Consideraciones analíticas

Margaret Masterman, su crítica a la noción de paradigma y su invisibilización
María Carolina Álvarez Puerta

El papel de las heurísticas en el desarrollo de una Inteligencia Artificial General
Luz Marina Barreto

El *Dilema de Benacerraf* como una objeción al Platonismo tradicional en matemática
Ricardo Da Silva y Solani Rivas

Fundamentos técnicos y dilemas éticos en los modelos de lenguaje natural en las Inteligencias Artificiales
Nahir Hurtado

An interpretation of the I/Me distinction in G. H. Mead's "The definition of the psychical" and H. Joas' criticism
Leopoldo Márquez Velasco

¿Hay matemática más allá de la lógica?
Edgar Valenzuela

U.C.V EPISTEME NS ENERO - JUNIO 45 N°1 2025

EPISTEME NS

Revista del Instituto de Filosofía

ISSN: 07984324 Dep. Leg.: pp198102DF423

45

Enero - Junio

N°1

2025

Universidad Central de Venezuela
FACULTAD DE HUMANIDADES Y EDUCACIÓN

Revista Semestral ISSN: 0798–4324 Depósito Legal: pp198102DF423

EPISTEME NS

Revista del Instituto de Filosofía

45
Enero - Junio
N° 1
2025

Universidad Central de Venezuela
FACULTAD DE HUMANIDADES Y EDUCACIÓN

© by Instituto de Filosofía
Facultad de Humanidades y Educación
Universidad Central de Venezuela
Publicado en línea en abril 2026
Caracas–Venezuela

EPISTEME NS

EPISTEME NS es una revista de crítica e investigación en filosofía, abierta a todas las corrientes y estilos de pensamiento y a la reflexión en todos los ámbitos del saber filosófico, sin más requisitos que la originalidad, seriedad y rigor argumentativos. Los números salen alternativamente dedicados a temas de Filosofía, Historia de la Filosofía y Filosofía Social (Serie Azul) y de Lógica, Análisis del Lenguaje y Filosofía de las Ciencias (Serie Roja), además de las ediciones especiales (Serie Gris). La Revista, de periodicidad semestral y arbitrada según los procedimientos *ad usum*, es una publicación de circulación internacional indexada en *The Philosopher Index, Latindex* y *Clase*, y registrada en *Ulrich´s International Periodical Directory*; también aparece en versión electrónica en las plataformas SciELO (www.scielo.org.ve) y Saber UCV (www.saber.ucv.ve). Todas las colaboraciones deben ser inéditas y ser enviadas por triplicado para ser sometidas a la consideración del Comité de arbitraje. Los trabajos podrán ser de tres tipos: Artículos (no deben exceder de 10.000 palabras), Notas y Discusiones (5.000 palabras) y Reseñas. Los Artículos y las Notas y Discusiones deben ser presentados en cuartilla tamaño carta a doble espacio con notas a pie de página (también pueden ser enviados en *attachment* a las siguientes direcciones electrónicas: revista.epistemens@gmail.com y/o nnunezm@gmail.com y deberán estar acompañados de sendos resúmenes, en inglés y en castellano, de una extensión no mayor de diez (10) líneas, acompañados con tres palabras claves. Para cualquier uso posterior de los materiales publicados en EPISTEME NS, es necesaria la autorización escrita del Editor de la Revista.

EPISTEME NS/Revista del Instituto de Filosofía, Vol. 45, (enero-junio, 2025). Caracas: Ediciones de la Facultad de Humanidades y Educación, 1981.

V./ 22 cm.
2 veces al año.
Continúa: Episteme NS
Título de cubierta.
Director –fundador Anuario Episteme: 1957 – Juan David García Bacca
Director–fundador Episteme NS: 1981 – Juan Nuño
Directora: diciembre 2021 – María Carolina Álvarez
ISSN: 0798–4324
1. Filosofía – Publicaciones periódicas.
Depósito Legal: pp198102DF423

EPISTEME NS

Semestral

Director Fundador: Juan David García Bacca†

Directora: María Carolina Álvarez Puerta

Comité Editorial:

Editora: Nancy Núñez

Secretaría de Redacción: María Carolina Álvarez Puerta

Secretaría de Administración: Benjamín Sánchez

Consejo Consultivo: Omar Astorga (UCV), Francisco Bravo (UCV) †, Luz M. Barreto (UCV), Sabine Knabenschuh de Porta (LUZ), Carlos Kohn (UCV), Vincenzo P. Lo Monaco, (UCV), Nancy Núñez (UCV), Benjamín Sánchez (UCV), Gustavo Sarmiento (USB), Fabiola Vethencourt (UCV), Corina Yoris (UCAB)

Consejeros Internacionales: J. J. Acero (UGR), Francisco Ayala (UCLA), Mario Bunge (McGILL) †, Hugo Calello (UBA), Marisa Kohn de Beker (UCV) †, Pedro Lluberes (USB), M. Reyes Mate (CSIC), Jesús Mosterín (CSIC) †, Ulises Moulines (LMU), José M. Rosales (UMA), Giulio F. Pagallo (Unipd) †.

Responsable de este número: Ricardo Da Silva

Asistentes: María Alejandra Pérez y Sebastián Bogoljubskij

Diseño de Portada: Lic. José Bifano

Arte Final: María Alejandra Pérez

EPISTEME NS

Volumen 45, N° 1, 2025

Contenido
Consideraciones analíticas

Presentación
DA SILVA, R.: Presentación..... *i*

Artículos *Págs.*

ÁLVAREZ, M.: Margaret Masterman, su crítica a la noción de paradigma y su invisibilización *(Margaret Masterman, her criticism of the notion of paradigm and its invisibility)*..... 01

BARRETO, L.: El papel de las heurísticas en el desarrollo de una inteligencia artificial general *(Heuristic's role on the development of a general artificial intelligence)*..... 25

DA SILVA, R. Y RIVAS, S.: El *Dilema de Benacerraf* como una objeción al Platonismo tradicional en matemática *(Benacerraf's Dilemma as an objection to traditional Platonism in mathematics)*..... 49

HURTADO, N..: Fundamentos técnicos y dilemas éticos en los modelos de lenguaje natural en las Inteligencias Artificiales *(Technical foundatinos and ethical dilemmas of natural language in Artificial Intelligence)*..... 81

Marquez, L.: An interpretation of the I/Me distinction in G. H. Mead's "The definition of the psychical" and H. Joas' criticism *(Una interpretación de la distinción Yo/Mí en "The definition of the psychical" de G. H. Mead y la crítica de H. Joas)*..... 95

Valenzuela, E.: ¿Hay matemática más allá de la lógica? (Is *there mathematics beyond logic?*)..... 121

Notas y discusiones

Blanco, F.: Notas introductorias al concepto de «Lógica» (*Introductory notes on the concept of «Logic»*)..... 137

Chávez, M.: ¿Es la *existencia* una propiedad, o una *modalidad*? (*Is existence a property, or a modality?*)..... 153

Recensiones

Bogoljubskij, S.: Posy, C - Intuicionismo matemático (*Mathematical Intuitionism)*..... 159

μiscelánea 165

EPISTEME NS, VOL. 45, N°1, 2025, pp. i-ii

Ricardo J. Da Silva Araujo

PRESENTACIÓN

Bajo la premisa de que la unidad filosófica no exige uniformidad temática, este número ofrece un panorama diverso donde la multiplicidad de problemas abordados cobra sentido bajo el rigor que el análisis filosófico confiere a la investigación académica actual.

Aunque es factible trazar conexiones entre ciertos núcleos temáticos que se comentarán más adelante, el verdadero denominador común de esta entrega es el compromiso de presentar los problemas desde una visión argumentativa que los sustente. De esta manera, cada texto aquí reunido destaca por la claridad conceptual empleada para desarrollar sus hipótesis, ofreciendo al lector un ejercicio de pensamiento crítico y estructurado.

Aunque el índice obedece a un orden lexicográfico, podemos reagrupar los artículos según ciertas similitudes. Los primeros serían aquellos que reflexionan directamente sobre la obra de un autor. Así, María Carolina Álvarez se propone revalorizar y visibilizar las críticas que Margaret Masterman le hiciese a la noción de "paradigma" de Thomas Kuhn en la edición de 1962 de *La estructura de las revoluciones científicas*; por su parte, Leopoldo Márquez propone una nueva lectura de la subjetividad en el trabajo temprano de George H. Mead, heredada luego por su obra madura, tratando así de sustentar una exégesis continuista sobre el pensamiento del autor.

En otra dirección, encontramos trabajos que giran en torno al tema de la inteligencia artificial, tales como su capacidad para capturar la semántica del lenguaje y los desafíos éticos que presentan estas nuevas herramientas. Luz Marina Barreto busca examinar el desarrollo de algunos métodos algorítmicos de procesamiento de lenguaje y explicar, especialmente, las limitaciones de los grandes modelos de lenguaje (LLM) para actuar como modelos de razonamiento sin la retroalimentación continua de anotadores humanos. Nahir Hurtado, por su parte, hace un recorrido elucidatorio por los mecanismos que han hecho

posible el surgimiento de modelos como GPT, pasando así revista a las redes neuronales y las arquitecturas transformers. Ambas autoras concluyen remarcando el énfasis en las consideraciones éticas y en las exigencias de legislación sobre el uso de estas nuevas tecnologías.

El último par de artículos trata sobre cuestiones inherentes a la filosofía de la matemática. Edgar Valenzuela revisa la persistencia de la lógica en el quehacer matemático a través del estudio de sistemas lógicos no clásicos, defendiendo, así, una tesis pluralista en contraposición al monismo lógico tradicional. A su vez, Solani Rivas y Ricardo Da Silva realizan una reconstrucción argumentativa del dilema de Benacerraf en contra del platonismo matemático clásico y presentan la propuesta de Mark Balaguer, el platonismo pleno, como un intento de solución.

El presente número se ve, además, enriquecido con dos notas ensayísticas y una reseña bibliográfica. La primera de las notas, escrita por Francisco Blanco, busca indagar en un concepto de lógica que se reconstruya desde sus compatibilidades e incompatibilidades semánticas con otros conceptos. En este sentido, el material puede servir de apoyo para las primeras sesiones de un curso de lógica en cualquier pregrado de humanidades. La segunda nota, de Marcel Chávez, revisa críticamente la idea de que la existencia es una propiedad de los individuos, es por ello que esboza una interpretación desde la ontología de Duns Escoto, desde la cual se concluye que la existencia no es una propiedad, sino una forma de modalidad. Finalmente, Sebastián Bogoljubskij reseña el texto *Mathematical Intuitionism* de Carl Posy. La obra reseñada forma parte de la colección *Cambridge Elements: The Philosophy of Mathematics* y tiene por objetivo introducir a los lectores en la filosofía intuicionista de la matemática desde consideraciones filosóficas y formales. Bogoljubskij reconoce la claridad analítica en la exposición propuesta por el autor; sin embargo, identifica que algunos capítulos, debido a su reconstrucción formal, no resultarían accesibles para estudiantes que no estén familiarizados con ciertos contenidos de lógica matemática.

Podemos concluir afirmando que la presente edición contribuye a la idea de que la *buena filosofía* —sea lo que sea— se sustenta en argumentos. Que resulte de provecho para todos.

EPISTEME NS, VOL. 45, N° 1, 2025, pp. 81-94

Nahir Hurtado

FUNDAMENTOS TÉCNICOS Y DILEMAS ÉTICOS EN LOS MODELOS DE LENGUAJE NATURAL EN LAS INTELIGENCIAS ARTIFICIALES

Resumen: El siguiente artículo tiene como objetivo mostrar los fundamentos de los generadores de lenguajes naturales dentro de las inteligencias artificiales, desde la elucidación de las *Redes Neuronales* hasta las arquitecturas *Transformers* de los modelos GPT. Explicar los fundamentos técnicos permitirá entender cómo funciona la "comprensión del lenguaje" bajo estas nuevas tecnologías, contrastándolos con los desafíos éticos que presentan estos nuevos avances y el impacto de su uso sin una adecuada supervisión humana.

Palabras clave: Modelo GPT, Redes Neuronales, Redes Transformers, ética de las IA.

TECHNICAL FOUNDATIONS AND ETHICAL DILEMMAS OF NATURAL LANGUAGE MODELS IN ARTIFICIAL INTELLIGENCE

Abstract: The following article has the goal of showing the foundations of natural language generators within artificial intelligence, from the elucidation of Neural Networks to the architectures of Transformers like GPT models. Explaining this technical foundations will allow to understand how "language comprehension" works within this new technologies, contrasting them with the ethical challenges that this new advances bring forth and the impact of their use without adecuate human supervision.

Keywords: GPT model, Neural Network, Trasformer Network, AI ethics.

Recibido 27-01-25 ☼ Aceptado 16-02-25

El procesamiento del lenguaje natural en el dominio de la IA ha avanzado velozmente. Ser testigo y usuario de los modelos de lenguaje avanzado permite ser parte de los hitos notables en la historia de la IA, al observar la rápida progresión que ha llevado la inteligencia artificial desde la máquina de Turing, hasta el desarrollo de los modelos GPT (Generative Pre-trained Transformer), capaces de generar conversaciones tal como lo haría cualquier humano.

Sin duda, este avance tecnológico ha llamado la atención en varios campos, que no competen únicamente al área de Sistemas, entre estos, el campo de la Filosofía del Lenguaje cuyos teóricos han sido fascinados por la generación de lenguaje natural. Es por ello que será indispensable mostrar los elementos esenciales en los fundamentos en los que se apoyan esas arquitecturas de sistemas para procesar el lenguaje natural y las implicaciones éticas que conlleva un adelanto tan significativo.

Comparemos este avance con el uso de una calculadora simple. Hasta hace poco, cuando queríamos ejecutar un algoritmo, teníamos que proporcionar datos precisos para obtener una respuesta única. Por ejemplo, en una calculadora, para sumar dos números solo teníamos que ingresar los números seguidos del símbolo más y el signo igual. Si por error colocábamos un signo de más o errábamos la cifra, tendríamos que empezar de nuevo.

Ahora, comparemos esto con el asistente virtual de cualquier Smartphone, este asistente, impulsado por IA, tiene la capacidad de entender y responder a preguntas en lenguaje natural. Cuando le haces una pregunta como "Siri, ¿puedes decirme el clima de esta semana? No, mejor solo dime el clima de hoy", el asistente utiliza un algoritmo de procesamiento de lenguaje natural para analizar las palabras clave de tu consulta y comprender su significado incluso si está formulado o no con claridad.

El asistente utiliza técnicas para interpretar sintáctica y semánticamente las palabras para descomponer la pregunta y extraer información relevante. Deberá entonces identificar las palabras como "clima", "semana", "no" y "hoy" para entender que se está buscando

información sobre las condiciones climáticas actuales. Pero ¿cómo logra una máquina entender lo que se quiere decir?.

Modelo GPT

El modelo GPT, que en sus siglas en español se traduciría como Transformador Generativo Preentrenado, se fundamenta en una arquitectura generativa que permite diversos modelos de lenguaje. Esta arquitectura es la estructura teórica basada en el diseño de la red neuronal que permite a la máquina "comprender" y reproducir frases de manera coherente al recibir un texto, y a diferencia de procesos similares como la función de autocompletado en los teléfonos móviles, el modelo GPT funciona de forma mucho más sofisticada.[1]

La supuesta capacidad de "comprensión del lenguaje" que exhibe el modelo GPT parece ser casi humana, pero es crucial desentrañar el mecanismo que lo sustenta. En realidad, el modelo funciona anticipando las palabras y completando frases, sustentado en los patrones que ha desarrollado y en lo que puede denominarse como la fase de entrenamiento. Las bases de este proceso son en realidad las llamadas *redes neuronales*, las cuales funcionan como aprendizaje automático que analiza una gran base de datos para generar respuestas estructuradas en esquemas específicos, entendiendo entonces que el proceso de "comprensión" de las IA, no es realmente una comprensión cognitiva, sino procesamiento probabilístico altamente desarrollado.

Aunque pueda parecer que la tecnología GPT tiene un profundo entendimiento del contenido y del contexto, en realidad está haciendo predicciones basadas en datos anteriores. La interacción con GPT se basa en su capacidad para predecir y generar texto coherente, utilizando la información adquirida durante su entrenamiento en lenguaje natural.

1 OPENAI. *Informe técnico de GPT-4*. [Documento en línea]. Disponible: file:///C:/Users/nhurtado/Downloads/GPT-4_Technical_Report.pdf [Consulta: 2024, Mayo 20].

Las redes neuronales

En términos sencillos, una *Red Neuronal* es una interconexión de capas que procesan información de forma jerárquica, capaz de simular los procesos sinápticos humanos para la identificación de patrones complejos. Las Redes permiten que se optimicen los datos internos dentro del entrenamiento, lo que permite mayor probabilidad al dar respuesta efectiva y coherente. Para ejemplificar este funcionamiento podemos imaginar una serie de filtros que nos permitirá identificar, por ejemplo, la figura de un gato: el primer filtro solo permite identificar que hay un cambio de líneas o colores; un segundo filtro es capaz de integrar figuras geométricas y nota que la figura tiene dos triángulos arriba y dos círculos, uno pequeño y uno grande; un tercer filtro más avanzado reconoce e identifica rasgos específicos: los triángulos son orejas y dentro del círculo pequeño identifica dos ojos, una nariz pequeña, un bigote, además de una cola pegada al final del círculo grande. Por último, un cuarto filtro que sintetiza todas las piezas, de tal manera que si encuentra una figura con orejas puntiagudas, ojos, nariz pequeña, bigote y con cola concluye que hay una alta probabilidad de que sea un gato.

Ahora bien, la probabilidad de acertar o no, está intrincadamente ligada a la cantidad de información que tengan dichos filtros, lo que permitirá que si durante el entrenamiento hay un error, se ajuste el "peso" de los datos para que acierte más en su predicción.

De esta forma, la llamada comprensión de lo que es un gato, no es un proceso cognitivo de la máquina, sino una sofisticada proyección estadística que simula el lenguaje humano.[2]

¿Cómo las IA logran comprender el lenguaje natural?

El proceso para crear GPT involucra varias etapas claves en el procesamiento del lenguaje natural. Comienza con la tokenización, donde cada palabra se asigna a un "token" representado por un número binario. Este paso simplifica los datos y sienta las bases para el modelo.

2 *Ibíd.*, p. 7.

El siguiente paso es el llamado *embedding*, que dentro de la inteligencia artificial es el método utilizado para representar frases y palabras como vectores numéricos. Estos vectores numéricos se dan dentro de un espacio multidimensional que representan las relaciones sintácticas y semánticas que tiene cada palabra. Por ejemplo, "perro" se relaciona con "mamífero" y con "ladrido", y la cercanía de estos vectores permite la coincidencia dentro del proceso del procesamiento del lenguaje. Así, estos modelos permiten que la inteligencia artificial comprenda y procese el lenguaje natural de forma más completa.

Supongamos que tenemos nuestro asistente virtual, que ha sido entrenado para clasificar reseñas positivas y negativas de películas. De esta manera, el modelo puede comprender una reseña detectando palabras positivas y negativas en la misma. Para lograr esto, se debe convertir cada palabra en un *token*. Estos vectores son generados mediante algoritmos de aprendizaje automático que analizan grandes cantidades de texto y capturan relaciones entre las palabras. Una vez que la palabra se convierte en un vector dentro del *embedding*, se puede proceder a relacionar las palabras por su cercanía y así el modelo procesa las reseñas de las películas. Por ejemplo, si la reseña contiene la palabra "fantástico", cuyo *embedding* es similar o cercano a la palabra "increíble", el modelo puede interpretar esto como algo positivo. En cambio, si la reseña tiene palabras como "horrible" o "aburrido", puede interpretarla como una reseña negativa.

Un ejemplo similar sería el considerar los vectores de las palabras “gato” y “perro", es probable que estos vectores estén cerca en el espacio del *embedding*, ya que ambos son animales domésticos comunes. En contraste, las palabras “gato” y “mesa” tendrán vectores en el *embedding* más distantes debido a su falta de una relación semántica clara.

Esta propiedad del *embedding* permite realizar operaciones interesantes en el espacio vectorial. Los *embedding* permiten realizar operaciones matemáticas, como sumar y restar vectores, para obtener un resultado, todo esto, como ya comentamos, dentro de un espacio multidimensional, lo que explicaría como las palabras con significados similares tienen coordenadas cercanas. Por ejemplo, es posible realizar operaciones con el vector "animal" y relacionarlo con el vector "perro" y luego sumarlo con el vector "compañero"; así obtendremos

un resultado cercano a la palabra "mascota". De esta forma, logramos que el sistema relacione tanto géneros como conceptos dentro de un contexto determinado mediante la vectorización de cada palabra.

De igual forma, esto no solo funciona con palabras individuales, el *embedding* también permite utilizar frases o documentos completos. Así, en lugar de tener vectores para palabras individuales, se puede promediar el vector de cada palabra en una oración. Esto permite que el modelo comprenda el significado general de una frase.

Otro de los conceptos fundamentales para entender como las IA procesan el lenguaje natural es la noción de *sampling*. Este es un engranaje fundamental que agiliza la información eliminando términos comunes, promoviendo un procesamiento eficiente. Este método ramifica varias respuestas posibles imitando el comportamiento dinámico del habla humana.

El *sampling* sienta sus bases en la producción probabilística de palabras independientes o secuencias de palabras donde cada siguiente elección se basa en la familiaridad con el contexto o iteraciones anteriores, esto basándose en las probabilidades aprendidas durante las sesiones de entrenamiento en la red neuronal.

El *sampling* permite una ventaja técnica en el procesamiento del lenguaje natural que da como resultados la generación de textos creativamente coherentes debido a que estos aspectos se combinan entre sí, pero no se limitan solo a ellos, además las interacciones permiten que el entrenamiento de la red neuronal mejore a partir de las respuestas de los usuarios.

Entonces el *sampling* permite respuestas genéricas que se adaptan a las representaciones que se necesitan en cada momento. La pérdida de información y la posterior reconstrucción contribuyen significativamente a la creatividad del modelo.

Sin embargo, este proceso no está exento de desafíos y uno de los problemas fundamentales es la limitación de memoria en la red neuronal. Después de cierto punto, la IA sufre una pérdida de contexto en secuencias largas, esto se traduce en el desgaste de detalle en el procesamiento de memoria, lo cual es un problema común en las redes neuronales recurrentes y afecta la retención de información a largo

plazo. Este obstáculo resalta la necesidad del sistema para mantener el contexto y recordar detalles durante conversaciones más extensas.

Las redes Transformers: El gran avance

Entender el procesamiento del lenguaje natural no estaría completo sin conocer la evolución que han tenido las redes neuronales tradicionales hasta los modelos más avanzados, como los llamados *Transformers*.[3] Las redes neuronales tradicionales tienen una incapacidad para mantener contextos extensos en textos largos, lo que ha llevado al desarrollo de soluciones más avanzadas. Así nacen las *Redes Transformer*, que son un tipo de arquitectura en el procesamiento del lenguaje natural y que, a diferencia de las redes neuronales tradicionales (RNN), utilizan estructuras de procesamiento paralelo que permiten capturar las relaciones entre las palabras y su secuencia dentro del texto.

En las *Redes Transformer* se utiliza una técnica matemática llamada "atención" que se basa en calcular la relevancia que cada palabra tiene con respecto a todas las demás palabras en la secuencia y dar un mayor peso a ella.

Esto permite que el modelo capture las conexiones a largo plazo y las relaciones de significado entre las palabras, lo cual facilita la comprensión del sentido en el lenguaje natural. La arquitectura *Transformer* es un modelo de red neuronal utilizada para que las inteligencias artificiales entiendan el lenguaje natural en amplios contextos.

Un ejemplo para entender cómo funcionan, sería con el siguiente caso: supongamos que deseamos que una computadora pueda responder preguntas sobre un cuento. Y que, para que sepamos más de este cuento, se descomponen por varias oraciones, cada una con un significado fundamental para comprender la trama general de la historia.

En una *Red Transformer*, cada oración de la historia se representa por un vector numérico dentro de *embedding* que captura el significado semántico de las palabras u oración. "Perro amigable", por ejemplo,

3 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. Y Polosukhin, I. *Attention is all you need*, 2017, [Documento en línea]. Disponible: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

podría asociarse con una incrustación que represente al animal de cuatro patas que ladra.

Luego, la *Red Transformer* utiliza el mecanismo de "atención" para calcular la relevancia de cada palabra con respecto a todas las demás palabras en la historia. Esto se hace comparando los vectores en el *embedding* y asignando un mayor peso a cada vector según su importancia para comprender la historia. Entonces es posible que nunca haya una palabra igual a otra.

Una vez que se ha calculado la "atención", la *Red Transformer* puede crear una representación de cada palabra basada en el contexto, dependiendo de su relación dentro de la narrativa. Esta representación contextual captura el significado de cada término en relación con otras palabras que componen la historia. Finalmente, la IA puede utilizar esta interpretación para responder consultas relacionadas con la narrativa. Por ejemplo, si se le pregunta "¿Qué animal se menciona en la historia?", la IA que utilice esta representación contextualizada podría identificar y responder "de un perro amigable" como la respuesta correcta.

Entrenamientos de Modelos de Lenguaje

Ahora bien, ¿cómo se construye la arquitectura de las redes neuronales? Las mismas pueden ser recurrentes o incluso más avanzadas, como las *Redes Transformer*. Este sistema consta de redes conectadas a través de un proceso de información que aprende y reconoce datos de una gran base de información adquirida. El primer paso es la recopilación de datos etiquetados. Para explicarlo podemos colocar el siguiente ejemplo, debemos recopilar reseñas de películas y su correspondiente etiqueta buena o mala. Estos datos se utilizan para enseñar a la *redneuronal* cómo asociar una reseña con una específica orientación positiva o negativa.

Luego, se procesan los datos, lo que implica limpiar el texto, eliminar la puntuación y convertir las palabras en vectores matemáticos que tal como mencionamos están dentro de un espacio multidimensional, es decir, la construcción de los *embedding*. Estas incrustaciones capturan el significado semántico de las palabras y permiten que la red neuronal trabaje con ellas.

Una vez que la arquitectura está lista, se inicia el entrenamiento de la red neuronal utilizando un conjunto de datos etiquetados. Este entrenamiento ajusta los pesos y conexiones internas para minimizar las diferencias entre las predicciones y las etiquetas reales. Esto es posible a través de algoritmos de optimización, también llamados descensos de gradiente, que actualizan los errores cometidos.

A medida que el entrenamiento avanza y mejora la capacidad para crear significado en lenguaje natural, es decir, mediante la aproximación de los vectores en el *embedding*, la IA o el modelo puede aprender a reconocer ciertas palabras o combinaciones de palabras asociadas con una valoración positiva o negativa, permitiendo así entrenar un sistema. Una vez entrenado, se puede utilizar para predecir, según una base de datos, por ejemplo, cuáles de las películas más recientes tendrán más éxito. De esta forma, la red tomará lo que ha aprendido en el entrenamiento y hará predicciones de las reseñas positivas o negativas.

Los grandes desafíos de las IA

La inteligencia artificial ha logrado grandes desafíos en la comprensión del lenguaje natural, aunque esto no la exenta de los inconvenientes que enfrenta. Uno de los más importantes es la limitación de memoria en las redes neuronales. Después de un tiempo, estas máquinas tienden a olvidar los detalles; es decir, existe una frontera en las redes recurrentes que afecta el almacenamiento a largo plazo. Adicionalmente, la IA se enfrenta a varios desafíos al intentar capturar significados del lenguaje en un contexto pragmático. Para comprender estas complicaciones, supongamos que queremos que la IA comprenda la frase "aquí hay gato encerrado".

En esta frase las palabras pueden tener muchos significados diferentes según la situación pragmática. Por ejemplo, la palabra “encerrado” puede hacer referencia a algo contenido o represado, pero el significado real de la frase lo que señala es que estamos próximos a ser engañados. Hay grandes complicaciones para que la inteligencia artificial sea capaz de comprender el significado correcto en la situación lingüística en el que se utiliza ciertas frases o palabras.

Otro desafío es el cambio de ciertos términos. Las personas pueden explicar las mismas ideas de diferentes maneras. Por ejemplo, la

frase "El gato está encerrado" pareciese tener el mismo significado lingüístico que la primera frase. La inteligencia artificial debe ser capaz de detectar estos cambios y entender que están describiendo distintas situaciones.

Por otro lado, el lenguaje natural puede contener contenidos metafóricos, sarcásticos o irónicos, complicando aún más la tarea de gestionar los significados. Por ejemplo, la frase "Eso es genial" podría significar lo opuesto en un tono sarcástico. Existe un reto significativo en la IA para ser capaz de reconocer estas sutilezas y comprender lo que realmente quiere decir el usuario.

Finalmente, el conocimiento y la experiencia son fundamentales para comprender el lenguaje natural. La gente tiene un mundo tan vasto en sus marcos lingüísticos que solo conociendo los mismos y la fuerza ilocucionaria[4] con las que se trasmiten puede permitir el conocimiento de significados. Una frase tan simple como "Mamá, voy a la tienda a comprar leche", puede suponer un aviso o una forma de solicitar dinero para comprar leche. Las IA necesitan tener una gran cantidad de información para capturar mejor estos significados.

Este avance se ve acentuado por desafíos éticos, especialmente en lo que respecta a la desinformación generada en el uso de estos generadores de lenguaje. La exigencia de ajustes manuales para corregir el problema enfatiza la importancia de mantener un control humano en el desarrollo y despliegue de estas tecnologías para garantizar un comportamiento responsable y ético.

El desafío ético

La inteligencia artificial enfrenta varios desafíos éticos al intentar entender los significados en lenguaje natural. Para comprender estos desafíos, cabe mencionar la investigación realizada por Catrin Misselhorn[5] y otros autores que estudian las cuestiones básicas de la Ética de las Máquinas, estos autores destacan que hay tres desafíos éticos fundamentales que enfrentan las IA.

4 Austin, J. L. *Cómo hacer cosas con palabras: Palabras y acciones* (Trad. G. Carrió y E. Rabossi). Barcelona, Editorial Paidós. 1982.

5 Misselhorn, C., *Emotional Machines: Perspectives from Affective Computing and Emotional Human-Machine Interaction.* Wiesbaden, Springer Fachmedien Wiesbaden. 2023.

1. Sesgos y prejuicios: los algoritmos de IA pueden aprender de una gran cantidad de datos, pero debido a que los datos están sesgados, la IA puede amplificar perpetuamente ese sesgo. Los datos pueden revelar prejuicios y discriminaciones existentes en la sociedad, lo que puede dar lugar a sesgos y resultados de discriminación en la comprensión del lenguaje natural. De esta forma, al entrenar un modelo de IA con datos que refuercen estereotipos de género, este puede dar lugar a respuestas sexistas, discriminatorias o incluso xenófobas.
2. Privacidad: El siguiente desafío tiene que ver con la seguridad y la protección de los datos personales. Los datos personales pueden ser almacenados por las IA, ya que los modelos artificiales suelen requerir grandes cantidades de información, como conversaciones de chat, foros, cache, historiales, etc. Esto hace que toda la información que le proporcionamos y que el sistema recopila pueda ser utilizada para poder predecir o capturar patrones. Esto genera preocupaciones sobre la privacidad y el abuso de los datos. Si los datos utilizados no se almacenan de forma segura, pueden producirse fugas de información que no permitan su uso indebido de la persona de los datos.
3. Transparencia y explicabilidad: Es importante entender cómo llegan a generarse las IA. Sin duda, una de las mayores brechas éticas es la poca capacidad que tenemos para entender cómo funciona el algoritmo y la forma cómo el mismo toma decisiones en la generación del contenido solicitado.

Un ejemplo para ilustrar estos dilemas éticos del uso de algoritmo de las IA puede ser en la contratación de personal. La creación de dichos algoritmos mediante capacitación y datos históricos que indican prejuicios raciales o de género en las decisiones de contratación puede perpetuar estos prejuicios y discriminar a grupos específicos donde los empleadores desconocen incluso esta situación. Esto puede tener un impacto negativo en la diversidad y la igualdad de oportunidades en el lugar de trabajo.

Los valores en la IA

Dada estas implicaciones éticas, es indispensable que los desarrolladores y sus usuarios sean conscientes de los valores que reflejan estos sistemas tecnológicos.

Latorre[6] en concordancia con lo planteado por Misselhorn,[7] identifica tres criterios importantes a considerar al diseñar e implementar programas de las IA:

1. La Seguridad: Los Sistemas de IA deben garantizar la seguridad de los usuarios y no deben permitir generar daños emocionales. Este valor es fundamental para la ética de la IA. Esto incluye proteger a los humanos de daños físicos, como lesiones o muerte, pero también de daños emocionales, como el estrés o la ansiedad.
2. La Privacidad: Los Sistemas de IA no deben recopilar ni utilizar datos personales sin el consentimiento de las personas. Los Sistemas deben ser diseñados de manera que no haya perturbación de la privacidad. Esto incluye proteger los datos personales de las personas contra la divulgación no autorizada, así como contra su uso con fines penales.
3. La Equidad: Los Sistemas de IA no deben discriminar a ningún grupo de personas. Este valor pretende garantizar que los Sistemas IA se utilicen de manera óptima y equitativa. Los sistemas de IA deben diseñarse de manera que no nieguen a las personas por motivos de raza, casta, religión, preferencia o cualquier otra protección especial.

Los sistemas de IA pueden utilizarse para controlar la autonomía de las personas o para tomar decisiones sobre sus vidas. Es por ello, que autores como Misselhorn y Poljanšek[8] insisten en la importancia de tener mecanismos de libertad en un mundo cada vez más controlado por la IA. Para lograr esto es necesario explicaciones transparentes del funcionamiento de estos sistemas para así permitir la toma de decisiones informadas sobre cómo implementar las mismas. Los Sistemas

6 Latorre, J. I. *Ética para máquinas*. Barcelona, Editorial Ariel. 2019.
7 Misselhorn, C. *Grundfragen der Maschinenethik*. Ditzingen, Springer Nature, 2018.
8 Cf. Misselhorn, C. *Emotional Machines: Perspectives*...

de IA deben ser claros en su funcionamiento, fomentando políticas que sean justas y equitativas.

La IA es un arma poderosa que puede ayudar al desarrollo de la humanidad o causar un gran daño a las personas. Es imperativo que los desafíos éticos que plantea la IA se aborden con firmeza para evitar degradación de la condición humana.

Es responsabilidad de sus creadores, gestores y usuarios afrontar los desafíos éticos que incluyan desarrollar y seguir nuevos marcos legales y regulatorios del uso de los dispositivos, conocer las limitaciones o sesgos que pueden presentar la IA y dar a conocer y entender el uso de los datos recopilados. Reglamentar nuevas directrices de uso pasaría por aclarar quién es responsable de las acciones de las máquinas hasta gestionar las limitaciones de uso que se pudieran dar a la IA.

Todo lo anterior solo puede sustentarse bajo los estándares de educación y sensibilización sobre ética de la IA. De ahí que los desarrolladores y usuarios de las IA seamos conscientes de los dilemas éticos que ellas plantean y conozcamos cómo funcionan y las consecuencias de su uso.

A modo de síntesis

La capacidad de la inteligencia artificial para captar significado en el lenguaje natural se basa en una combinación de enfoques, desde la representación numérica de palabras hasta el uso de redes neuronales y arquitecturas avanzadas como las llamadas *Redes Transformers*.[9] Conocer estos procesos permite una mayor comprensión del proceso y el funcionamiento de las IA.

Sin duda, aún queda un largo camino por recorrer, el viaje hacia la "comprensión del lenguaje natural" es mucho más complejo del que se ha mostrado, no obstante los avances tecnológicos y las nuevas tecnologías seguirán dando los indicios de cómo entendernos a nosotros mismos.

La IA marca el ritmo de avances notables, pero aún enfrenta desafíos, como la retención a largo plazo y los desafíos éticos. No

9 Vaswani et al. "*Attention Is All You Need*". 2017. [Documento en línea]. Disponible: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

obstante, estos desafíos también presentan oportunidades para una mayor investigación y mejora de estos sistemas tanto para su comprensión como para el manejo correcto de esta tecnología.

Nahir Hurtado
Universidad Central de Venezuela
nahirhurtado@gmail.com