Jethro Masís

Vino nuevo en odres viejos: el problema del mundo en los modelos más recientes de inteligencia artificial

Nota: *Una versión preliminar de este texto fue presentada el 26 de diciembre de 2025 en la conferencia de clausura organizada por la Escuela de Filosofía de la Universidad de Costa Rica. Esta publicación se presenta en conjunto con el texto de Sergio Martén Saborío, quien también expuso en dicha conferencia. Ambos textos buscan poner en diálogo El regreso de los modelos de mundo en la inteligencia artificial: discusiones filosóficas.*

1. Exordio

El vertiginoso avance de la inteligencia artificial (IA), impulsado por el ascenso sin precedentes de los modelos generativos de lenguaje de gran escala (LLMs), ha reavivado tanto el asombro como la aprensión de la comunidad global, en general, y de la académica, en particular. Causa asombro, en primer lugar, el hecho de que estos sistemas, capaces de generar lenguaje, imágenes y código con una fluidez que a menudo roza lo indistinguible de la creación humana, parecen estar en el umbral de una nueva forma de inteligencia. Modelos como GPT-4, Claude y Gemini han demostrado capacidades que hace apenas una década parecían pertenecer al ámbito de la ciencia ficción por su capacidad de sostener conversaciones coherentes sobre temas complejos, escribir ensayos académicos, programar software funcional y hasta aprobar exámenes profesionales. Sin embargo, a medida que la euforia inicial da paso a un escrutinio más riguroso (en la economía se habla cada vez más de una *burbuja de la IA* que está a punto de estallar), emerge una verdad incómoda: las limitaciones más profundas de la IA, que los usuarios experimentamos cuando los sistemas inteligentes con los que trabajamos alucinan, no son meramente técnicas, sino fundamentalmente filosóficas. Dicho de otra forma, la reflexión filosófica sobre conceptos fundamentales como mundo deviene ineludible a la hora de plantear posibles soluciones para forjar sistemas inteligente más robustos. Al fin y al cabo, como ha dicho Dennett, «la IA es en gran medida filosofía. La más de las veces se ocupa directamente con preguntas que reconocemos al instante como filosóficas» (1988, 283). Los conceptos fundamentales no deberían simplemente asumirse como si la tradición del pensamiento no nos dotara, en principio, de prejuicios a la hora de recurrir a ellos.

Los nuevos modelos generativos constituyen, sin duda, tecnologías radicalmente novedosas. Con todo, al ser examinadas bajo la lupa de los problemas filosóficos de la ciencia cognitiva y de la historia del diseño de programas inteligentes, caemos en la cuenta de que ciertos problemas fundamentales de la IA clásica (tanto de GOFAI como del conexionismo) no han sido superados enteramente, sino que parecen retornar enmascarados bajo la sofisticación estadística de los nuevos modelos y dejan, en nosotros los usuarios, la sensación de que nos hallamos ante algo genuinamente alienígena. Porque nos hallamos sin duda ante algo similar a la inteligencia, pero

Rev. Filosofía Univ. Costa Rica, LXV (171), 195-201, Enero - Abril 2026 / ISSN: 0034-8252 / EISSN: 2215-5589

no enteramente como aquella de la que estamos dotados nosotros, los seres humanos. Existe la sensación de estar ante algo humano y conocido, pero esta sensación en ocasiones se difumina y nos deja extrañados cuando el chatbot comienza a inventarse datos, da citas de obras inexistentes o comete yerros en los aspectos más nimios. Hay algo tonto en esa supuesta *inteligencia* que nos causa una profunda desconfianza.

En lo que sigue voy a argumentar que los desafíos más apremiantes en la IA contemporánea que gravitan en torno a la búsqueda de modelos del mundo (*World Models*) robustos son, en esencia, odres nuevos para un vino viejo: en efecto, los nuevos modelos parecen redescubrir el viejo problema del mundo. El problema central que pretendo abordar se resume en preguntas como las siguientes: ¿Qué significa para un sistema simular la inteligencia humana sin poder habitar un mundo? ¿Qué puede enseñarnos el retorno del concepto del mundo en los problemas atinentes a la IA generativa? ¿Con qué sistemas al final nos vamos a quedar? ¿Qué impacto van a tener sistemas de lenguaje desencarnados y desmundanizados en nuestras vidas?

2. La alternativa heideggeriana

Los orígenes de la IA heideggeriana se remonta a la década de 1960, cuando el filósofo Hubert Dreyfus ejercía la docencia en el Instituto Tecnológico de Massachusetts. En aquel entonces, los estudiantes de IA que asistían a su seminario sobre Heidegger le informaron que los científicos cognitivos habían tomado las riendas de la cognición y estaban teniendo éxito donde los filósofos habían fracasado estrepitosamente con sus especulaciones metafísicas de escritorio. Dreyfus no se fiaba de la advertencia: un estudio pormenorizado de los reportes de investigación que emanaban del laboratorio de IA lo llevó a concluir que los investigadores de la IA —sin saberlo— estaban trabajando arduamente para convertir la filosofía racionalista en un programa científico. La historia es conocida en su obra *What Computers Can't Do* (1972), las implicaciones filosóficas de la investigación en IA se expusieron de la siguiente manera:

> si la razón puede programarse en una computadora, esto confirmará una comprensión del hombre como un objeto, que los pensadores occidentales han estado buscando a tientas durante dos mil años, pero que solo ahora tienen las herramientas para expresar e implementar. La encarnación de esta intuición cambiará drásticamente nuestra comprensión de nosotros mismos. Si, por otro lado, la inteligencia artificial resultara ser imposible, entonces tendremos que distinguir la razón humana de la artificial, y esto también cambiará radicalmente nuestra visión de nosotros mismos. (1992, 78-79)

La temprana disputa de Dreyfus con la nueva concepción de la razón artificial le permitió lanzar una carrera intelectual muy fructífera al profundizar en cuestiones con implicaciones filosóficas de gran alcance. Dreyfus fue el primer pensador en notar que conceptos como *mundo, agencia, inteligencia, afrontamiento (coping), mente y conocimiento* (por nombrar solo algunos), no podían dejarse a la libre en manos de ingenieros, que los asumían sin tener conciencia de su naturaleza problemática y de su historia milenaria.

Dreyfus seguía en lo suyo a principios de los años 90: hizo publicar una segunda edición de su libro al que intituló más agresivamente *What Computers Still Can't do* (1992) y agregó una larga digresión sobre los sistemas conexionistas, a los que les espetaba una hueste de problemas. Sin embargo, la influencia de Dreyfus cayó en tierra fértil en el laboratorio de IA del MIT. Terry Winograd ha señalado la ironía de que fuera el laboratorio de IA del MIT el que se convirtiera en cuna de la IA heideggeriana:

> fue en el MIT donde Dreyfus formuló por primera vez su crítica, y durante veinte años la atmósfera intelectual en el laboratorio de IA fue abiertamente hostil a reconocer las implicaciones de lo que dijo. Sin embargo, parte del trabajo que ahora se realiza en ese laboratorio parece haber caído bajo la influencia de Heidegger y Dreyfus. (cita de Dreyfus, 1992, xxxi)

El propio Winograd fue el primer investigador de IA de alto nivel en admitir que trabajaba explícitamente bajo la influencia de la fenomenología del mundo desarrollada en *Ser y tiempo* de Heidegger. Winograd pensaba que las

computadoras debían entenderse como parte de un entramado de equipos dentro del cual encontramos nuestro propio ser en el comportamiento inteligente diario. Se necesitaba un cambio de paradigma para diseñar programas basados en la actividad en curso, en lugar de escribir planes basados en la existencia de una sede central de procesamiento que representara el mundo. La nueva concepción era radical por su rechazo rotundo al neurocentrismo: la idea de que la inteligencia era solo una cuestión de procesos cognitivos localizados intracranealmente que interactuaban con sistemas biológicos, químicos y físicos extracraneales. En este sentido, Winograd hablaba de la interacción humano-computador en términos genuinamente existenciales.

El roboticista del MIT Rodney Brooks propuso una arquitectura de subsunción con el fin de establecer un vínculo de la percepción con la acción, incrustando así a los robots concretamente en el mundo. Dicha arquitectura de subsunción podría permitir a los robots usar el mundo como su propio modelo, por medio de la interacción continua con sus sensores en lugar de con un modelo interno del mundo que definía los parámetros de antemano. Brooks estaba convencido de que nociones jerárquicas como pensamiento y razón eran solo metáforas de la inteligencia basadas en el modelo de un computador de von Neumann, que consiste en un procesador (que se supone funcionalmente idéntico a la mente humana) desde el cual se ejecutaban instrucciones que recuperan datos de la unidad de memoria principal. Existía la tentación (a menudo sucumbida) de diseñar programas basados en una noción jerárquica de la inteligencia concebida como planificación exhaustiva, y que planteaba la resolución de problemas bajo un modelo de representación previa. En lugar de diseñar programas que partieran de *mentes*, el nuevo paradigma en ciernes concibió su tarea como la de modelar la propia actividad de afrontamiento (*coping*) en un mundo donde se encarna el comportamiento inteligente. En contraste con la hipótesis de los sistemas físico-simbólicos defendida en la década de 1970 por Newell y Simon, este nuevo paradigma se opuso a la opinión de que la computación digital era necesaria y suficiente para el *comportamiento inteligente general*. El punto de Brooks era eliminar por completo la idea de una unidad de representación central para construir sistemas autónomos, cuyo repertorio de acciones inteligentes es solo el resultado de una serie de comportamientos competentes, y por lo tanto no depende principalmente del diseño mental jerárquico y descendente.

Philip Agre, el discípulo más prominente de Brooks por aquel entonces, publicó un libro de profundas implicaciones filosóficas: *Computation and Human Experience* (1997). Su principal argumento era que se necesitaba una práctica técnica crítica que incluyera la reflexión sobre el lenguaje y la historia, las ideas y las instituciones, como parte del trabajo técnico mismo. La relevancia de la crítica filosófica debería volverse fundamental dentro del trabajo tecnológico. El punto no era invocar la filosofía como una suerte de autoridad exógena para suplantar los métodos técnicos, sino expandir los horizontes de la investigación computacional, principalmente porque las ideas de la IA tienen sus raíces genealógicas en ideas filosóficas. Del mismo modo, se podría decir que la mayoría de los impasses y dificultades que encontraron los investigadores de la IA se derivan de tensiones internas en los sistemas filosóficos subyacentes, a veces sin que los investigadores fueran conscientes de ellas. En cierto modo, los practicantes de la IA también deberían convertirse en filósofos dotados de una sofisticada destreza teórica. Los ingenieros deberían volverse como el propio Agre, un ávido lector de Wittgenstein, Husserl y Heidegger, sin dejar de ser un talentoso científico computacional.

En resumen, la expresión heideggeriana *estar-en-el-mundo* fue por cierto tiempo la nueva moda en algunos círculos de la IA, y prometía romper el nudo gordiano de los problemas más esquivos de la cognición.

La herencia heideggeriana —a través de Dreyfus— del nuevo paradigma apenas es materia de discusión, aunque Brooks se apresuró a señalar que su enfoque no se basaba en la filosofía alemana o, para el caso, en la obra de Heidegger, aunque tiene ciertas similitudes con el trabajo inspirado por este filósofo alemán. ¿No es sorprendente ver al roboticista más prominente del MIT invocando el pensamiento de Heidegger? El

caso de Brooks es el de un maestro que cae bajo la influencia de su notable discípulo, Phil Agre. La principal contribución de Dreyfus fue mostrar cómo la filosofía juega un papel fundamental al tratar con cuestiones que giran en torno a la definición misma de la cognición humana. El cambio de paradigma heideggeriano buscaba abandonar la idea previa (mal concebida) del mundo típicamente sostenida por los científicos cognitivos como representación mental intracraneal, para reemplazarla con una concepción del mundo en la que las cosas se perciben y se trata con ellas en términos de las posibilidades de acción que ofrecen. La IA heideggeriana fue el proyecto de comprender los horizontes del comportamiento inteligente humano más allá de las limitaciones del cognitivismo: la idea —para usar la irónica definición de Putnam (1988)— de que el ser humano es solo una computadora que resulta estar hecha de carne y hueso.

3. El regreso actual de la noción de un *modelo de mundo*

Si ponemos los pies en la situación actual de la IA nos encontramos con la revolución del machine learning, iniciada con el éxito de las redes neuronales convolucionales en visión computacional y acelerada dramáticamente con la arquitectura Transformer y los modelos de lenguaje a gran escala; modelos que han cambiado radicalmente el panorama de la IA en nuestros días. Estos modelos, entrenados en cantidades masivas de datos textuales y visuales mediante técnicas de aprendizaje auto-supervisado, no operan sobre símbolos explícitos y reglas lógicas, sino sobre patrones estadísticos en espacios vectoriales de alta dimensión. Podría pensarse que estos modelos han superado las críticas de Dreyfus al no depender de reglas pre-programadas ni de bases de conocimiento manualmente codificadas. Aprenden de la experiencia (aunque sea experiencia textual), exhiben una forma de generalización y pueden manejar la ambigüedad y el contexto de maneras que eludieron a GOFAI y al conexionismo. Pero ¿han resuelto realmente el problema del mundo? Argumento que solo lo han reformulado en un nuevo y más sutil lenguaje técnico, y que los viejos fantasmas filosóficos retornan con nueva urgencia.

Los Transformers, la arquitectura detrás de modelos como GPT, BERT, Claude y Gemini son el ejemplo paradigmático. Su éxito se basa en la hipótesis de que la tarea de predecir la siguiente palabra (o *token*) en una secuencia, si se realiza a una escala suficientemente masiva (billones de tokens de texto, miles de millones de parámetros), puede llevar a la emergencia de una comprensión profunda del lenguaje y, por extensión, del mundo que el lenguaje describe. Sin embargo, una creciente cantidad de evidencia sugiere que, si bien estos modelos son extraordinariamente buenos para capturar y reproducir correlaciones estadísticas complejas, caen en problemas a la hora de exhibir un conocimiento causal, estructurado y composicional del mundo. Son modelos que carecen de habilidades cognitivas esenciales como el razonamiento abstracto, la comprensión contextual, la comprensión causal y el juicio matizado. Su aparente razonamiento es a menudo una apuesta, una simulación que se desmorona cuando se enfrenta a tareas que requieren una verdadera comprensión del mundo subyacente, especialmente cuando estas tareas difieren de la distribución de entrenamiento.

Como usuarios de estos sistemas ya nos hemos enterado de que no son enteramente confiables porque suelen fallar estrepitosamente. ¿A qué se deben estos disparates? Según personas tan autorizadas como Yann LeCun de Meta y Demis Hassabis de Google Deep Mind el problema radica en que los sistemas generativos carecen de un modelo de mundo que les permita contar con la robustez necesaria para una cognición más certera. La idea de un modelo de mundo se remonta al concepto de «modelos a pequeña escala» propuesto por Kenneth Craik en *The Nature of Explanation* (1943). Según Craik, un organismo inteligente debe poseer un modelo interno, predictivo y simulable de su entorno:

> Si el organismo lleva un «modelo a pequeña escala» de la realidad externa y de sus propias acciones posibles dentro de su cabeza, es capaz de probar varias alternativas, concluir cuál es la mejor de ellas, reaccionar

> a situaciones futuras antes de que surjan, utilizar el conocimiento de eventos pasados al tratar con el presente y el futuro, y en todos los sentidos reaccionar de una manera mucho más completa, segura y competente a las emergencias que enfrenta. (1943, 51)

Se especula que un modelo del mundo robusto permitirá eventualmente a la IA razonar sobre causa y efecto, planificar a largo plazo en entornos novedosos y generalizar a situaciones que no ha encontrado explícitamente en su entrenamiento. Para un agente robótico, esto podría ser un modelo de la física de los objetos (cómo caen, ruedan, se rompen). Para un agente de lenguaje, podría ser un modelo de las intenciones, creencias y normas sociales de los humanos. Para un sistema de conducción autónoma, sería un modelo de la dinámica del tráfico, el comportamiento de otros conductores y peatones, y las leyes físicas del movimiento vehicular.

Un estudio reciente particularmente revelador (Vafa et al. 2025) utilizó «sondas de sesgo inductivo» (*inductive bias probes*) para determinar si un modelo fundacional, entrenado para predecir trayectorias planetarias, había aprendido la mecánica newtoniana subyacente. La pregunta era: ¿ha descubierto el modelo la ley de la gravitación universal, o simplemente ha memorizado patrones específicos? Los resultados fueron aleccionadores: aunque el modelo predecía las órbitas con alta precisión dentro de la distribución de entrenamiento, el modelo interno de la física que había desarrollado era una ley matemáticamente sin sentido que no se parecía en nada a la ley de Newton. Cuando se le pedía generalizar a configuraciones fuera de la distribución (por ejemplo, sistemas con más planetas o diferentes masas), fallaba dramáticamente. Solo los modelos que incorporaban explícitamente el sesgo inductivo de la mecánica newtoniana lograban capturar la estructura causal subyacente.

Los sistemas artificiales pueden tener un rendimiento casi perfecto en la predicción del siguiente *token* y, sin embargo, parece que carecen de un modelo causal del mundo, lo cual quiere decir que el sistema puede simular superficialmente el mundo sin haberlo comprendido estructuralmente. Esto demuestra de manera concluyente que la simple predicción estadística deviene una métrica frágil e insuficiente para evaluar la comprensión genuina. Salta a la vista que términos filosóficos de raigambre hermenéutica como *comprensión y sentido* no deben dejarse de lado en esta nueva aventura de dotar a los sistemas de inteligencia artificial generativa de un modelo de mundo.

4. El valle inquietante en los modelos generativos

La experiencia al interactuar con estos sistemas de la IA generativa está lejos de ser fluida. Muchas veces se experimenta una disonancia persistente. El problema central acá es la paradoja que reside en el corazón de la interacción humano-computador: una tensión profunda e inquietante entre las capacidades sobrehumanas de la tecnología y sus fallos simultáneos, a menudo cómicos, en dominios que los seres humanos encuentran triviales. Esto no es simplemente una cuestión de errores o fallos ocasionales en un sistema complejo; es una característica fundamental y definitoria de la tecnología en su forma actual. Este perfil de rendimiento errático provoca una forma única y potente de frustración en el usuario, una que se siente cualitativamente diferente de la molestia de lidiar con un *software* con errores o una interfaz mal diseñada.

Esta forma de frustración, novedosa en su escala y carácter, tiene profundas raíces en la historia de la interacción humano-computador. El programa ELIZA de Joseph Weizenbaum de 1966 ofrece una parábola fundante. ELIZA, un simple *chatbot* de coincidencia de patrones diseñado para imitar a un psicoterapeuta rogeriano, provocó una reacción emocional en algunas personas que interactuaban con el programa (Weizenbaum 1966). Weizenbaum se horrorizó al ver a sus colegas, que conocían la naturaleza mecanicista del programa, confiando en él como si fuera un ser existente. La posterior reacción negativa contra ELIZA surgió cuando su destreza conversacional se reveló como una forma de mímica, un truco de salón en lugar de un pensamiento genuino. Esta decepción es

como aquella en la que caemos cuando se nos explica un truco de magia.

El punto de inflexión llegó con la arquitectura Transformer que reemplazó el procesamiento del lenguaje paso a paso con un mecanismo global de auto-atención, permitiendo a los modelos construir representaciones asociativas de alta dimensión a través de secuencias completas simultáneamente (Vaswani et al. 2017). Esto produce capacidades que se sienten menos como una mímica inteligente y más como una cognición alienígena. Los mapas asociativos construidos por máquinas son abstracciones estadísticas destiladas de vastos *corpus* de texto, ponderadas por patrones de co-ocurrencia en lugar de la experiencia vivida, la fundamentación sensorial y el contexto encarnado que dan forma al pensamiento humano Para un Transformer, el lenguaje es la totalidad de su mundo, una topología de relaciones divorciada de las limitaciones encarnadas de la vida humana. Es por eso que los científicos cognitivos y los eticistas describen la IA moderna como cognitivamente alienígena, poderosa precisamente porque no piensa como nosotros (Sandini, Sciutti, & Morasso 2024).

Estas dos perspectivas —el alienígena en la máquina y el humano en el modelo— son dos caras de la misma moneda compleja. Un camino productivo a seguir requiere un marco que tenga en cuenta el estado único del Transformer como un *alienígena en el espejo*, un sistema que refleja nuestro lenguaje y patrones cognitivos a través de un proceso fundamentalmente no humano. La disonancia cognitiva generada por esta tecnología se intensifica en proporción a su capacidad percibida. La decepción con ELIZA fue simple porque el sistema era simple. La IA generativa es un paso radical más allá: un socio activo que genera contenido. En consecuencia, la frustración que engendra no se trata de descubrir un truco, sino de lidiar con una entidad que es simultáneamente más capaz y más incomprensible que cualquier tecnología anterior. El problema ha pasado de ser técnico, en relación con el rendimiento del sistema, a ser ontológico, en relación con la naturaleza misma del sistema. Un error en el *software* tradicional es un fallo de su diseño. Los fallos de la IA generativa, sin embargo, a menudo se perciben como fallos de comprensión o intención, lo que obliga al usuario a preguntar no ¿*Qué es esta cosa*?.

Pero ¿qué se concluye de todo esto? En primer lugar, que la urgencia de dotar a los programas generativos de un modelo de mundo constituye, de cierta forma, una admisión tácita de que la predicción estadística por sí sola es insuficiente para generar tal modelo. Los sistemas artificiales generativos andan a tientas, especulan con datos, pero no comprenden absolutamente nada. La necesidad de incorporar conocimiento causal, estructurado y composicional es, en esencia, la necesidad de un *mundo* —un marco de referencia coherente que permita la generalización, la planificación y el razonamiento. Sin embargo, siempre está la tentación de intentar objetivar el *estar-en-el-mundo*: es decir, de hacer explícito lo que funciona como trasfondo silente, no llamativo. Por ello, quienes abogan por un modelo de mundo causal pueden ya estar operando bajo el embrujo de esta abstracción. Parece que la cognición humana, al menos, está fundamentalmente incrustada (*embedded*) en una estructura de posibilidades de acción (*affordances*), pre-teóricas, pre-proposicionales. ¿Pre-causales también?

Así que queda la pregunta: ¿se resolverá el entuerto dotando a los programas generativos de un modelo de mundo que le permita entender las relaciones causales de los eventos que acaecen en el mundo real? Seguramente, un modelo de mundo semejante hará que los sistemas generativos sean más robustos y que alucinen en mucho menor medida. Empero, esto supone que los seres humanos estamos dotados de un modelo de mundo semejante, lo cual nos devuelve a una teoría computacional de la mente de vieja data. Precisamente el tipo de teoría que la alternativa heideggeriana quiso desmontar. En arreglo con la filosofía fenomenológica, no solo los modelos generativos carecen de un modelo de mundo, sino que también nosotros mismos. Un modelo causal de mundo —podría afirmarse heideggerianamente— debe pasar primero por lo-a-la-mano. *Modelo de mundo* es así una expresión muy desafortunada.

Pero eso es lo de menos. Estamos ante un monstruito que habla como nosotros, dibuja como nosotros, crea historias como nosotros y aparenta

ser nosotros. Sin embargo, eso no es nosotros. El problema de las nuevas tecnologías en IA es que dan la impresión de ser nosotros sin serlo. Esta disonancia entre la apariencia de la inteligencia humana y la naturaleza de su mecanismo evoca poderosamente el concepto de uncanny valley (valle de lo inquietante). Originalmente acuñado en la robótica por Masahiro Mori (1979/2012), el valle de lo inquietante o siniestro describe la repulsión que sentimos hacia los robots que se parecen casi perfectamente a los humanos, pero no del todo. En ese punto de casi-perfección, los pequeños fallos se magnifican, transformando la familiaridad en una profunda inquietud.

Me parece que con las nuevas tecnologías de lenguaje a gran escala, este fenómeno se traslada del ámbito visual y físico al ámbito cognitivo y existencial. Ya no es la apariencia física lo que nos perturba, sino la mímesis casi perfecta de la producción de nuestro Dasein —el lenguaje, la creatividad, el razonamiento— que emana de una fuente fundamentalmente acausal y desmundanizada. La IA generativa se sitúa en lo que podría llamarse un valle de lo inquietante ontológico. Algo que valdría la pena explorar en otro lugar.

Referencias

Agre, Philip. 1997. *Computation and Human Experience*. Cambridge University Press.

Craik, Kenneth. 1943. *The Nature of Explanation*. Cambridge University Press.

Dreyfus, Hubert. 1992. *What Computers Can't Do: The Limits of Artificial Intelligence*. MIT Press.

Dennett, Daniel. 1988. «When Philosophers Encounter Artificial Intelligence». En The Artificial Intelligence Debate. False Starts, Real Foundations, editado por Stephen Graubard, 283-295. Cambridge, MA/London: The MIT Press.

McCarthy, John, Minsky, Marvin, Rochester, Nathaniel, & Shannon, Claude. 2006. «A proposal for the Dartmouth Summer Research Project on Artificial Intelligence», August 31, 1955. *AI Magazine*, 27, no. 4.

Mori, Masahiro. 1970/2012. «The Uncanny Valley». Traducido por Karl F. MacDorman & Norri Kageki. *IEEE Robotics & Automation Magazine*, 19, no. 2, 98-100.

Putnam, Hilary. 1988. «Artificial Intelligence: Much Ado About Not Very Much». *Daedalus*, 117, no. 1, 269-281.

Sandini, Giulio, Sciutti, Alessandra, & Morasso, Pietro. 2024. «Cognitively alien AI: A challenge for human-robot interaction». *Frontiers in Robotics and AI*, no. 11, 1345678.

Vaswani, Ashish, Shazeer, Noam, Parmar, Nikki, Uszkoreit, Jacob, Jones, Llion, Gomez, Aidan, Kaiser, Lukasz & Polosukhin, Illia. 2017. «Attention is all you need». *31st Conference on Neural Information Processing Systems* (NIPS 2017), Long Beach, CA, USA. https://arxiv.org/pdf/1706.03762

Vafa, Keyon, Chen, Justin, Rambachan, Ashesh, Kleinberg, Jon & Mullainathan, Sendhil. 2024. «Evaluating the world model implicit in a generative model: Formalizing world-model coherence using Myhill-Nerode theory». arXiv preprint arXiv:2406.03689

Vafa, Keyon, Chang, Peter, Rambachan, Ashesh & Mullainathan Sendhil. 2025. «What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models». arXiv preprint arXiv:2507.06952.

Weizenbaum, Joseph. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9, no. 1, 36-45.

Jethro Masís (jethro.masis@ucr.ac.cr) es profesor catedrático en la Escuela de Filosofía de la Universidad de Costa Rica, San José, Costa Rica, donde dirige el Instituto de Investigaciones Filosóficas y el Programa de Posgrado en Filosofía. Se doctoró en la Universidad de Wurzburgo, Alemania, con la tesis *The Primacy of Phenomenology Over Cognitivism. Towards a Critique of the Computational Theory of Mind* bajo la tutela de Karl-Heinz Lembeck.

Recibido: 3 de diciembre, 2025.
Aprobado: 12 de diciembre, 2025.
DOI: 10.15517/revfil.2026.2022