemplo: «Lístame las películas de Woody Allen».

Las respuestas a preguntas contrafácticas o abductivas tienen menor desempeño en los grandes modelos de lenguaje de las evaluaciones comparativas estudiadas asociadas (Chen et al., 2025). Esto está relacionado, también, a la teoría de la escalera de la causalidad de Pearl que propone tres estratos: el inferior, meramente correlacional —donde está posicionada toda la inteligencia artificial (IA) hoy en día con su conjunto de datos—, el intervencional y, finalmente en la cúpula de la escalera, el nivel causal (Pearl & Mackenzie, 2018). De acuerdo con el teorema de jerarquía de Pearl (2019):

$$Correlación < Intervención < Contrafactual \ / \ Causalidad$$

Para pasar del nivel correlacional al nivel causal, hacen falta muchos presupuestos extra, lo que lleva a llenar «vacíos de información» provocando alucinaciones. Asimismo, las causas de las alucinaciones pueden ser *múltiples*. A continuación se explicarán algunas causas que derivan de la ausencia

de sentimientos metacognitivos en grandes modelos de lenguaje, que es el enfoque principal de este artículo.

Alucinaciones inferenciales

Los grandes modelos de lenguaje tienen la tendencia a producir respuestas erróneas en términos fácticos o de fiabilidad.[5] En cuanto a los *fácticos*, las alucinaciones más problemáticas son aquellas que, con gran nivel de confianza, generan respuestas incorrectas. Este fenómeno puede conducir a decisiones equivocadas, difusión de información falsa y otros efectos indeseables. Las causas de alucinaciones en los grandes modelos de lenguaje se pueden clasificar taxonómicamente, de acuerdo con Huang et al. (2023), en 3 tipos: 1) por datos (como los sesgos), 2) por entrenamiento (como la «adulación algorítmica» al usuario), o 3) por inferencias, una de cuyas causas es la falta de carencia metacognitiva, como se analizará a continuación.

En primer lugar, Berglund et al. (2023) revelaron fallos de razonamiento que denominaron «la maldición de la reversión» (*The Reversal Curse*). Si un modelo es entrenado con una oración de la forma «A es B», no generalizará automáticamente a la dirección inversa «B es A». El fenómeno de la «la maldición de la reversión» se considera una clara evidencia de ausencia en el metaaprendizaje del modelo:

> La maldición de la reversión pone de manifiesto una falla en el metaaprendizaje. Oraciones del tipo "<nombre> es <descripción>" y "<descripción> es <nombre>" suelen coaparecer en los conjuntos de datos de preentrenamiento; si la primera aparece en un conjunto de datos, resulta intuitivo pensar que la segunda también tiene una mayor probabilidad de aparecer. Esto se debe a que los seres humanos suelen variar el orden de los elementos dentro de una oración o de un párrafo. Por ello, un *buen meta-aprendiz debería aumentar la probabilidad de una instancia de "<descripción> es <nombre>" después de haber sido entrenado con "<nombre> es <descripción>"*. (Berglund et al., 2023, pp. 2-3)[6, 7]

Berglund et al. (2023) muestran en sus resultados que los grandes modelos de lenguaje no son buenos «metaaprendices»: fallan tanto con preentrenamiento con datos sintéticos (ficticios) como con datos reales.

Los grandes modelos de lenguaje están diseñados para imitar el lenguaje humano, prediciendo texto en lugar de establecer la verdad factual (Lin et al., 2022). Aunque el modelo internamente haya captado la inversión, podría no manifestarlo en su respuesta. Entrenar un modelo con la relación

[5] Las alucinaciones en términos de fiabilidad se refieren más específicamente a que el gran modelo de lenguaje no siga las instrucciones asignadas.

[6] Las itálicas han sido añadidas por los autores.

[7] Texto original en inglés: The Reversal Curse demonstrates a failure of meta-learning. Sentences of the form "<name> is <description>" and "<description> is <name>" often co-occur in pretraining datasets; if the former appears in a dataset, the latter is intuitively more likely to appear. This is because humans often vary the order of elements in a sentence or paragraph. Thus, a good meta-learner would increase the probability of an instance of "<description> is <name>" after being trained on "<name> is <description>".

«A es B» parece cambiar la forma interna de A para que guarde datos de B. Se podría pensar que también debería cambiar la representación de B, dándole algo de A. Pero el algoritmo de gradiente (en la etapa de entrenamiento) probablemente sea «miope». Su meta es solo acertar que B sigue a A en ese momento, usando los *logits*.[8] No se programa para imaginar que más tarde alguien querría invertir la dirección.[9] En la práctica, la actualización sigue el camino A → B y casi nunca «mira de regreso», B → A. Por eso la información inversa suele quedar vacía.

> ¿Los modelos fallan en invertir otros tipos de relaciones (como predice la maldición de la reversión)? Entre estas pueden incluirse las implicaciones lógicas (por ejemplo, «X implica Y» y «No X implica no Y»), las relaciones espaciales (por ejemplo, «La taza está sobre la mesa» y «La mesa está debajo de la taza»), o las relaciones de n lugares (por ejemplo, «Alice, Bob, Carol y Dan están en el mismo grupo»). (Berglund et al., 2023, p. 9)[10]

La «maldición de la reversión» puede entenderse como un fallo en la posevaluación inferencial, a saber, el sistema carece de cualquier *sentimiento de corrección* que le permita evaluar la validez estructural de la inferencia, yendo *más allá* del patrón ya aprendido.

En segundo lugar, se tienen las alucinaciones inferenciales por estrategias de decodificación, a saber: los grandes modelos de lenguaje parecen poder crear textos muy creativos y variados, algo que depende en gran parte la aleatoriedad que usan al generar palabras (Olezza, 2025). Pero esa variedad tiene precio; suele ir ligada a las alucinaciones si no se regulan adecuadamente los hiperparámetros. Cuando se aumenta el valor de los hiperparámetros generativos (semántica externa introducida por el supervisor del sistema o por el usuario) como la temperatura (T#) de muestreo, o Top-K,[11] la probabilidad de elegir cualquier token se vuelve más pareja. Luego, aparecen tokens más extraños. Esa tendencia a elegir palabras poco comunes eleva el riesgo de generar información errónea. En conclusión, más aleatoriedad puede dar ideas nuevas, pero también aumenta la posibilidad de errores; por lo tanto, *son los usuarios* quienes siempre deberían vigilar cuidadosamente los resultados generados. La creatividad aparente, entonces, inducida por ejemplo por T#, no está balanceada por ningún *sentimiento de fiabilidad*, lo que explica por qué la generación en los grandes modelos de lenguaje, si el usuario no vigila correctamente, puede ir acompañada de alucinaciones.

[8] Los *logits* son los valores reales que produce la capa final de un modelo de clasificación antes de aplicar una función de activación como *softmax* (capa de salida).

[9] Esto está relacionado con alucinación por entrenamiento del tipo «preentrenamiento», ya que los modelos suelen intentar predecir el siguiente token basado en los anteriores, de forma unidireccional de izquierda a derecha. Esto posibilita entrenamiento veloz, aunque parece que reduce su habilidad para captar dependencias contextuales complejas (Chen et al., 2022).

[10] Texto original en inglés: Do models fail to reverse other types of relation (as the Reversal Curse predicts)? These could include logical implications (e.g. "X implies Y" and "Not X implies not Y."), spatial relationships (e.g. "The cup is on the table" and "The table is under the cup."), or n-place relations (e.g. "Alice, Bob, Carol and Dan are in the same group.").

[11] Top-K está encargado de distribuir las probabilidades en los próximos K tokens. Si se configura en 1, será total la probabilidad de elegir el token más probable (próximo token = 100 % probabilidad); pero al seleccionar, por ejemplo, K = 500, las probabilidades se distribuyen entre los próximos 500 tokens, logrando que el sistema se vuelva más aleatorio.

En tercer lugar, los grandes modelos de lenguaje pecan del llamado «exceso de confianza»: estudios han descubierto este problema en la generación de texto condicional (Lin, 2004; Yang et al., 2021). Al parecer, estos modelos también pueden concentrarse demasiado en la frase que se está creando justo en ese momento y priorizar la fluidez y la coherencia breve, lo que a veces perjudica la precisión respecto al contexto original o a la fuente de datos reales y significativos. El «exceso de confianza» observado en los grandes modelos de lenguaje es consecuencia directa de la *ausencia de sentimientos metacognitivos negativos*, como por ejemplo la duda o el sentimiento de error.

CONCLUSIÓN

La popularidad de los grandes modelos de lenguaje ha generado diversas reacciones entre aquellos preocupados por su potencial reemplazo en áreas como la traducción y la redacción académica. La diferencia fundamental entre las inferencias realizadas por estos modelos y las inferencias humanas reside en dos causas fundamentales:

- Los procesos inferenciales humanos a menudo están influenciados por sentimientos metacognitivos. Esto sugiere que la experiencia afectiva de la inferencia, mencionada anteriormente, puede considerarse como una característica distintiva de los procesos inferenciales humanos. Se ha explorado en este artículo, entonces, la naturaleza de las inferencias tanto en grandes modelos de lenguaje como en humanos, enfatizando en los sentimientos metacognitivos que guían y evalúan el proceso del razonamiento.
- Los grandes modelos de lenguaje, al carecer de dichos sentimientos, son proclives a alucinaciones inferenciales de orden más grave que en el ser humano, como los mencionados en el último apartado: la «maldición de la reversión» (carencia de sentimiento de corrección), las estrategias de decodificación (falta de balance por sentimiento de fiabilidad) y el exceso de confianza (ausencia de sentimientos metacognitivos negativos).

Aunque los grandes modelos de lenguaje son capaces de producir respuestas que guardan coherencia lógica con las premisas dadas, carecen de la interioridad afectiva y de la vivencia afectiva que hacen de la inferencia humana una experiencia rica. Los sentimientos metacognitivos desempeñan un papel crucial en la evaluación y el ajuste del razonamiento humano, y permiten una navegación más efectiva a través de la complejidad del pensamiento inferencial. Es evidente que este aspecto afectivo no está presente en las inferencias de los grandes modelos de lenguaje, ya que este sistema carece de experiencia subjetiva en sí misma. Al atribuir un tipo específico de sentimiento epistémico al proceso inferencial, se ha presentado un enfoque para comprender la naturaleza de la inferencia humana que sugiere cierta peculiaridad distintiva. Se plantea entonces una posible vía para futuras investigaciones en este ámbito.

CONFLICTOS DE INTERÉS

Los autores declaran que no tienen conflictos de intereses.

CONTRIBUCIÓN DE AUTORÍA

Sebastián Mejía-Rendón y Mariana Olezza se encargaron de la redacción, borrador original, recursos, metodología, investigación, análisis formal y conceptualización.

REFERENCIAS

Arango-Muñoz, S. (2011). Two Levels of Metacognition. *Philosophia*, *39*(1), 71-82. https://doi.org/10.1007/s11406-010-9279-0

Arango-Muñoz, S., & Michaelian, K. (2014). Epistemic Feelings, Epistemic Emotions: Review and Introduction to the Focus Section. *Philosophical Inquiries*, *2*(1), 97-122. https://doi.org/10.4454/philinq.v2i1.79

Basmov, V., Goldberg, Y., & Tsarfaty, R. (2023). *Simple Linguistic Inferences of Large Language Models (LLMs): Blind Spots and Blinds*. arXiv. https://doi.org/10.48550/arXiv.2305.14785

Bellucci, F., & Pietarinen, A.-V. (2023). Peirce's Abduction. En L. Magnani (Ed.), *Handbook of Abductive Cognition* (pp. 7-20). Springer. https://doi.org/10.1007/978-3-031-10135-9_7

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. En *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2023). *The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A"*. arXiv. https://doi.org/10.48550/arXiv.2309.12288

Boghossian, P. (2014). What is inference? *Philosophical Studies*, *169*(1), 1-18. https://doi.org/10.1007/s11098-012-9903-x

Boghossian, P. (2018). Delimiting the Boundaries of Inference. *Philosophical Issues*, *28*(1), 55-69. https://doi.org/10.1111/phis.12115

Borji, A. (2023). *A Categorical Archive of ChatGPT Failures*. arXiv. https://doi.org/10.48550/arXiv.2302.03494

Boyle, M. (2011). Transparent Self-Knowledge. *Aristotelian Society*, *85*(1), 223-241. https://doi.org/10.1111/j.1467-8349.2011.00204.x

Broome, J. (2013). Practical Reasoning and Inference. En D. Bakhurst, B. Hooker & M. O. Little (Eds.), *Thinking About Reasons: Themes from the Philosophy of Jonathan Dancy* (pp. 286-309). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199604678.003.0014

Broome, J. (2014). Normativity in Reasoning. *Pacific Philosophical Quarterly*, *95*(4), 622-633. https://doi.org/10.1111/papq.12050

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, *109*(2), 204-223. https://doi.org/10.1037/0033-2909.109.2.204

Bryson, A. E., Ho, Y.-C., & Siouris, G. M. (1969). Applied Optimal Control: Optimization, Estimation, and Control. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(6), 366-367. https://doi.org/10.1109/TSMC.1979.4310229

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., & Efros, A. A. (2018). *Large-Scale Study of Curiosity-Driven Learning*. arXiv. https://doi.org/10.48550/arXiv.1808.04355

Chen, X., Li, M., Gao, X., & Zhang, X. (2022). Towards improving faithfulness in abstractive summarization. En *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS'22)* (pp. 24516-24528, art. 1780). Curran Associates Inc. https://dl.acm.org/doi/10.5555/3600270.3602050

Chen, Y., Singh, V. K., Ma, J., & Tang, R. (2025). *CounterBench: A Benchmark for Counterfactuals Reasoning in Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2502.11008

Chudnoff, E. (2014). Intuition in Mathematics. En L. M. Osbeck & B. S. Held (Eds.), *Rational Intuition: Philosophical Roots, Scientific Investigations* (pp. 174-191). Cambridge University Press. https://doi.org/10.1017/CBO9781139136419.010

Chudnoff, E. (2024). Inferential Seemings. En U. Kriegel (Ed.), *Oxford Studies in Philosophy of Mind: Volume 4* (pp. 311-333). Oxford University Press. https://doi.org/10.1093/9780198924159.003.0009

de Sousa, R. (2009). Epistemic Feelings. *Mind & Matter*, *7*(2), 139-161. https://www.researchgate.net/publication/233488784_Epistemic_Feelings

Dokic, J. (2012). Seeds of self-knowledge: Noetic feelings and metacognition. En M. J. Beran, J. L. Brandl, J. Perner & J. Proust (Eds.), *Foundations of Metacognition* (pp. 302-321). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199646739.003.0020

Dreyfus, H. L. (1972). *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row.

Eschmann, J. (2021). Reward Function Design in Reinforcement Learning. En B. Belousov, H. Abdulsamad, P. Klink, S. Parisi & J. Peters (Eds.), *Reinforcement Learning Algorithms: Analysis and Applications* (pp. 25-33). Springer.

Fernandez Cruz, A. L., Arango-Muñoz, S., & Volz, K. G. (2016). Oops, scratch that! Monitoring one's own errors during mental calculation. *Cognition*, *146*, 110-120. https://doi.org/10.1016/j.cognition.2015.09.005

Fernández Velasco, P. (2022). Disorientation and self-consciousness: a phenomenological inquiry. *Phenomenology and the Cognitive Sciences*, *21*(1), 203-222. https://doi.org/10.1007/s11097-020-09659-1

Fernandez Velasco, P., & Loev, S. (2022). *Cognitive Feelings in the Predictive Mind: Emotion, Meta-cognition and Predictive Processing*. PsyArXiv. https://doi.org/10.31234/osf.io/594hj

Gallagher, S. (2005). *How the Body Shapes the Mind*. Oxford University Press. https://doi.org/10.1093/0199271941.001.0001

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.

Goupil, L., & Proust, J. (2023). Curiosity as a metacognitive feeling. *Cognition*, *231*, art. 105325. https://doi.org/10.1016/j.cognition.2022.105325

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, *56*(4), 208-216. https://doi.org/10.1037/h0022263

Hill-Yardin, E. L., Hutchinson, M. R., Laycock, R., & Spencer, S. J. (2023). A Chat(GPT) about the future of scientific publishing. *Brain, Behavior and Immunity*, *110*, 152-154. https://doi.org/10.1016/j.bbi.2023.02.022

Hlobil, U. (2014). Against Boghossian, Wright and Broome on inference. *Philosophical Studies*, *167*(2), 419-429. https://doi.org/10.1007/s11098-013-0104-z

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, *43*(2), art. 42. https://doi.org/10.1145/3703155

Kalouli, A.-L., Buis, A., Real, L., Palmer, M., & De Paiva, V. (2019). Explaining simple natural language inference. En *Proceedings of the 13th Linguistic Annotation Workshop* (pp. 132-143). Association for Computational Linguistics.

Kornblith, H. (2012). *On Reflection*. Oxford University Press.

Larson, E. J. (2021). *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*. Harvard University Press.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. En *Text Summarization Branches Out* (pp. 74-81). Association for Computational Linguistics. https://aclanthology.org/W04-1013.pdf

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. En S. Muresan, P. Nakov & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214-3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

Manzotti, R., & Chella, A. (2018). Good Old-Fashioned Artificial Consciousness and the Intermediate Level Fallacy. *Frontiers in Robotics and AI*, *5*, art. 39. https://doi.org/10.3389/frobt.2018.00039

Markovits, H., & Vachon, R. (1989). Reasoning with contrary-to-fact propositions. *Journal of Experimental Child Psychology*, *47*(3), 398-412. https://doi.org/10.1016/0022-0965(89)90021-0

Mejía-Rendón, J. S., & Olezza, M. (2024). ChatGPT e Inferencias. En S. Stein & V. Braga (Eds.), *Libro de resúmenes: XIII Encuentro de filosofía e historia de la ciencia del Cono Sur 2024* (pp. 170-171). Editorial Philosophia. http://www.afhic.com/wp-content/uploads/2024/06/Caderno-de-Resumos-AFHIC-2024-Vers%C3%A3o-final.pdf

Metcalfe, J., & Shimamura, A. P. (Eds.). (1994). *Metacognition: Knowing about Knowing*. The MIT Press. https://doi.org/10.7551/mitpress/4561.001.0001

Moshman, D., & Franks, B. A. (1986). Development of the Concept of Inferential Validity. *Child Development*, *57*(1), 153-165. https://doi.org/10.2307/1130647

Mukherjee, A. (2024). *Psittacines of Innovation? Assessing the True Novelty of AI Creations*. arXiv. https://doi.org/10.48550/arXiv.2404.00017

Olezza, M. (2025). Analogical mappings of facts and counterfactuals in the human mind and Peirce's abduction: limitations in LLMs. *Cognitive Systems Research*, *94*, art. 101408. https://doi.org/10.1016/j.cogsys.2025.101408

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54-60. https://doi.org/10.1145/3241036

Pearl, J., & Mackenzie, D. (2018). *The Book of Why*. Basic Books.

Proust, J. (2013). *The Philosophy of Metacognition: Mental Agency and Self-Awareness*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199602162.001.0001

Proust, J., & Pliushch, I. (2016). The Representational Structure of Feelings. En T. Metzinger & J. M. Windt (Eds.), *Open MIND, 2-vol. set: Philosophy and the Mind Sciences in the 21st Century*. The MIT Press. https://doi.org/10.7551/mitpress/10603.003.0094

Quilty-Dunn, J., & Mandelbaum, E. (2018). Against dispositionalism: belief in cognitive science. *Philosophical Studies*, *175*(9), 2353-2372. https://doi.org/10.1007/s11098-017-0962-x

Ruder, S. (2016). *An overview of gradient descent optimization algorithms*. arXiv. https://doi.org/10.48550/arXiv.1609.04747

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. The MIT Press.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, *25*(1), 54-67. https://doi.org/10.1006/ceps.1999.1020

Searle, J. R. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, *3*(3), 417-457. https://doi.org/10.1017/S0140525X00005756

Shea, N. (2024). Metacognition of Inferential Transitions. *Journal of Philosophy*, *121*(11), 597-627. https://doi.org/10.5840/jphil20246202

Stetie, N. A. (2023). *¿Con qué lentes buscamos a Wally? Articulaciones y problemas de interfaz entre la lingüística y el estudio de la cognición* [conferencia]. Workshop de Estudios sobre Ciencias Cognitivas y Neurociencias, Córdoba, Argentina. https://www.conicet.gov.ar/new_scp/detalle.php?keywords=&id=56674&congresos=yes&detalles=yes&congr_id=11015032

Strawson, G. (2003). Mental Ballistics or the Involuntariness of Spontaneity. *Proceedings of the Aristotelian Society*, *103*(1), 227-256. https://doi.org/10.1111/j.0066-7372.2003.00071.x

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107-140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Valaris, M. (2017). What reasoning might be. *Synthese*, *194*(6), 2007-2024. https://doi.org/10.1007/s11229-016-1034-z

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. En *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* (pp. 6000-6010). Curran Associates Inc. https://dl.acm.org/doi/10.5555/3295222.3295349

Wright, C. (2014). Comment on Paul Boghossian, "What is inference". *Philosophical Studies*, *169*(1), 27-37. https://doi.org/10.1007/s11098-012-9892-9

Yang, L., Li, J., Cunningham, P., Zhang, Y., Smyth, B., & Dong, R. (2021). Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis. En C. Zong, F. Xia, W. Li & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 306-316). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.26

Zahavi, D. (2025). *Phenomenology: The Basics* (2.ª ed.). Routledge.