CUADERNOS SALMANTINOS DE FILOSOFÍA
Vol. 52, 2025, 43-58
ISSN: 0210-4857, E-ISSN: 2660-9509

ASPECTOS EPISTEMOLÓGICOS DEL USO DE GRANDES MODELOS EN INTELIGENCIA ARTIFICIAL

EPISTEMOLOGICAL ASPECTS OF THE USE OF LARGE LINGUISTIC MODELS IN ARTIFICIAL INTELLIGENCE

FERNANDO BRONCANO
Doctor en Filosofía
Catedrático de Filosofía emérito
Universidad Carlos III de Madrid
Getafe, Madrid/ España
fernando.broncano@gmail.com
Orcid: 0000-0001-8316-8885

Recibido: 14/09/2024
Revisado: 08/01/2025
Aceptado: 06/02/2025

Resumen: Los dos grandes fenómenos de la era digital que presentan nuevas cuestiones epistemológicas son la extensión de internet en la versión Web 2.0 y la muy reciente aparición de la inteligencia artificial generativa y el aprendizaje profundo que ofrece un uso generalizado de los grandes modelos lingüísticos. La hipótesis en este trabajo es que la epistemología de orientación analítica que ha trabajado sobre el fenómeno de internet lo ha hecho tomando como modelo el testimonio como centro de gravedad de la epistemología social. Este modelo no es el adecuado para los nuevos dispositivos. El argumento es que la autoridad (o falta de autoridad) epistémica de estos ingenios no está basada en algún contrato explícito o implícito de reconocimiento mutuo en los intercambios de información o conocimiento, sino en los amplios procesos sociotécnicos de diseño, formateo y alimentación de datos, entrenamiento dirigido y feedback con un uso social masivo. Ello nos lleva a la necesidad de una epistemología de *agentes híbridos*.

Palabras Clave: Inteligencia artificial generativa; Epistemología de la inteligencia artificial; Mente extendida; Epistemología de internet.

Abstract: The two great phenomena of the digital era that present new epistemological questions are the extension of the Internet in the Web 2.0 version and the very recent emergence of generative artificial intelligence and "deep learning" that offers a generalized use of large linguistic models. The hypothesis in this paper is that the analytically oriented epistemology that has worked on the phenomenon of the Internet has done so taking as its model the testimony as the centre of gravity of social epistemology. This model is not adequate for the

https://doi.org/10.36576/2660-955X.52.43

new devices. The argument is that the epistemic authority (or lack of authority) of these devices is not based on some explicit or implicit contract of mutual recognition in exchanges of information or knowledge, but on the extensive socio-technical processes of design, formatting and data feeding, directed training and feedback with massive social use. This leads us to the need for an epistemology of massive hybrid agents.

Keywords: Generative artificial intelligence; Epistemology of artificial intelligence; Extended mind; Epistemology of the Internet.

1. NEOPIRRONISMOS DIGITALES Y EPISTEMOLOGÍA DE INTERNET

Hay una cierta analogía entre el escepticismo en la era digital y el originario de la modernidad. El escepticismo con el que se inaugura la modernidad nace en experiencias históricas que transformaron el entorno epistémico de las sociedades occidentales: la difusión del libro a través de la imprenta, las discusiones teológicas sobre la autoridad de la fe, entre la Iglesia o la interioridad y la gracia, el desarrollo de la experimentación cuidadosa y la consiguiente distinción entre cualidades primarias y secundarias, produjeron grandes controversias sobre la robustez de las creencias que se resumen en la metáfora de quien explora una cueva oscura llena de objetos algunos de oro y otros de latón y debe decidir qué llevarse con el objeto de hacerse rico. La mayoría de la creciente literatura sobre internet tiene una coloración escéptica al resaltar los muchos riesgos que aceptan quienes basan sus creencias en informaciones obtenidas de la red, muchas de las cuales son producto de *fake news* o de puras opiniones sin fundamento. Karen Frost-Arnold (Frost-Arnold, 2023) subraya la amenaza a la epistemología social que suponen todos estos nuevos fenómenos de usos malévolos o no fundamentados de la Web 2.0. Esta web, que sustituye a la vieja internet orientada únicamente a la comunicación profesional, se caracteriza por la ilimitada producción de páginas elaboradas por usuarios de toda índole, por la presencia de medios (redes) sociales en los que se discuten todo tipo de asuntos públicos y privados y por el inmenso poder de las plataformas digitales que dirigen estas redes, que proporcionan sistemas de "navegación" por las páginas, o que ofrecen todo tipo de servicios (Apps) comerciales o instrumentales. La Web 2.0 ha devenido en un auténtico entorno epistémico que compite, tal vez con un poder similar, si no mayor, con el entorno de la modernidad basado en la difusión de libros y artículos, los sistemas educativos y los actos comunicacionales de las instituciones epistémicas (congresos, seminarios, etc.).

La epistemología analítica contemporánea centró su trayectoria principal en la respuesta a la amenaza del escepticismo pirrónico moderno. A las escuelas fundacionalista, coherentista o pragmatista les ha sucedido una serie de variantes del fiabilismo entre las que destaca sin duda la epistemología de virtudes,

que ha derivado en formas de epistemología anti-riesgo epistémico, centrada en las capacidades de control epistémico por parte del agente tanto de sus facultades como de la riqueza o penuria epistémica de la situación cognitiva. Un segundo estadio en la pequeña historia de la epistemología analítica ha sido la extensión de estas estrategias antiescépticas a la epistemología social. Aunque Thomas Reid ya había observado que la inmensa mayoría de nuestras creencias provienen de la palabra de los otros, el individualismo metodológico había persistido en la tradición analítica hasta que diversas presiones desde la epistemología feminista y las aproximaciones críticas de los estudios de ciencia, tecnología y sociedad abrieron el campo de lo colectivo al análisis epistemológico. Fruto de este cambio fue traer a primera línea de la escena la epistemología del testimonio, en tanto que fuente básica y sustrato de la epistemología colectiva. La epistemología del testimonio se distingue de la epistemología de grupos en que se enfoca en las interacciones dinámicas interpersonales de compartir y aceptar información y, con ello aceptar las mutuas autoridades epistémicas de hablantes y oyentes (o escritores y lectores) bajo la condición normativa de testimonio.

En un tercer momento, la epistemología analítica ha girado hacia el fenómeno del ciberespacio e internet, en la versión Web 2.0 –o internet participativa– en la que hemos vivido en las dos últimas décadas de una forma masiva. La filosofía continental había producido una apreciable cantidad de literatura en el último tercio del siglo pasado sobre lo virtual, llegando incluso a la cultura popular en la forma del mito cartesiano del demonio maligno en la aclamada película *The Matrix* (Lilly y Lana Wachowski, 1999). Esta vuelta del pirronismo y la confusión entre lo virtual y lo real fue una de las marcas características del posmodernismo filosófico y su desprecio por la epistemología (Richard Rorty, Michael Williams). El abordaje analítico de la epistemología de internet, afortunadamente, ha ido por otros derroteros más interesantes, centrados en el carácter social y sus riesgos de la Web 2.0 o web de las plataformas de compartición de textos e imágenes y los motores de búsqueda.

La literatura sobre epistemología de internet ha crecido notablemente en los últimos años y se ha centrado principalmente en dos temas: la agencia epistémica y el riesgo epistémico generado por la toma de información basada en internet. El modelo teórico que ampara el tratamiento de estos temas es básicamente la amplia literatura sobre testimonio y, en general, sobre epistemología social. En esta sección resumiré brevemente los problemas que se han tratado en la epistemología de internet y cómo el modelo de testimonio se ha convertido en hegemónico en lo que respecta al problema del escepticismo. Algo distinta es la línea que han mantenido los análisis relacionados con la agencia desde la perspectiva de la mente extendida, que, en mi opinión, abren nuevas líneas para el tratamiento de la epistemología de los grandes modelos lingüísticos de inteligencia artificial generativa (en adelante, LLMs por sus siglas en inglés)

El foco de la epistemología de internet ha sido la confiabilidad de las informaciones obtenidas en las búsquedas en la gran mayoría de la literatura. El modelo teórico aplicado para este examen ha sido el del testimonio. Se basa en la idea de que un acto testimonial correcto desde la perspectiva epistémica es una fuente de conocimiento si se cumplen ciertas condiciones normativas establecidas según dos concepciones. En la primera, la que podemos denominar "individualista" o, según la literatura, "evidencialista", de origen humeano (Lackey, 2008), estipula que el testimonio es correcto si el hablante realiza una aserción que sabe verdadera y la realiza sin intención de manipular malévolamente las creencias del oyente o lector (se puede engañar diciendo la verdad). Si este, por su parte, ya tiene razones evidenciales para confiar en la palabra del hablante, aceptará la aserción y actualizará con ella sus creencias. En la segunda, denominada "interpersonal", se considera que el testimonio es una acto con un contenido normativo tal como lo son los contratos y otros actos similares. En este caso, es lo que se ha llamado una "acción conjunta" (Gilbert, 2013), que crea derechos y deberes en ambas partes. Si el oyente solicita una información *bona fide*¸ y el hablante se considera posesor del conocimiento demandado, ofrecerá esta información respaldando su palabra con su compromiso de compartir conocimiento y con su experticia para hacerlo. El oyente, por su parte, que ha concedido autoridad epistémica al hablante, está obligado a aceptar su palabra sin despreciar ni devaluar dicha autoridad.

No es este el lugar para discutir las respectivas bondades de ambas concepciones, pero sí cabe afirmar que en ellas cumple una función primordial el lazo de la confianza, una relación que es a la vez epistémica y emocional que liga a las dos partes del testimonio. La confianza, a diferencia de la fiabilidad que puede ofrecer un dispositivo mecánico como un automóvil o un ordenador, implica la voluntad y las capacidades cognitivas de las partes, especialmente la del prestador del testimonio en este caso. ¿Cumple el uso epistémico de internet las condiciones suficientes para que la consideremos una ampliación de las capacidades epistémicas o de la agencia epistémica (Gunn, Lynch, 2021)? Muy pronto se tomó como ejemplo de uso testimonial de internet el caso de Wikipedia (Tollefsen, 2009). Algunos autores adoptan una perspectiva optimista acerca de este ejemplo y otros consideran que el uso de internet puede ampliar las virtudes intelectuales de los agentes (Smart, Clowes, Heersmink, 2017; Heersmink, 2018, Schwengarer, 2021a y 2021 b; Smart y Clowes, Robert, 2021). Sus argumentos, independientemente de lo convincentes que sean, coinciden bastante con nuestras prácticas diarias, en las que confiamos a internet nuestras búsquedas de información con objetivos teóricos o prácticos, y seguimos confiadamente sus instrucciones, como ocurre, por ejemplo, en la orientación a través de Maps de Google.

No han faltado voces, sin embargo, que han criticado de forma generalizada los riesgos epistémicos de internet: Frost-Arnold 2014, 2021 y 2023 es una

conocida militante de los fallos de confiabilidad y el grave riesgo epistémico que generan las extendidas malas prácticas de internet: la "infoxicación" producida por las *fake news*, las intencionalmente falsas informaciones, las teorías de la conspiración o los filtros burbuja y cámaras de eco que han proliferado por un uso estratégico de internet para fines de propaganda (McKinnon, 2018). Se ha culpado a la anonimia que caracteriza muchas intervenciones en los medios (redes) sociales a esta pérdida de confianza (Ivy, 2021), lo que parece poner en peligro el uso del modelo de testimonio, que exige la confianza en la fuente, sea por la trayectoria pasada, sea por el compromiso explícito de dar la palabra. Lo relevante de esta ya larga discusión es que hay una deriva desde el análisis puramente epistemológico hacia la ética o política epistemológica, bien recordando las condiciones normativas de la práctica de la aserción (Goldberg, 2015), bien planteando la cuestión de quiénes deberían intervenir en internet, limitando la anonimia o las prácticas pasivas (*lurking*), tal como demanda Karen Frost-Arnold.

La discusión se alargaría mucho, pero hay razones para creer que no es este modo de mirar la interacción con las tecnologías digitales el mejor modo de tratar la epistemología de los LLMs, aunque los pocos trabajos que se encuentran en la literatura académica bajo el rubro de "epistemología de los LLMs o, más general, de la inteligencia artificial" (Coeckelbergh, 2020, Stevens, 2024, Heersmink, et al., 2024) parecen seguir centrados en el problema de la responsabilidad y la confianza en las producciones de la inteligencia artificial. No es extraño que el modelo de testimonio sea la manera popular y preteórica de enfrentarse a la inteligencia artificial. En la explosión informativa de los dos últimos años, que ha causado una inusitada ansiedad en todos los ámbitos sociales, especialmente en los de la educación, dejando al lado los múltiples temores al abandono del trabajo de aprendizaje por los alumnos, se ha generado una cierta ansiedad epistémica por la tasa de fallos, de "alucinaciones" y sesgos generados por las inteligencias artificiales. El modelo de testimonio se ofrece, por el contrario, como un marco aceptable para el análisis epistemológico de estas producciones generadas automáticamente: permite una discusión sobre los méritos distribuidos entre agentes respecto a la consecución de creencias verdaderas, sopesa las virtudes y vicios intelectuales y trata de la autoridad epistémica respectiva que tienen las partes en el contrato testimonial. Con estos instrumentos, la literatura sobre internet y la epistemología se centró en los estereotipos de la Web 2.0 como son los trabajos colaborativos que sostienen Wikipedia o la generosidad epistémica de las publicaciones online en blogs, la ciencia abierta; en el lado de los vicios, en lo malévolo de las intervenciones anónimas en redes, los *bots* que producen *fake news* o la proliferación de conspiracionismos y negacionismos. ¿Funciona con igual rendimiento este modelo en los ingenios lingüísticos o multimodales de la inteligencia artificial?

2. ACERCAMIENTOS EPISTÉMICOS A LOS LLMS

Aunque los más populares LLMs usan como fuente de datos lo que se puede encontrar en la Web 2.0, y por ello cabe pensar en que se pueden trasladar las consideraciones epistemológicas de uno a otro dominio, lo cierto es que hay diferencias notables que llevan a pensar en que debemos pensar en otros modos de análisis epistemológicos. La inteligencia artificial se ha dedicado históricamente al aprendizaje automático (*machine learning*), a la gestión, clasificación y uso de ingentes cantidades de datos (*data mining*) y, en sus versiones más especulativas, a la búsqueda de una inteligencia general autónoma. De hecho sin inteligencia artificial y aprendizaje automático no habría sido posible la Web 2.0, que necesitaba el recurso a los algoritmos clasificatorios para los motores de búsqueda o para generar el descomunal número de aplicaciones con funciones concretas, las populares *apps*. Aunque la influencia de la inteligencia artificial en internet ha tenido poco eco en los análisis epistemológicos, más centrada en la cuestión de la fiabilidad humana, en el caso de los LLMs, la automatización y la integración humanos-máquinas cambia de forma notable el marco de análisis epistemológico.

Comencemos por notar algunas diferencias entre la inteligencia artificial clásica simbólica (GOFAI, en la jerga introducida por John Haugeland en 1985) y los actuales modelos de aprendizaje profundo. La importancia de este cambio es relevante para el análisis epistemológicos de los nuevos dispositivos. La inteligencia artificial simbólica, y sus máquinas de aprendizaje automático estaban basadas en la mimetización del conocimiento experto humano en estructuras simbólicas, de las que se ocupaba (y ocupa) la ingeniería del conocimiento. Sus resultados a lo largo del último tercio del siglo pasado han tenido una repercusión extraordinaria en la filosofía de la mente y en la epistemología naturalizada como la representada por los primeros trabajos de Alvin Goldman. El panorama, sin embargo, cambió con la introducción de las redes neuronales a partir de los años ochenta y los desarrollos exponenciales del presente siglo. A diferencia de la IA clásica, la IA generativa no pretende tanto imitar patrones de acción ya conocidos como generar otros nuevos, en forma de textos, códigos o imágenes y sonidos en los sistemas multimodales. La capacidad generativa no nace de variaciones sobre esquemas o guiones de acción estructurados que le hayan sido suministrados a la máquina provenientes de la experticia humana, sino que son producciones basadas en patrones probabilísticos que el sistema encuentra en ingentes cantidades de datos que le son suministrados. Estos productos lo son a petición de humanos en la forma de preguntas o problemas (*prompts*) que, a su vez, son datos estructurados que constriñen las posibles respuestas. La reacción humana a estas producciones les sirve como entrenamiento para nuevas producciones, por lo que en realidad hay que entender que los modelos deben ser evaluados no en actos singulares cuanto en dinámicas largas de aprendizaje.

Lo central de los LLMs es la extraña composición de agencia humana y mecanismos automatizados, en la que se produce una mezcla de opacidad y transparencia. Andrada et alt. 2023 han analizado las diversas formas de opacidad que generan los nuevos modelos de inteligencia artificial y cuáles pueden ser sus consecuencias, principalmente éticas. El problema epistemológico, sin embargo, tiene que ver con lo que podría llamarse la "opacidad algorítmica" parcial que introduce la arquitectura y funcionamiento de estos modelos (véase Fig. 1). Los modelos más conocidos como GPT 4 pertenecen a una variedad que usa "*transformers*", que son dispositivos automáticos para el autocontrol y autoaprendizaje basado en la interacción con entrenadores y usuarios. Merece la pena observar cómo la mediación técnica no es inocua para las propiedades epistemológicas del compuesto humanos-máquina que son estos ingenios.

El funcionamiento más o menos eficiente de estos modelos depende de numerosos factores entre los que cabe señalar:

1.- La *disponibilidad* de datos. Este es un factor esencial. Los datos se han convertido en la fuente más importante de diferencias tecnológicas y epistémicas en el siglo XXI. Todo puede ser un dato, en tanto en cuanto pueda ser representado digitalmente, almacenado y pre-tratado para que pueda operar en las entrañas de un dispositivo de inteligencia artificial. No es fácil saber cuántos y cuál es la procedencia exacta de los datos que alimentan a los LLMs más populares como GPT 4, Gemini, Llama y otros, pero cabe pensar razonablemente que han sido alimentados con todo lo disponible en la Web, fundamentalmente textos, en lo que se refiere a los modelos puramente lingüísticos y con imágenes y otras modalidades a los ampliados a la multimodalidad. El complejo de datos ya no es una representación fiel de la realidad, sino un conjunto dependiente de numerosas fuentes fiables o no fiables que alimentan los algoritmos de los modelos.

2. El proceso de *representación*: los datos, ya digitalizados, son convertidos a través de un tratamiento informático en *tokens* o unidades que puedan ser situadas en los nodos de las redes neuronales. Estos *tokens*, a su vez, son convertidos en *vectores* para que puedan ser operados por los dispositivos de tratamiento (*transformers*) cuya función es crear contextos de probabilidades de relación asociadas a un *token* de modo que se acorte el tratamiento masivo en las redes. Aquí ya se produce un segundo alejamiento entre la fuente y la unidad de tratamiento, que depende de cuál es la arquitectura informática del modelo.

3. *Entrenamiento*: el buen rendimiento funcional, cognitivo, de los modelos lingüísticos o multimodales depende de su capacidad de aprendizaje que, a su vez, depende del entrenamiento. Este puede ser supervisado por los ingenieros del sistema o por personal contratado específicamente para ello o no supervisado y dependiente de la interacción continua con usuarios y a través de sus dispositivos internos de auto-aprendizaje o auto-control. Aunque en todas las

fases hay intervención humana, es en el entrenamiento en donde aparece claramente el carácter mixto, híbrido, de estos sistemas, que no funcionarían sin la adecuada interacción humanos-máquinas en las diversas formas de entrenamiento y corrección de errores.

4. *Ajuste* (*fine-tunning*): Aunque buena parte del éxito mediático de estos dispositivos se debe a las expectativas abultadas artificiosamente sobre el carácter "general" de la inteligencia de los modelos, lo cierto es que su valor pragmático y comercial depende de formas de ajuste fino de los sistemas para objetivos prácticos muy específicos de orden empresarial, militar, científico, educativo, etc., donde los datos son seleccionados y sobre todo se generan formas de aprendizaje por refuerzo (de nuevo la importancia del entrenamiento) orientadas a elegir las venas respuestas y a sustentar la fiabilidad del sistema. Así, por ejemplo, los usos de estos modelos en traducción y su creciente fiabilidad depende mucho de estos ajustes.

Estos cuatro puntos y quizás otros que podrían tenerse en cuenta nos llevan al convencimiento de que la inteligencia generativa es un producto tanto del artefacto como, para decirlo en términos vigotskyanos, de la *zona de desarrollo próxima*, es decir, de la interacción con un entorno inteligente como es el de los ingenieros, los entrenadores masivos y los mucho más masivos usuarios. La fiabilidad de estos modelos varía y cambia con los progresivos entrenamientos. Las primeras impresiones de los usuarios novatos, como el que escribe, son a veces de fascinación por los resultados, pero hay que esperar a las evaluaciones de los expertos en las distintas áreas y aplicaciones. En todo caso, sus producciones parecen ser bastante acertadas en tareas como la traducción, en preguntas no demasiado complicadas, cuyas respuestas se encuentren ya representadas en la Web y en algunas otras tareas de experticia no abierta debido a los contratos confidenciales de uso.

Las críticas más usuales respecto a la fiabilidad de estos sistemas son la alta tasa de "alucinaciones", un término que se ha generalizado para indicar las producciones falsas o incorrectas del sistema. Por ejemplo, la invención de referencias bibliográficas inexistentes suele ser una queja usual en los contextos académicos, tanto de estudiantes como de investigadores, que emplean estos modelos para información inicial en sus trabajos. Junto a la tasa de alucinaciones, que alcanza tantos por ciento notables, al decir de algunos expertos respecto a modelos populares como GPT4, pero que varía por temas a lo largo del tiempo, son también habituales las quejas por los sesgos identitarios que a veces se observan en las consultas, especialmente en las preguntas sobre interpretación de imágenes (uno de los campos más atrasados en los modelos generativos). Estas quejas son las que, unidas al trasfondo del modelo de testimonio que está en el trasfondo de las evaluaciones sustenta un generalizado escepticismo filosófico que contrasta con el exagerado y propagandístico entusiasmo de una gran parte de la comunicación científico-técnica.

(Prem, 2023) acierta al observar que mucha de esta crítica nacida del marco testimonial se sustenta sobre un supuesto erróneo: el de que una inteligencia general generativa es algo así como un modelo de la realidad y que, por ello, sus fallos predictivos son fallos directamente epistémicos. El autor sostiene, por el contrario, que son dispositivos que no crean mapas del mundo sino mapas de los enormes almacenes de textos e imágenes que los alimentan. Su tesis es que un LLM podría parecerse, más que a un sistema de testimonio, a una ficción literaria que no refiere directamente al mundo sino al complejo de experiencias, memorias y textos del que la escritora extrae un relato. Se crea así un modelo oblicuo y ficcional de un universo que no existiría sin la realidad, cierto, pero que no es un mapa de lo real como puede serlo una teoría o modelo formulados con una intención referencial y veritística. Las alucinaciones y sesgos no son algo extraño como no lo son en la ficción: son parte de la construcción del sistema.

Mirados desde esta perspectiva, los LLM parecen mucho más humanos de lo que son. Pues, al igual que un cerebro creando textos o imágenes, hay un grado de opacidad y falta de autoconocimiento notable en su proceso de producción; hay también un elemento combinatorio en el que las afinidades de textos no se crean mediante cercanías por taxonomías conceptuales o lógicas, como podrían ser los que estructuraban la arquitectónica de la inteligencia artificial simbólica, sino que son producto de perfiles y distancias terminológicas que son generados combinatoriamente por los pesos computacionales de las redes neuronales, a veces más cercanas a las dinámicas de los sueños freudianos que a los modelos matemáticos de los sistemas físicos que relacionan variables que representan propiedades y magnitudes reales.

¿No nos hemos al otro extremo desde la epistemología a la doxología al considerar los LLMs como producciones ficcionales? Sí, en cierta manera. Pero, al fin y al cabo, este modo de considerar los dispositivos de modo general, sin cualificar sus tareas prácticas concretas, es mucho más cercano a su naturaleza que los estereotipos que producen toda la ansiedad epistémica que recorre ahora el mundo de la educación o el de una parte de las empresas. Recordemos que los LLMs no pueden separarse de las bases de datos que los alimentan, de forma que son estas y no la realidad las que utilizan para crear sus perfiles y patrones. De hecho, nuestra experiencia cotidiana con internet, ya movida por estos ingenios, corrobora muy bien este modelo de ficción: si durante unas horas tenemos el capricho de buscar si hay abrigos de lana para perros chiguagua, sin que tengamos ningún interés práctico en ello, nuestras nuevas exploraciones se llenarán de anuncios de perros, de ropa y alimentos caninos, de videos de perritos adoptados, de anuncios de aplicaciones para adoptar perros abandonados, …, y estaremos sumergidos unos días en un universo canino que puede que no nos importase lo más mínimo al comienzo.

Desafortunadamente, nuestro realismo sobre el funcionamiento de la inteligencia artificial generativa nos devuelve a un horizonte pirrónico como el del mundo Matrix del posmodernismo del siglo pasado. Si queremos persistir en nuestra resistencia al escepticismo tal vez debamos cavar algo más en este jardín y tener en cuenta otros aspectos que hemos dejado de lado.

3. LOS LLMS COMO INGENIOS EPISTÉMICOS HÍBRIDOS

La ansiedad epistémica que suscitan los LLMs, como otros casos de ansiedad –ya hemos aludido a la que recorre las evaluaciones de internet, a la que podríamos añadir la que suscitan en general los medios de comunicación–, tiene una base correcta, en tanto que detecta una alta tasa de falibilidad, y una equivocada expectativa sobre lo que estos modelos tendrían que hacer. Lamentablemente estas expectativas han sido creadas por la popularidad que han adquirido, muchas veces fomentada por el propio diseño de los LLMs cuando funcionan en abierto y de modo no propietario, induciendo la idea falsa de que son un paso adelante en la consecución de una inteligencia artificial general que imitará o sustituirá a la humana. Pocas controversias sociales son más lamentables que esta.

Los LLMs son buenos en lo que son, y en ello residen sus posibles virtudes epistémicas. (Alvarado, 2023) ha sugerido que debemos considerar las inteligencias artificiales como tecnologías epistémicas, y en particular como "mejoradores epistémicos" (*epistemic enhancers*). La idea es muy general, tal como se expone en el artículo, pero creo que va en la dirección correcta: la manera adecuada de examinar el valor epistemológico de estos dispositivos debería ser mucho más pragmática, al modo como enfocamos la tecnología que nos rodea sin los agobios heideggerianos sobre si es una degeneración del *dasein* en su olvido del ser y una caída en el pensamiento instrumental. Sí, quizás, cierto funcionalismo, al menos como el que mantiene la teoría dual de los artefactos (Kroes, Meijers, 2006), podría ayudarnos a resolver las cuestiones escépticas sobre las inteligencias generativas.

La teoría dual de los artefactos sostiene que debemos considerar un dispositivo técnico como un mecanismo que cumple una función en tanto que ha sido diseñado para que los usuarios entiendan y usen el artefacto para esa función. Hay aquí, pues, la doble perspectiva de lo mecánico y lo intencional, que da cuenta de muchos de los debates que se han producido en la filosofía de la tecnología contemporánea. Mi propuesta es continuar en esa línea aplicada ahora a intenciones y funciones epistemológicas.

El primer punto de esta hipótesis es que no es correcto examinar las propiedades epistemológicas de las inteligencias artificiales tomando como unidades de evaluación el programa y el acto concreto de un particular que aplica un

promt para que la máquina le dé una respuesta epistemológicamente correcta que él tome como base para formar una creencia justificada. Este modo de evaluar los LLMs no atiende a su funcionamiento real, sino a un estereotipo creado artificialmente para propósitos comerciales que no importan en este trabajo. Al contrario, hay que tener en cuenta que un modelo lingüístico o multimodal debe ser valorado epistémicamente teniendo en cuenta los cuatro momentos que he señalado anteriormente: la base de datos, el proceso de representación, las fases de entrenamiento supervisado, no supervisado y el producido por el uso y, por último, y muy relevante, el ajuste a tareas específicas para las que se demanda socialmente uno de estos modelos.

Entendidos de este modo, estos dispositivos tienen una complejidad sociotécnica y una dinámica de funcionamiento, uso y corrección que son centrales en la evaluación epistemológica. Podemos considerarlos sistemas epistémicamente híbridos y masivos, compuestos por numerosos agentes humanos con una gran división social del trabajo cognitivo y técnico que suministran datos, que los preparan para su introducción en las redes neuronales y transformadores, por el propio sistema algorítmico, la base material y energética que sostiene la adquisición, almacenamiento y tratamiento de datos, el campo de intereses y dominio social para el que se ha diseñado el ajuste fino, y el progresivo entrenamiento generado por las derivas del uso.

Nos encontramos, pues, ante un complejo sociotécnico híbrido sobre el que ahora podemos preguntarnos por la agencia epistémica, por la posible autoridad y, consecuentemente, por las virtudes y vicios y el riesgo epistémico que asumimos al incorporarlo a nuestras vidas. ¿Son artefactos epistémicos los LLMs? La respuesta a esta pregunta no puede ser independiente de qué unidad e intervalo temporal consideremos como base para la evaluación y, en segundo lugar, cuándo y cómo se introduce la demanda de conocimiento en el sistema. La misma doble consideración es la que debe guiarnos para responder a la pregunta de la existencia y naturaleza de la autoridad epistémica de estos ingenios y de sus virtudes y vicios.

La respuesta a estas preguntas no puede ser incondicionada, sino dependiente de diseños y usos y modelos concretos. Se trata de complejos sociotécnicos en donde el conocimiento se encuentra distribuido (Hutchins, 1995) y en los que, por ello, es en el conjunto de la circulación de datos, procesamiento y correcciones en el que tenemos que plantearnos las preguntas epistemológicas. Muchos de los usos y diseños no tendrán ninguna función epistémica o no será central, por ejemplo en los usos lúdicos o estrictamente empresariales para manipular hojas de cálculo. En otros, como por ejemplo en lo que parece que son bastante buenos, en la traducción, la consideración epistémica es más bien de contenido que de referencia: importa el grado en que ha sido preservado el sentido y el significado del texto. En otros, sin embargo, en los usos científicos, educativos, militares o gubernamentales, la función epistémica es central y hay

que considerar no tanto los posibles fallos sino la dinámica de aprendizaje del sistema.

Tenemos, por tanto, una agencia epistémica distribuida, mucho más que en el caso del testimonio, mucho más masiva en todos los órdenes de cantidad de humanos y de recursos materiales y, por ello, con una autoridad epistémica también distribuida que debe situarse entre el dogmatismo y el escepticismo como una autoridad falible y en continua corrección. El carácter distribuido aquí es central. No se trata simplemente de una distribución entre cerebros, al modo en que se considera esta cuestión en epistemología social, sino de una distribución más cercana a lo que se considera en las reflexiones sobre la mente extendida: las máquinas realmente hacen inferencias, general probabilidades y con ellas hacen predicciones cuya relevancia epistémica es examinada y corregida en los procesos tanto de entrenamiento como de uso.

Entre el entusiasmo comercial y el pirronismo digital, ambos fenómenos comunicacionales del momento presente mi propuesta es considerar los LLMs (o los MLLMs multimodales) como dispositivos sociotécnicos falibles para los que necesitamos una epistemología aplicada que no es la epistemología social, tal como está siendo desarrollada en el contexto académico analítico, ni tampoco las derivas del tratamiento de la inteligencia artificial hacia la ética de la tecnología, sino que debe abrirse un campo específico que, usando el término luminoso de Alvarado,2023 podría ser el de la epistemología tecnológica o tal vez la tecnología epistémica.

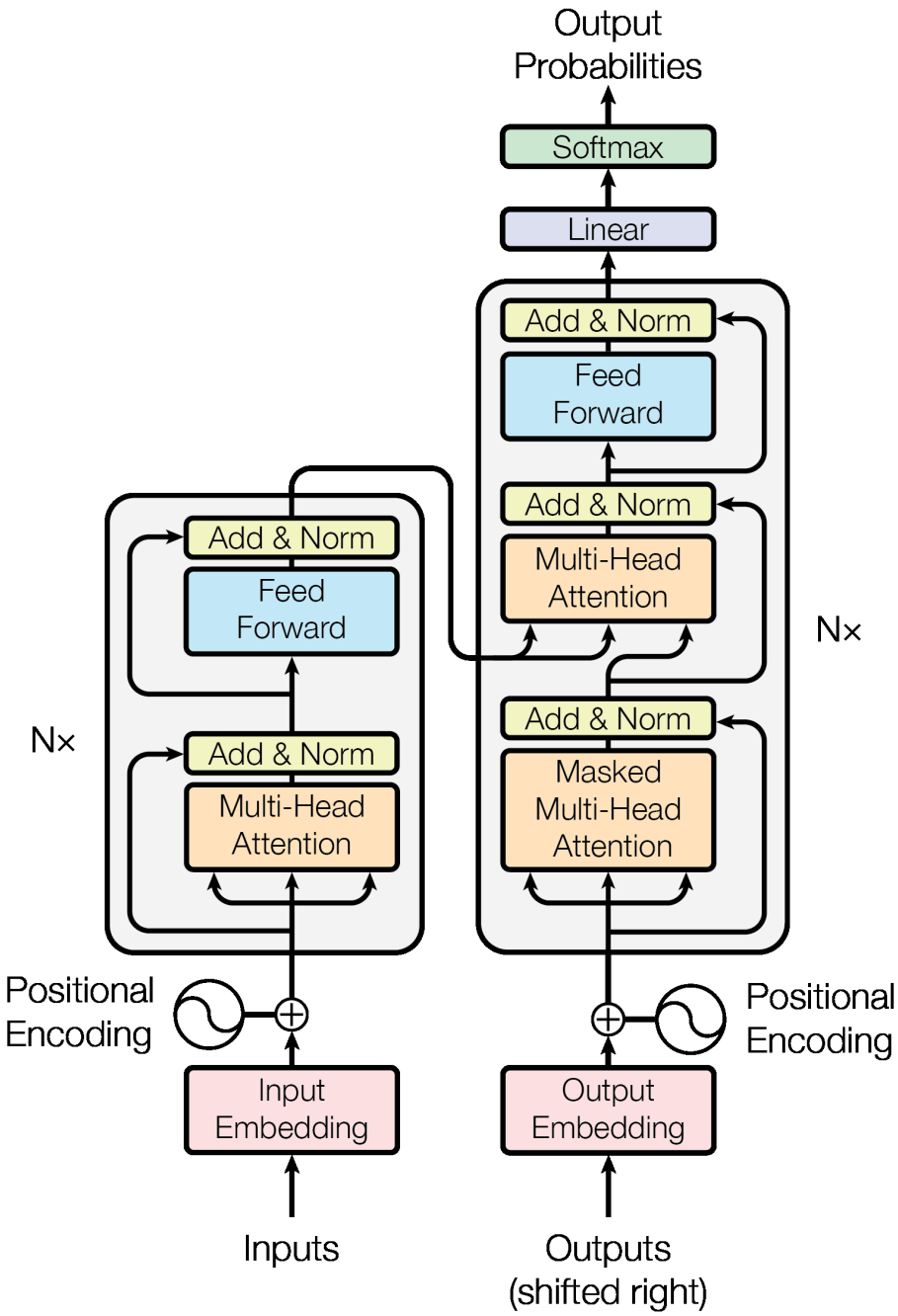

Fig 1[1]

1 Fuente: https://www.run.ai/guides/generative-ai/transformer-model (recuperado el 31/08/2024) y Yuening Jia - DOI:10.1088/1742-6596/1314/1/012186. Licencia CC BY-SA 3.0

REFERENCIAS BIBLIOGRÁFICAS

ALVARADO, RAMÓN (2023) "AI as an Epistemic Technology. Science and Engineering Ethics 29 (5):1-30.

ANDRADA, GLORIA; CLOWES, ROBERT WILLIAM & SMART, PAUL (2023). Varieties of transparency: exploring agency within AI systems. AI and Society 38 (4):1321-1331.

BAI, H. (2022) "The epistemology of machine learning", *Filosofija,Sociologija*. 33/ 1, p. 40–48, en https://www.lmaleidykla.lt/ojs/index.php/filosofija-sociologija/article/view/4668 (recuperado el 31/08/2024

BLUMER, A. (2024) "LLMs have revolutionized AI. Do we still need knowledge models and taxonomies, and why?" https://www.linkedin.com/pulse/llms-have-revolutionized-ai-do-we-still-need-models-why-blumauer-rfkkf/ (recuperado el 31/08/2024)

CARTER, J. ADAM & GORDON, EMMA C. (2020). "Is searching the internet making us intellectually arrogant?" en Alessandra Tanesini & Michael P. Lynch (eds.*), Polarisation, Arrogance, and Dogmatism: Philosophical Perspectives*. London, UK: Routledge.

CARTER, J. ADAM; AND, & SIMION, MONA (2020). "The Ethics and Epistemology of Trust". Internet Encyclopedia of Philosophy.

COECKELBERGH, M. (2020) "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability". *Science and Engineering Ethics 26* (4):2051-2068.

COECKELBERGH, M. (2021) *Ética de la inteligencia artificial*, trad. Lucas Álvarez Canga, Madrid: Cátedra

COECKELBERGH, M. (2023) *La filosofía política de la inteligencia artificial: Una introducción,* trad. Lucas Álvarez Canga, Madrid: Cátedra.

FROST-ARNOLD, K. (2021) "The Epistemic Dangers of Context. Collapse online", en Lackey J. (ed.) (2021), *Applied Epistemology*. New York, NY: Oxford University Press.

FROST-ARNOLD, KAREN (2014). "Trustworthiness and truth: The epistemic pitfalls of internet accountability". *Episteme 11* (1):63-81.

FROST-ARNOLD, KAREN (2023). "Who Should We Be Online? A Social Epistemology for the Internet". New York: Oxford University Press.

GILBERT, MARGARET (2013). *Joint Commitment: How We Make the Social World*. New York, NY: Oup Usa.

GOLDBERG, SANFORD (2015). *Assertion: On the Philosophical Significance of Assertoric Speech*. New York, NY: Oxford University Pres

GUEGUEN, GAEL & YAMI, SAID (2010). "Internet in the process of data collection and dissemination". en Bernard Reber & Claire Brossaud (eds.), *Digital cognitive technologies: epistemology and the knowledge economy*. Hoboken, NJ: Wiley.

GUNN, HANNA & LYNCH, MICHAEL P. (2021). "The Internet and Epistemic Agency." en Jennifer Lackey (ed.), Applied Epistemology. New York, NY: Oxford University Press. pp. 389-409.

HEERSMINK, RICHARD (2018). "A virtue epistemology of the Internet: Search engines, intellectual virtues and education". *Social Epistemology 32* (1):1-12.

HEERSMINK, RICHARD; DE ROOIJ, BAREND; CLAVEL VÁZQUEZ, MARÍA JIMENA & COLOMBO, MATTEO (2024). "A phenomenology and epistemology of large language

models: transparency, trust, and trustworthiness". *Ethics and Information Technology 26* (3):1-15.

HUTCHINS, EDWIN (1995). *Cognition in the Wild*. Cambridge MA: MIT Press.

KROES, PETER & MEIJERS, ANTHONIE (2006). The dual nature of technical artefacts. Studies in History and Philosophy of Science Part A 37 (1):1-4.

IVY, V. (2021) Yikkity Yak, "Who Said That?' The Epistemology of Anonymous Assertions, en Lackey J. (ed.) (2021), *Applied Epistemology*. New York, NY: Oxford University Press.

IVY, VERONICA (2021) Part Eight: Epistemology and the Internet. The Internet and Epistemic Agency / en Lackey, J. (2008) *Learning form Words: Testimony as a Source of Knowledge*, Oxford: Oxford University

MCKINNON, R. (2018). The Epistemology of Propaganda. Philosophy and Phenomenological Research 96 (2):483-489.

MÖßNER, NICOLA & KITCHER, PHILIP (2017). "Knowledge, Democracy, and the Internet". *Minerva 55* (1):1-24.

PREM E. (2023) "The Semantics, Ethics, and Epistemology of Large Language Models", en https://caiml.org/dighum/summerschool2023/program/llm-semantics-ethics-epistemology.pdf (recuperado el 31/08/2024)

SCHMIDT, C. T. A. (2007). "Artificial Intelligence and learning, epistemological perspectives". *AI and Society 21* (4):537-547.

SCHWENGERER, LUKAS (2021) a. "Online Intellectual Virtues and the Extended Mind". *Social Epistemology 35* (3):312-322.

SCHWENGERER, LUKAS. (2021)b. "Revisiting Online Intellectual Virtues." *Social Epistemology Review and Reply Collective* 10 (3): 38-45. https://wp.me/p1Bfg0-5JX

SMART, P. R., CLOWES, R. W., & HEERSMINK, R. (2017). "Minds Online: The Interface between. Web Science, Cognitive Science and the Philosophy of Mind". *Foundations and Trends in Web Science*, *6*(1–2), 1–232. https://doi.org/10.1561/1800000026

SMART, PAUL & CLOWES, ROBERT (2021). "Intellectual Virtues and Internet-Extended Knowledge". *Social Epistemology Review and Reply Collective 10* (1):7-21.

SMART, PAUL (2017). "Extended Cognition and the Internet: A Review of Current Issues and Controversies". *Philosophy and Technology 30* (3):357-390.

SMART, PAUL; HEERSMINK, RICHARD & CLOWES, ROBERT (2017). "The Cognitive Ecology of the Internet". In Stephen Cowley & Frederic Vallée-Tourangeau (eds.), *Cognition Beyond the Brain: Computation, Interactivity and Human Artifice* (2nd ed.). Springer. pp. 251-282.

SMART, PAUL R. (2012). "The Web-Extended Mind." *Metaphilosophy* 43 (4):446-463

SMART, PAUL R. (2022). "Toward a Mechanistic Account of Extended Cognition". *Philosophical Psychology 35* (8):1107-1135

SPENCE, E.H. (2009). "A universal model for the normative evaluation of internet information." *Ethics and Information Technology 11* (4):243-253.

STEVENS, I (2024). "The Epistemological Consequences of Artificial Intelligence, Precision Medicine, and Implantable Brain-Computer Interfaces". *Voices in Bioethics 10*. DOI https://doi.org/10.52214/vib.v10i.12654 (https://doi.org/10.52214/vib.v10i.12654)

STEVENS, IAN (2024). "The Epistemological Consequences of Artificial Intelligence, Precision Medicine, and Implantable Brain-Computer Interfaces". *Voices in Bioethics 10*, en https://journals.library.columbia.edu/index.php/bioethics/article/view/12654 (recuperado el 31/08/2024)

TOLLEFSEN, D.P. (2009). "WIKIPEDIA and the Epistemology of Testimony". *Episteme 6* (1):8-24.