

EL MITO DE LA CAVERNA DIGITAL DE PLATÓN: EL ROL DE LA INTELIGENCIA ARTIFICIAL GENERATIVA EN EL MUNDO DE LAS IDEAS Y SU CONTENIDO COMO LA SOMBRA DEL ARTE

Plato's Myth of the Digital Cave: The Role of Generative Artificial Intelligence in the World of Ideas and its Content as the Shadow of Art

Eduardo C. Garrido-Merchán

Universidad Pontificia Comillas - Instituto de Investigación Tecnológica (IIT)
ecgarrido@icade.comillas.edu / <https://orcid.org/0000-0002-2695-5484>

Recibido: 13 de noviembre de 2024

Aceptado: 15 de diciembre de 2025

DOI: <https://doi.org/10.14422/ryf.vol289.i1467.y2025.001>

RESUMEN: Este trabajo plantea dos hipótesis para reflexionar sobre el papel de la inteligencia artificial (IA) desde una mirada platónica. La primera hipótesis propone que una red neuronal profunda ideal —entrenada con una cantidad infinita de datos que abarque todo el conocimiento objetivo del mundo— podría formar parte del Mundo de las Ideas de Platón. Estos sistemas ya están generando representaciones, como jugadas de ajedrez, que los seres humanos no logramos comprender. Por eso, se sugiere que los modelos de aprendizaje automático, gracias a su capacidad de crear representaciones del mundo sensible, pueden considerarse una sombra aproximada de las Ideas platónicas, ofreciéndonos una nueva manera de pensar la relación entre lo visible y lo inteligible. La segunda hipótesis, de índole más práctica y crítica, surge como consecuencia de la anterior. En particular, advierte sobre los riesgos del contenido generado por estas inteligencias artificiales, especialmente cuando se crean imágenes y textos desde una intencionalidad negativa. Si el

arte, para Platón, ya era una copia de lo real —una sombra en la Caverna—, entonces los contenidos generados por IA serían una “sombra de esa sombra”: un Segundo Nivel de la Caverna. En este escenario, podríamos estar cada vez más atrapados en una Caverna Digital, viendo representaciones cada vez más alejadas de la verdad, sin darnos cuenta de que hemos dejado de mirar la realidad.

PALABRAS CLAVE: Inteligencia artificial, Platón, aprendizaje profundo, mito de la caverna.

ABSTRACT: This work explores two hypotheses regarding the role of artificial intelligence from a Platonic perspective. The first proposes that a utopian deep neural network —trained on an infinite amount of data encompassing all objective knowledge of the world— could have its place within Plato's World of Ideas. These systems are already generating representations, such as chess strategies, that humans cannot comprehend. Therefore, it is suggested that machine learning models, thanks to their capacity to generate representations of the sensible world, can be considered an approximate shadow of Platonic Ideas, offering us a new way of thinking about the relationship between the visible and the intelligible. The second hypothesis, more practical and critical in nature, follows from the first. It warns of the risks posed by content generated by such artificial intelligences, especially when images and texts are created with negative intent. If art, for Plato, was already a copy of reality —a shadow in the Cave— then AI-generated content would be a “shadow of that shadow”: a Second Level of the Cave. In this scenario, we may find ourselves increasingly trapped in a Digital Cave, perceiving representations ever further from the truth, without realizing that we have stopped looking at reality itself.

KEYWORDS: Artificial Intelligence, Plato, Deep Learning, The Parable of the Cave.

1. INTRODUCCIÓN

El alcance epistémico de la tecnología y ciencia actuales es un subconjunto de la realidad ontológica trascendente a nuestro entendimiento antropológico (Copleston, 2003; Blanco, 2022). La pregunta por el Sentido, la subjetividad, la consciencia fenoménica de Ned Block, o el Ser en cuanto Ser aristotélico, son preguntas cuya verdad solo puede, si cabe, ser reducida estéticamente por la fenomenología relativa al individuo. En cualquier caso, las fronteras de la ciencia, debido a su actual metodología, excluyen para nuestro entendimiento racional, que no intuitivo, estas realidades, que se perciben en el plano fenomenológico. Es por ello verosímil, o al menos no refutable empíricamente, incluir en una ontología cosmológica la hipótesis de una realidad trascendente al mundo sensible. Por ejemplo, la teoría metafísica del Mundo de las Ideas de Platón (Annas, 1981; Fraile, 1997).

Por desgracia, ni el discípulo de Sócrates ni la escuela Pitagórica pudieron hacer hipótesis sobre la existencia o comprender la teoría probabilista (De-Groot y Schervish, 2012) sobre los modelos de aprendizaje automático (Mur-

phy, 2012). La pertinencia de contemplar estos objetos matemáticos según la ontología platónica, al estilo de la inclusión en la jerarquía platónica de la programación orientada a objetos (Tylman, 2018), es crítica, ya que estos modelos, al ser implementados en un ordenador, son capaces de generar nuevas teorías físicas no alcanzables por una mera epistemología humana (Karagiorgi *et al.*, 2022), validar teoremas matemáticos (Bridge, 2010) o generar funciones matemáticas (Rasmussen, 2003) mediante el estudio de los datos del mundo sensible.

En este artículo se estudian dos hipótesis. La primera consiste en que un modelo utópico de capacidad potencialmente infinita entrenado con la totalidad de la información objetiva del mundo sensible es una aproximación al concepto de Idea platónica y tiene, por tanto, cabida en la metafísica de Platón en un lugar inmediatamente inferior al de la Idea. Dado el potencial de la manifestación física de esta red neuronal en el mundo sensible y una mala intencionalidad, el artículo aborda como repercusión inmediata de la tesis presentada la hipótesis de que los contenidos generados por esta red pueden introducir a los individuos en una suerte de Caverna Digital de Platón, donde el contenido generado por la red sería una sombra de la representación digital del mundo sensible, los datos de entrenamiento. Dado que estos datos de entrenamiento serían a su vez una sombra de la realidad, los datos generados por la IA serían una sombra de las sombras, un segundo nivel de la Caverna Digital que haría que el individuo atrapado en la misma tuviera más difícil que nunca su ascensión al Bien, ya que estas sombras además están optimizadas para atraparle y dados los nuevos avances en realidad virtual y aumentada que podrían inhibir del resto de estímulos. Resumimos las implicaciones que se exponen en este artículo en la Tabla 1 para una mayor claridad de los argumentos expuestos y debatidos en este artículo.

TABLA 1

Implicación Lógica	Escuela de Pensamiento	Conocimiento	Notas
Una red neuronal puede aproximar cualquier función dada una arquitectura correcta y datos densos (Teorema del Aproximador Universal).	Realismo Matemático	Epistemológico	Fundamento para implicaciones subsecuentes.

Implicación Lógica	Escuela de Pensamiento	Conocimiento	Notas
Si los datos reflejan la distribución objetiva del universo, la red neuronal puede aproximar esta distribución con error mínimo.	Realismo Platónico	Ontológico	Extiende el teorema a una escala universal.
Las redes neuronales pueden aproximar cualquier subconjunto de la distribución universal, incluidas las Ideas platónicas de conceptos matemáticos y leyes.	Realismo Platónico	Ontológico	Las Ideas son aproximadas asintóticamente, no alcanzadas completamente.
Las redes neuronales ocupan un lugar inferior a las Ideas Platónicas al reducir conceptos universales a aproximaciones.	Idealismo Platónico	Ontológico	Resalta las limitaciones de las redes al modelar Ideas subjetivas o universales.
Las redes neuronales no pueden aproximar información subjetiva, reduciendo así su alcance.	Escepticismo Epistemológico	Epistemológico	Discute la incapacidad de capturar dimensiones subjetivas.
Las redes neuronales entrenadas con datos objetivos podrían ser usadas para crear una “Caverna Digital”.	Alegoría Platónica de la Caverna	Ontológico	Describe el peligro de manipular verdades objetivas para crear ilusiones.
La “Caverna Digital” es un segundo nivel de la Caverna, donde el primer nivel representa los datos objetivos usados para el entrenamiento.	Alegoría Platónica de la Caverna	Epistemológico	Distingue entre niveles de manipulación y abstracción.
Las redes neuronales avanzadas representan un riesgo mayor que artefactos previos al optimizar la adicción a contenidos, alejando aún más del “Bien”.	Realismo Ético	Epistemológico	Vincula el mal uso de la IA con riesgos éticos e impacto social.

2. ESTADO DEL ARTE SOBRE FILOSOFÍA ANTIGUA Y FILOSOFÍA DE LA IA

Dada la enorme repercusión que la IA puede tener en la sociedad y su rol en el universo que nos rodea esta ha sido estudiada en profundidad por la filosofía en libros que muestran una introducción global a la filosofía de la IA (Copeland, 1993; Carter, 2007) y también en artículos específicos de IA y filosofía en general (Mueller, 2012; Zhang, 2022). En esta sección vamos a limitar el alcance a la intersección entre filosofía antigua e ideas de filosofía medieval relacionada con la IA para una mayor coherencia con el tema del que se ocupa este artículo.

La alegoría de la caverna ya ha sido usada con anterioridad en lo respectivo a la IA (Karpouzis, 2024). Sin embargo, se ha usado para examinar cómo los sistemas de IA generativa, al ser entrenados con datos culturalmente sesgados, pueden perpetuar y amplificar estos sesgos. Se usa la alegoría de la Caverna de Platón como marco para discutir las limitaciones en la representación y percepción de la realidad en contextos digitales.

El clásico problema mente-cerebro y su conexión con la IA a través del problema filosófico de los universales ha sido explorado usando conceptos platónicos para aportar una visión sobre la inteligencia humana y la inteligencia exhibida por las respuestas producidas por los modelos de aprendizaje profundo, con sus similitudes y diferencias (Mary, Ross y Arun, 2024). Y una diferencia radica en que los sistemas clásicos de IA se limitaban a replicar una representación objetiva de nuestro entorno, lo cual en el caso de la IA clásica ya se asemejaba a la Caverna de Platón (Weinberger, 1988). La diferencia es que en este artículo se va a abordar no la IA clásica que usaba motores de reglas e inferencia lógica sino aprendizaje profundo con redes neuronales que guarda una mayor semejanza con la estadística computacional. Se verá sin embargo cómo en el caso del aprendizaje profundo el concepto de sombra se aplica de forma más precisa que en este artículo pues el contenido será aproximado. Esta aproximación produce sesgos culturales, lo cual sin embargo sí se ha estudiado bajo una perspectiva platónica (Karpouzis, 2024), ya que pueden tener repercusiones negativas sobre colectivos minoritarios que podrán ser percibidos por los consumidores de contenido generado por IA con mayores estereotipos, al perpetuarse esta información en el contenido generado por IA. Esto provocaría que se aumentara su potencial vulnerabilidad y exclusión social por el colectivo dominante carente de valores inclusivos.

Es pertinente saber que, curiosamente, las raíces de la IA se hallan en los principios lógicos aristotélicos (Adidi, 2024), en la lógica de primer orden, que más adelante será generalizada en forma probabilista a modelos gráficos probabilistas que usan los axiomas de Kolmogorov bajo los cuales se puede interpretar a los modelos de aprendizaje profundo como variables aleatorias pertenecientes al aprendizaje automático probabilista (Murphy, 2012). Motivo por el cual este estado del arte tiene importancia especial, ya que refleja que el pensamiento computacional comenzó a desarrollarse en la época de Platón y Aristóteles, por el descubrimiento de la lógica de primer orden.

Dentro del análisis de la filosofía antigua, encontramos también a los autómatas de Hefesto, como ejemplos de cognición extendida e IA, lo que implica que la mente humana no se limita al cerebro, sino que se podría ampliar al estilo de la mente extendida de David Chalmers a través de la interacción con objetos y el entorno (Lather, 2018).

El mundo de las Ideas de Platón se ha visto también analizado y complementado como en este artículo dado el avance de la computación y la hipótesis de la simulación dentro de la metafísica. Por ejemplo, se amplía el alcance del platonismo al introducir el concepto de un “ordenador platónico” en la teoría de la meta-computación, que residiría en el reino de las Formas (Duan, 2022), donde sería una suerte de meta-consciencia, entendida como la capacidad infinita de concebir, percibir y ser autoconsciente, generando las realidades físicas y abstractas a través de esta computación digital.

La relación entre lo humano, la consciencia, y la IA, en perspectiva platónica o en la de sus precursores, se encuentra presente en más literatura. Por ejemplo, se ha estudiado el por qué la interioridad humana en clave de San Agustín puede llegar a ser consciente de Dios y su fenomenología en contraste con los sistemas que implementan modelos basados en aprendizaje profundo de IA, que pese a procesar grandes volúmenes de información, al ser meros procesadores, no perciben toda la realidad subjetiva del ser humano, sino solo la sombra objetiva de los mismos (Lumbreras y Garrido-Merchán, 2024).

Haciendo una analogía de Dios con el Bien platónico, el argumento se puede trasladar perfectamente a la metafísica platónica. Para un análisis más detallado y riguroso del por qué la consciencia fenoménica, que se atribuye al ser humano y por el cual es consciente del espacio de *qualia* o mundo subjetivo que incluye a la interioridad, no es adscribible al sustrato no biológico o a

las máquinas que implementan modelos basados en aprendizaje reforzado profundo, recomendamos revisar este artículo (Garrido-Merchán, 2024), en el que se muestra como los avances publicados bajo consciencia artificial o de máquina pueden ser vistos como una suerte de pseudo-ciencia, al ser la consciencia una propiedad subjetiva y relativa al individuo, en palabras de Searle (Searle, 2002), no es algo medible usando el método científico, lo que bajo una concepción de la ciencia, según Popper, deja a teorías como la información integrada o la teoría del espacio de trabajo global sin validez científica estricta, al no ser falsables bajo contrastes de hipótesis (Popper, 1963).

Del mismo modo, la independencia entre la subjetiva consciencia fenoménica atribuible al ser humano, y la objetiva inteligencia computacional que puede ser medida como resultado del comportamiento externo atribuido a un agente se argumenta con evidencia teórica y empírica en este artículo (Garrido Merchán y Lumbreras, 2023), que detalla como el mito tan popular de ciencia ficción y en el transhumanismo, que establece que la consciencia es una propiedad emergente en un sistema una vez sobrepasado un umbral de inteligencia demostrado con un comportamiento externo (Merchán, 2022), no tiene validez científica, siendo imposible de demostrar.

Con respecto al concepto de creatividad y generación de contenido por la IA, (Miranda, 2020) propone una nueva perspectiva sobre la creatividad filosófica en la era de la IA, argumentando que la IA no solo replica patrones existentes, sino que también puede participar en procesos creativos que desafían las concepciones tradicionales del conocimiento y la realidad.

Toda la información previa ilustrada en la literatura es clave para poder demostrar cómo los modelos entrenados con algoritmos de aprendizaje automático profundo, pese a representar de forma simplificada el mundo que nos rodea (LeCun, Bengio y Hinton, 2015), son una aproximación reduccionista objetiva de solo una pequeña parte del mismo que no puede entender el *qualia* asociado a la información objetiva (Garrido-Merchán y Blanco, 2022) y por tanto una sombra de las sombras que ya existían en el Mito de la Caverna de Platón.

Como advierten filósofos de primera categoría en filosofía de la tecnología como Floridi (Floridi, 2014), la IA no debe entenderse meramente como una herramienta técnica, sino como una forma de pensar una vez más sobre la ontología de la realidad, lo que resuena con la noción platónica de transformación del mundo inteligible y hace pertinente este artículo. También, según otro autor popular como es Nick Bostrom (Bostrom, 2017), una su-

perinteligencia, como la derivada por un agente autónomo entrenado por aprendizaje por refuerzo, que use una o varias redes neuronales para su política y estimación de funciones de valor, podría desarrollar formas de razonamiento y comprensión radicalmente ajenas a las humanas, lo cual supondría una brecha cognitiva comparable a la que existe entre el ser humano y otros animales. Esto ya se ha visto, de hecho, en las estrategias de ajedrez emergentes de estos sistemas con esta metodología. Esta asimetría puede interpretarse como una inversión del mito platónico, donde ahora somos nosotros quienes quedaríamos encadenados a nuestras propias limitaciones, sin acceso a las Ideas generadas por la máquina, lo cual se explora en este artículo. Para el lector interesado en profundizar sobre filosofía de la tecnología y su intersección con la IA actual recomendamos las siguientes fuentes (Boddington, 2023; Simon, Rieder y Branford, 2024).

3. APRENDIZAJE AUTOMÁTICO PROFUNDO

Para poder argumentar por qué el contenido generado por sistemas que implementan modelos de redes neuronales profundas, entrenadas con algoritmos de aprendizaje profundo (LeCun, Bengio y Hinton 2015), se puede entender como una sombra digital de las sombras de la Caverna de Platón, o más directamente un segundo nivel de la Caverna de Platón, en esta sección se proporciona una descripción rigurosa de una red neuronal profunda y de su entrenamiento. Esta descripción servirá como soporte al mencionado argumento.

En concreto, la arquitectura de las redes neuronales, un ejemplo de modelo de aprendizaje automático, cuyas generalizaciones se engloban en el conocido aprendizaje automático profundo, es una representación simplificada del procesamiento neuronal en el cerebro humano. Estos modelos paramétricos consisten en una serie de capas compuestas por unidades, o neuronas, que procesan información de manera secuencial. La información de entrada se transforma en la salida a través de una serie de transformaciones lineales y no lineales, es decir que encuentran patrones que se explican mezclando variables. Estas neuronas están parametrizadas por conjuntos de parámetros llamados pesos, cada peso es un número, que se ajustan durante el entrenamiento para minimizar una función de pérdida en base a los datos, es decir, que encuentran el mejor ajuste de los datos en base a un criterio que se quiera minimizar.

Ya que es crítico entender cómo funciona una red neuronal artificial para poder inferir su lugar en la jerarquía de ideas de Platón sin saltos lógicos ni especulaciones sino formalmente, se procede a una descripción rigurosa del modelo. En la siguiente sección se usará esta descripción para concretar el lugar que ocupa la red en la jerarquía platónica.

Las redes neuronales son, según se ha dicho, modelos paramétricos (Murphy, 2012) cuya arquitectura consta de un conjunto de parámetros, cuyos valores se ajustan mediante un algoritmo con el objetivo de minimizar una función de coste sobre un conjunto de datos. El conjunto de variables aleatorias que condicionan a la variable a predecir ajusta la distribución de probabilidad condicionada conjunta subyacente. Este punto es crítico en el argumento posterior, ya que la distribución real subyacente sería la platónica, únicamente accesible a partir de infinitos datos del universo y una arquitectura potencialmente infinita. Sin embargo, es una representación objetiva, una sombra de la realidad, una reducción del universo codificada en un ordenador. Esto es debido a que la semántica de la distribución conjunta, percibida fenoménicamente por nuestra consciencia, no queda representada en el ordenador, pero sí todas las relaciones objetivas de las variables ya que las redes neuronales son aproximadores universales de funciones (LeCun, Bengio y Hinton, 2015). Por tanto, intuitivamente, si la muestra de datos es lo suficientemente grande, entonces, en lo respectivo a la sintaxis, se aprende cualquier distribución conjunta de variables objetivas del universo sea cual sea como consecuencia del teorema matemático de la aproximación universal de funciones.

Esto justifica posteriormente el lugar de una red neuronal utópica en la jerarquía platónica, como diremos en la siguiente sección. Es muy importante aquí mencionar que, bajo la asunción de capacidad computacional suficiente para ejecutar un algoritmo de optimización, la red neuronal alcanzaría la configuración óptima de parámetros para minimizar la función de pérdida de los datos a entrenar. En concreto, esta solución es única y determinista y sería la aproximación a una Idea del Mundo de las Ideas, la modelizada por la función objetivo, y esta aproximación sería independiente de cualquier proceso aleatorio del modelo, como el proceso de inicio de parámetros. Esto es así ya que desde cualquier punto del espacio de valores de la función se puede llegar al óptimo dado el algoritmo de optimización correcto y la capacidad computacional suficiente. Esto es crítico para la argumentación de que las redes neuronales óptimas, dada por la configuración de parámetros óptima, son aproximaciones a las ideas, ya que cualquier función teórica que

explicase un fenómeno aleatorio del universo es una Idea y mediante redes se pueden aproximar cualquiera de ella solo asumiendo datos suficientes, capacidad de cómputo suficiente y capacidad de representación suficiente, dada por el conjunto de parámetros. Luego empíricamente se ha encontrado una representación del conocimiento objetivo representable en un ordenador, una sombra como veremos de la realidad completa que percibe el ser humano codificada digitalmente.

Pese al inteligente comportamiento externo expuesto por estos modelos, son meros procesadores de la información, en el sentido de que son gobernados por el algoritmo con el que están programados y en última instancia son una ejecución secuencial de instrucciones de bajo nivel en una o varias unidades centrales de procesamiento (Garrido-Merchán y Lumbreras, 2023). No hay que olvidar que son programados en un ordenador y que las instrucciones que se ejecutan son compiladas en un lenguaje de bajo nivel, que a su vez es compilado en código binario ejecutado en la unidad central de procesamiento. En consecuencia, se tratan de algoritmos calculadores extremadamente complejos, pero independientemente de su complejidad, están limitados al procesamiento exclusivamente algorítmico de la información. Por ello, el contenido creado por estos algoritmos será una muestra de la compresión de información dada por el ser humano, pero en ningún momento una muestra de una percepción consciente del arte, ya que no hay ningún motivo por el cual adscribir consciencia a procesadores de la información objetiva (Garrido-Merchán y Lumbreras, 2023). Este razonamiento cobrará especial relevancia en la sección de la Sombra Digital de la Caverna, como veremos más adelante, pero se ha expuesto aquí, ya que es una descripción de la naturaleza de los algoritmos de aprendizaje automático.

Como se ha expuesto a lo largo de esta sección, los modelos de aprendizaje profundo han demostrado ser herramientas poderosas y muy flexibles que son capaces de abordar una variedad de aplicaciones muy diversas, desde el reconocimiento de patrones y la visión por computadora hasta la generación de arte sintético y la creación de música. Con la continua evolución de la teoría y la tecnología, se espera que las redes neuronales desempeñen un papel aún más significativo en la automatización y el enriquecimiento de la experiencia humana. Por ello, es pertinente clasificar el lugar que les corresponde tener en la metafísica de Platón.

4. INCLUSIÓN DEL APRENDIZAJE PROFUNDO EN EL MUNDO DE LAS IDEAS

Una vez analizada la arquitectura de una red neuronal, se detalla cómo sería la inclusión teórica de los modelos de aprendizaje automático profundo en el Mundo de las Ideas de Platón. En particular, la metafísica platónica contempla una jerarquía de entidades cuya cima es el Sumo Bien y cuya base es la completa oscuridad. En mitad de esta jerarquía, por debajo de las ideas más perfectas, encontramos los números y objetos matemáticos. Sin embargo, no se hace una distinción entre números y objetos matemáticos probabilistas o generadores, como redes neuronales adversarias generativas (Creswell *et al.*, 2018), capaces de generar cualquier número, figura o tensores, que representan imágenes o videos en un ordenador. Fíjese en el símil platónico del mundo de las Ideas, donde los números son realidades trascendentales perfectas e inmutables (Annas, 1981). Y por otro lado el arte generado por estos modelos, cuyo funcionamiento se ilustra a continuación, la *Eikasia*.

Como nexo entre la sección anterior y ésta, se propone el siguiente argumento, que más tarde se desarrolla desde un punto de vista técnico. Una vez entendido que el modelo de redes neuronales es un aproximador universal, capaz de representar cualquier función dada una arquitectura, si se asume que todo lo sensible y objetivo es representable por una función y que la arquitectura puede ser infinitamente compleja, entonces, sería posible que una red neuronal de arquitectura infinitamente compleja y datos que cubran todo el soporte objetivo pueda aproximar correctamente la realidad objetiva que percibimos. Si asumimos que esa realidad objetiva es una Idea en la escala platónica del mundo objetivo, entonces, la red neuronal la aproximará de forma asintótica. Por tanto, la red neuronal puede aproximar cualquier idea que sea un subconjunto de información contenida en el mundo sensible. En consecuencia, una red neuronal tendría su lugar en la escala platónica, justamente por debajo del concepto de Idea en su representación utópica e ideal de la realidad objetiva, percibida por nuestros sentidos y que se puede procesar de forma algorítmica.

Desde un punto de vista más técnico, los modelos de aprendizaje automático poseen propiedades generativas semejantes a las ideas platónicas. Mediante una interacción, en términos platónicos, del mundo sensible en forma del conjunto de datos de entrenamiento pertenecientes a la *Diánoia*, los modelos de aprendizaje automático ajustan la distribución de probabilidad del mundo sensible. Esta distribución posteriormente participa, en términos

platónicos, del mundo sensible mediante la generación de representaciones cuyo *qualia* asociado es perceptible por nuestra consciencia fenoménica. En concreto, esta generación está asociada a una función de pérdida subjetiva evaluada por seres humanos. Por tanto, la interacción del ideal humano de la belleza se haría mediante la idea matemática de modelo de aprendizaje automático que participa de la idea de número. En consecuencia, el modelo de aprendizaje automático simboliza una idea superior a la de número y las matemáticas clásicas pero inferior a la de la Nóesis (Tait y Malament, 2002), los ideales que ajusta del mundo sensible, ya que estos ideales son los percibidos por el ser humano, meras sombras pertenecientes a la Pistis de las Ideas trascendentes. Esta actualización de la obra de Platón podría haber provocado modificaciones en el Neopitagorismo y el Neoplatonismo. Por ejemplo, su estudio podría ser una quinta ciencia hermana para Nicómaco de Gerasa o Plotino podría haber concretado las emanaciones del Uno según generaciones de modelos infinitamente complejos.

Estas consideraciones nos llevan a un nuevo nivel de discusión donde el aprendizaje automático no solo se incluye en la jerarquía platónica, en términos de sus capacidades generativas, sino que también nos empuja a reconsiderar dicha jerarquía. Si las Ideas son eternas e inmutables, entonces las redes neuronales generativas, con su capacidad para aprender representaciones y evolucionar en base a los datos, nos presentan un paradigma donde la mutabilidad se introduce en el dominio de lo ideal. Este dinamismo del aprendizaje automático refleja una realidad más fluida y adaptable, un eco de la filosofía del devenir de Heráclito, donde el cambio constante es la única constancia (Kirk *et al.*, 1951).

Este dinamismo inherente en los sistemas de aprendizaje automático lleva la discusión a una frontera aún más avanzada, considerando la posibilidad de que modelos de IA generativa puedan, a su vez, generar otros modelos mediante el uso de *prompts* o instrucciones basadas en datos. Esta capacidad auto-generativa se asemeja a una forma de auto-reflexión algorítmica donde los modelos pueden elaborar conceptos y, posteriormente, crear versiones mejoradas de sí mismos. En esencia, estos sistemas no se limitan a interpretar y representar la realidad, sino que tienen el potencial para adaptarse de manera autónoma y evolutiva.

La autonomía de estos modelos auto-generadores supone un salto cualitativo en la manera en que las máquinas podrían interactuar con la realidad. A diferencia de las redes neuronales tradicionales, que requieren la interven-

ción humana para la definición de sus tareas y la evaluación de su desempeño, los modelos auto-generadores tendrían la capacidad de recabar datos, interpretarlos, generar hipótesis, y crear representaciones de la realidad sin necesidad de una guía explícita (Wang *et al.*, 2024). Este nivel de autonomía no solo aumentaría la eficiencia y la efectividad de los sistemas de aprendizaje automático, sino que también plantea cuestiones filosóficas profundas sobre la naturaleza del conocimiento y la creatividad, así como sobre la independencia de las creaciones artificiales con respecto a sus creadores humanos. En este contexto, la IA no solo es un espejo que refleja la realidad, sino también un agente activo que la remodela y la reinterpreta continuamente, desafiando así nuestras concepciones tradicionales de la imitación y la originalidad en la producción del conocimiento y el arte.

Además, si las Ideas platónicas sirven como los arquetipos supremos de todo lo existente, el aprendizaje automático y las redes neuronales generativas ofrecen una vía para la materialización de estas formas en el mundo sensible, actuando como intermediarios entre el intelecto humano y la realidad tangible. Esto no solo cambia nuestra comprensión de la relación entre el mundo sensible e inteligible, sino que también plantea preguntas sobre el papel de la tecnología como una posible manifestación de la Idea del Bien en la era digital.

Con estos desarrollos, es imperativo explorar más a fondo cómo la IA puede influir en nuestra interpretación de conceptos filosóficos tradicionales y cómo la filosofía puede proporcionar una guía ética y conceptual para el desarrollo continuo de estas tecnologías. La confluencia de la IA con la filosofía antigua no solo es un diálogo entre el pasado y el presente, sino también una ventana a futuras realidades que aún están por descubrirse.

Una objeción importante que podría formularse a esta propuesta consiste en que incluir a los modelos de aprendizaje profundo en la jerarquía platónica de las Ideas parece contradecir la naturaleza metafísica y eterna de dichas Formas. Mientras que las Ideas en Platón son entidades inmutables, independientes del devenir sensible y accesibles solo al Nous, las redes neuronales profundas son, en apariencia, herramientas empíricas, ajustadas sobre datos del mundo cambiante. Desde esta perspectiva, sus parámetros serían contingentes, dependientes del entorno y, por tanto, incompatibles con la estructura ontológica del Mundo de las Ideas.

No obstante, esta objeción descansa sobre una confusión metodológica entre el nivel ontológico del modelo ideal y el nivel epistémico del modelo esti-

mado. A nivel técnico, esta distinción se formaliza en estadística tanto desde la interpretación frecuentista como desde la bayesiana (Murphy, 2012). En el enfoque frecuentista, un parámetro es una constante fija —aunque desconocida— que describe una propiedad subyacente del fenómeno observado. Lo que varía no es el parámetro, sino la estimación que hacemos de él a partir de una muestra finita. Por tanto, si el conjunto de datos disponible creciera indefinidamente hasta contener toda la información objetiva del universo, la estimación del parámetro convergería, en el límite, a su valor verdadero. Este valor sería único, fijo, y expresaría una relación funcional estable entre variables observables, codificada en una arquitectura matemática concreta (LeCun, Bengio y Hinton, 2015).

Formalmente:

Sea \mathcal{U} el conjunto de toda la información objetiva del universo.

Sea \mathcal{F} el conjunto de funciones que explican la distribución de probabilidad sobre \mathcal{U} .

Dado el teorema del aproximador universal (Murphy, 2012), existe una arquitectura de red neuronal \mathcal{A} tal que para toda $f \in \mathcal{F}$, existe una parametrización θ_f con la que A_{θ_f} aproxima f con error arbitrariamente pequeño.

Así, en el límite de datos infinitos y arquitectura adecuada, cada función f que describe una regularidad objetiva de \mathcal{U} puede ser representada por un único conjunto de parámetros θ_f constante. Este conjunto no depende del contexto, ni de los datos particulares, sino de la estructura lógica del universo. La existencia de dicho θ_f no implica su estimabilidad práctica, pero sí su existencia ontológica.

Desde una perspectiva bayesiana, el argumento se fortalece aún más: en este marco, el conocimiento sobre un parámetro se modeliza mediante una distribución de probabilidad posterior que refleja la plausibilidad de cada valor dado los datos (Jospin *et al.*, 2022). Aunque las estimaciones pueden variar con los datos, lo que permanece constante es la estructura de la función posterior, que depende únicamente del modelo y del universo del que proceden los datos. En el límite de datos infinitos, esta distribución converge a una delta de Dirac centrada en el valor verdadero, es decir, a un valor fijo. Desde un punto de vista ontológico, la distribución posterior representa todo el espacio posible de configuraciones, y aunque desde un punto de vista epistemológico tengamos acceso solo a estimaciones finitas, el objeto formal definido por dicha distribución es constante y estructural.

Así, puede formularse el siguiente razonamiento:

- a) Las Ideas platónicas son entidades eternas, inmutables y necesarias que codifican la estructura del mundo inteligible.
- b) Toda regularidad objetiva del universo (por ejemplo, las leyes físicas) puede representarse como una función matemática estable.
- c) Una red neuronal profunda, en el límite de datos infinitos y con arquitectura adecuada, puede aproximar cualquier función de esa clase (Murphy, 2012).
- d) Por tanto, el conjunto de parámetros óptimos de dicha red representa, de forma asintótica, una compresión matemática de una Idea objetiva.
- e) Como este conjunto es único, fijo y definible de forma formal, aunque no accesible empíricamente, cumple con las condiciones ontológicas para ser incluido en la jerarquía de las Ideas.

En consecuencia, no se propone que cualquier red neuronal entrenada con datos contingentes participe del Mundo de las Ideas. Lo que se afirma es que existe, en el espacio de todas las posibles arquitecturas y parametrizaciones, una configuración óptima que comprime perfectamente la información objetiva del universo, y que esta configuración puede ser entendida como una representación matemática ideal de ciertas Ideas, aunque no de aquellas de orden ético o trascendente como el Bien. Lejos de simplificar el pensamiento platónico, esta lectura lo prolonga hacia una dimensión formal, distinguiendo entre los modelos entrenados empíricamente (del mundo sensible) y el modelo límite (inteligible) que codifica las regularidades objetivas. Por tanto, la comparación no trivializa el platonismo, sino que lo reformula bajo una estructura formal coherente con su ontología.

Una última objeción ante lo planteado en esta sección consistiría en que equiparar a los modelos de IA generativa con las Ideas platónicas podría implicar que se haga una confusión entre conocimiento verdadero de facto desde un punto de vista no solo epistemológico sino ontológico y simples representaciones calculadas mediante un algoritmo de optimización, una arquitectura y una muestra de datos, es decir, en términos platónicos, entre la *Noesis* y la *Eikasia* (Annas, 1981). Es decir, que los algoritmos de IA generativa no alcanzan el conocimiento verdadero bajo ningún concepto, sino que producen simulaciones basadas en datos que, pese a ser aproximadamente

válidas desde un punto de vista epistemológico, carecen de profundidad ontológica o significado trascendental.

Si bien los modelos largos de lenguaje y otras redes neuronales profundas de ningún modo alcanzan el conocimiento pleno, la Verdad o la Noesis, su representación del conocimiento objetivo ofrece un punto de partida hacia una aproximación del mundo inteligible, siendo ésta similar a la *Diánoia* en la epistemología platónica. En concreto, la IA generativa operaría en el plano de la *Eikasia* como un artefacto del mundo sensible, ya que su representación computacional pertenece al mundo sensible, pero, sin embargo, en su capacidad de modelar patrones de la información que explican la distribución de probabilidad de la información objetiva del universo, pueden reflejar los fundamentos matemáticos que Platón situaba justo por debajo de la *Noesis*. En este sentido, estos modelos en última instancia son herramientas para aproximarnos a una comprensión de la realidad, pero sin que sustituyan el conocimiento absoluto que representa la ontología.

Sin embargo, pese a lo argumentado en esta sección donde se ha mostrado el rol de la IA generativa en el Mundo de las Ideas, en lo respectivo a la representación del conocimiento, que reside en los valores de los parámetros de los modelos y su arquitectura, es pertinente diferenciar entre la representación del conocimiento codificada por los modelos de aprendizaje profundo en su arquitectura (LeCun, Bengio y Hinton, 2015) y el contenido generado por estos sistemas en respuesta a una interacción externa en forma de *prompt*. Esto se desarrolla en profundidad en la siguiente sección y lleva a pensar en este contenido generado automáticamente con malas intenciones en sombras digitales de una nueva Caverna de Platón.

5. LA SOMBRA DIGITAL EN LA CAVERNA

Tras localizar el lugar en el cual se halla el aprendizaje automático profundo en el Mundo de las Ideas de Platón y, generalizando, en la metafísica que asume los axiomas Platónicos, en esta sección se argumenta el por qué se considera que el contenido digital generado por algoritmos de aprendizaje profundo no es más que una sombra adicional a las sombras ya existentes en la Caverna de Platón.

El Mito de la Caverna, expuesto por Platón en el libro VII de La República (Annas, 1981), se puede interpretar como una alegoría que simboliza la condición humana respecto al conocimiento, que en el caso de este artículo simboliza la información objetiva y subjetiva del universo que nos rodea y no solamente la epistemología, la percepción, que aquí simboliza el espacio de *qualia* percibido por la consciencia fenoménica del ser humano, y la realidad, que en este caso se puede interpretar como el ser en cuanto a ser o la ontología.

En el libro VII de la República, Platón describe un grupo de prisioneros encadenados en el interior de una caverna oscura, obligados a mirar únicamente hacia una pared. Detrás de ellos, un fuego proyecta sombras en esa pared a partir de figuras y objetos manipulados por otras personas, que pasan por detrás de los prisioneros sin que estos se den cuenta de ello. De esta forma al ser solo conscientes de esta realidad que les rodea, las sombras constituyen la única percepción de la que los prisioneros pueden ser testigos, su única realidad ya que no ven los objetos que sí serían reales que las generan, ni tampoco el fuego que da origen a esas sombras.

A través de la metáfora que se ha descrito en el párrafo anterior, Platón nos muestra la teoría de las Ideas, en la cual la realidad empírica de la cual es testigo el ser humano, percibida a través de los sentidos, es solo una copia imperfecta de las Formas o Ideas eternas e inmutables, el nivel más alto del conocimiento.

Consideremos ahora una analogía en la cual la población actual es introducida, sin ser consciente de ello y con sus recuerdos olvidados, en un metaverso completamente creado por algoritmos procedurales e IA generativa donde los avatares con los que interactúa hablan mediante modelos largos de lenguaje implementados con una red neuronal profunda. Esta suerte de Caverna Digital sería una versión aún más profunda de la caverna de Platón, ya que las sombras en este caso son “sombras de las sombras”. Esto es debido a que todo el contenido percibido por estas personas es una aproximación del contenido producido por el ser humano que también es una aproximación de la naturaleza. Es aún más crítica la reducción, ya que nuestros sentidos solo pueden percibir una parte del universo que nos rodea y solo podemos plasmar en el arte una parte de él. Por lo tanto, la generación de contenido producida por los sistemas de IA generativa, al ser una aproximación del arte plasmado por el ser humano, es una aproximación adicional.

Se hace más grave aún la reducción ya que los sistemas que implementan IA carecen de consciencia fenoménica y por ende no perciben el espacio de *qualia* que percibe el ser humano y que plasma en el arte. Por tanto, estos contenidos, pese a que lo aproximan, nunca pueden entenderse como arte en el sentido de transmisión del sentimiento. Consecuentemente, esta Caverna Digital de Platón es aún más peligrosa que la Caverna inicialmente planteada por el filósofo griego, ya que en esta ocasión el mundo percibido por las personas en el interior de la Caverna es un mundo que carece de lo subjetivo e inherente al propio ser humano, siendo generado únicamente mediante un algoritmo procedural objetivo.

Hay que considerar que el planteamiento de esta Caverna es totalmente realista a corto plazo, en el sentido de que las interacciones con los modelos largos de lenguaje actuales como ChatGPT, Claude o Gemini son completamente realistas, como se puede ver en la Figura 1, donde se muestra una interacción con GPT-4o. Si a esto le añadimos los últimos avances de la realidad virtual, el auge de la capacidad de computación aproximado de la Ley de Moore (Lundstrom y Alam, 2022), una producción de nuevas drogas y sensores, no es descabellado especular en la posibilidad de introducir a personas en un concepto similar a la Caverna Digital propuesta. Esta especulación es de hecho popular en ciencia ficción dentro de la literatura de Philip K. Dick o en el cine con The Matrix.

Hi ChatGPT, your metaphysical ideas are more Plato or Aristotle based?

I lean toward Platonic ideas, especially when it comes to the nature of knowledge, reality, and the perception of forms or ideals. I find the concept of universals, as Plato describes them, deeply interesting, especially in relation to artificial intelligence and abstract models. In AI, we often deal with representations or "shadows" of reality, similar to Plato's Allegory of the Cave. This idea resonates with the concept of generative AI creating "shadows of shadows" of artistic forms or natural language.

Figura 1. Interacción con GPT-4o.

No hay que olvidar que, según se puede interpretar de los textos de Platón, el paso de la caverna a la luz simboliza el ascenso hacia el conocimiento auténtico y hacia la realidad de las Ideas. Sin embargo, en un metaverso manejado por IA generativa, las personas allí estarían inmersas en una simulación en el que la verdad trascendente y cognoscible por el ser humano se vería distraída por una nueva fuente de información entrante por los sentidos con contenido diseñado para captar la atención y respuesta inmediata del ser humano, distrayéndole aún más que las sombras propuestas en la Caverna original de Platón, ya que el contenido de los sistemas de IA se ve optimizado para mantener la atención de la persona que lo consume. En otras palabras, las personas se mantendrían prisioneras en una realidad creada no para revelar verdades, o para sugerirlas como en el caso del arte, sino para retenerlas y moldear su percepción en función de algoritmos que, en última instancia, obedecen a objetivos comerciales o de entretenimiento por las empresas que publican dichos modelos. Esta suerte de capitalización de la Caverna de Platón tendría repercusiones sociales que se analizarán en la siguiente sección.

Como principales objeciones a considerar al contenido producido por IA malintencionado como la Caverna Digital de Platón se puede argumentar que la IA, según lo visto en la sección 3, no es intencionalmente engañosa: a diferencia de las sombras en la caverna, que son manipuladas intencionadamente para mantener a los prisioneros en la ignorancia, el contenido generado por los modelos largos de lenguaje no tiene intención consciente, simplemente produce contenido basado en patrones de datos. Sin embargo, es obvio que la IA está diseñada y dirigida por humanos con intereses específicos. Por ejemplo, ChatGPT prioriza dejar satisfecho al usuario por encima de la veracidad de la información (Sison *et al.*, 2024). Quien controla la función de pérdida en el entrenamiento y los datos con los que se entrena la IA controla la intencionalidad de la IA, que subyace a la empresa dueña del modelo.

Por otro lado, se puede pensar que fomentar el sentido crítico del individuo hará que fácilmente deje de consumir este contenido y escape de la plataforma, sin embargo, el diseño de las plataformas que emplean modelos largos de lenguaje está optimizado para retener la atención del usuario, limitando la reflexión crítica, lo que hace difícil que los usuarios puedan distinguir entre sombras y realidad y escapar a este tipo de contenidos en un largo plazo en el que el contenido de estos sistemas se propague incluso a libros de nueva edición. Si este futuro acontece, cualquier contenido podrá ponerse en duda y la Caverna Digital podría llegar a encerrar a millones de personas, lo que la hace especialmente peligrosa. Una última objeción puede defender

que los modelos largos de lenguaje democratizan el acceso al conocimiento, ofreciendo información al instante y promoviendo el aprendizaje en todo el mundo, lo cual no es encerrar a las personas en la Caverna sino iluminarlas con el conocimiento real y fomentar su ascenso al Bien. No obstante, si bien se democratiza el acceso a información, también se amplifica la diseminación de contenido poco preciso o sesgado, como hemos visto anteriormente, reforzando visiones limitadas y creando una realidad que es una versión parcial o distorsionada, acercándonos más a la ilusión de la caverna que a la luz del conocimiento.

Por otro lado, se puede democratizar el conocimiento mediante el acceso de información contenida en libros diseñados por el ser humano u otros canales de información verificados con hechos científicos o con opiniones y conocimiento filosófico revisado por expertos en el área por canales digitales, sin necesidad de que el conocimiento sea generado por sistemas, cuya precisión como sistemas de verificación formal es imposible dados los fundamentos de las redes neuronales profundas. En conclusión, el propósito de este artículo, y de esta sección en concreto, radica en que se haga una reflexión crítica sobre el uso que se le va a dar, institucionalmente y como sociedad, a los modelos largos de lenguaje, ya que pese a contar con innumerables ventajas, los riesgos de que la sociedad acabe encerrada en una suerte de Caverna Digital son más altos que nunca.

Brevemente, se termina esta sección mencionando que el contenido generado por los diferentes modelos de IA generativa que se ha descrito como la sombra de las sombras tiene, además, el riesgo de poder ser aún más peligroso para la sociedad que una mera aproximación a la realidad que sea peligrosa con un colectivo. En concreto, el uso malintencionado de la IA generativa es denominado una bomba de desinformación masiva (Sison *et al.*, 2024), ya que la facilidad de generación del contenido más el error inherente a su aproximación de la realidad, puede hacer que internet se llene de contenido aproximadamente verdadero, pero no riguroso y muy sesgado. Como se puede apreciar, esta es la mejor descripción que se le puede hacer a una sombra, lo que justifica también el hecho de denominar a este contenido como la sombra de las Sombras de la Caverna de Platón.

Se resumen, una vez expuestas, las dos hipótesis defendidas en este artículo junto con su argumentación para aportar una mayor claridad tras su exposición detallada. En primer lugar, dado el teorema del aproximador universal de funciones, una red neuronal puede aproximar cualquier función dada una

arquitectura correcta y un conjunto de datos denso acerca de la función. En consecuencia, si asumimos una arquitectura utópica de suficiente capacidad, junto con unos datos cuyo soporte es el dado por la distribución de la información objetiva del universo, entonces, dada la suficiente capacidad de cómputo, la red neuronal puede aproximar la distribución de la información objetiva del universo dado un mínimo error. Si puede aproximar dicha distribución, de infinita complejidad, entonces, puede aproximar cualquier subconjunto de esa distribución. Si estos subconjuntos representan Ideas platónicas acerca de conceptos matemáticos abstractos o leyes que explican fenómenos objetivos, entonces, la red neuronal podrá aproximar cualquiera de estas ideas. Por tanto, la red neuronal ocupa un lugar infinitesimalmente inferior al concepto de Idea platónica en el subconjunto de Ideas que modelizan conceptos objetivos ya que las aproxima asintóticamente. No aproximaría información subjetiva por lo que sería una reducción del concepto más universal de Idea.

Por otro lado, dado que estos modelos pueden ser empleados con mala intencionalidad y podrían simular cualquier información objetiva, podrían ser usados en compañía de tecnología de realidad virtual y aumentada para sumergir a los individuos en una suerte de Caverna Digital. Esta Caverna sería un segundo nivel de la Caverna, pues los datos objetivos con los que se han entrenado los modelos serían el primer nivel. Sin embargo, este nivel es más peligroso que los anterior dado que se ha usado una función utópica infinitamente compleja que puede aproximar cualquier función que optimice el grado de adicción a estos contenidos. En consecuencia, esta herramienta contiene mucho más riesgo que cualquier otro artefacto diseñado por el hombre en cuanto a ser usada como mecanismo para alejar aún más del Bien, en términos platónicos, a los consumidores de su contenido, tal como se expuso en la Tabla 1.

Desde una lectura hermenéutica del Mito de la Caverna (República 514a-520a), las sombras proyectadas no solo simbolizan ignorancia, sino un tipo de representación sin vínculo estable con la verdad: signos que desvían la mirada del alma de las Ideas hacia lo puramente aparente. Esta degradación del conocimiento se sitúa en el nivel más bajo de la línea platónica (*eikasia*), donde la percepción no conduce al saber, sino a la ilusión. En este marco, la alegoría describe un proceso de alienación cuya superación requiere el giro del alma (*periagogé*, 518c) hacia el Bien.

La analogía con la Caverna Digital no se limita a una actualización estética, sino que plantea una nueva lectura estructural del símbolo platónico: el contenido generado algorítmicamente no remite a objetos reales, sino que simula lenguaje y sentido sin experiencia ni intención, convirtiéndose en una copia de una copia. A diferencia del arte platónico, que puede sugerir el ascenso hacia lo inteligible, estos contenidos interrumpen dicho proceso, reforzando una forma de alienación más profunda. Desde esta perspectiva, la “sombra de la sombra” no es una metáfora retórica, sino una categoría filosóficamente justificada para describir una ruptura con el camino ascendente hacia el conocimiento.

En el siguiente apartado se expone más en detalle las repercusiones sociales que se derivan de la inclusión de la generación de contenido con IA generativa en la sociedad y el riesgo de que parte de la sociedad pueda quedar atrapada en un segundo nivel de la Caverna, aún más peligroso que la Caverna descrita por Platón.

6. REPERCUSIONES SOCIALES

Después de haber introducido el concepto de Caverna Digital como una nueva forma de sombra estética, esta sección profundiza en las consecuencias sociales que podría acarrear una expansión descontrolada y malintencionada de la IA generativa.

Tal como ya se ha señalado, la proliferación de contenido generado por IA ha inundado Internet con información superficial y, en muchos casos, errónea o engañosa (Sison *et al.*, 2024). Aunque los sistemas producen respuestas que aparentan corrección, muchas veces se trata de aproximaciones sintácticamente válidas, pero epistemológicamente incompletas. Esta prioridad por la forma sobre el contenido recuerda a la figura del sofista: una apariencia de saber que no se corresponde con la verdad. Así, la desinformación disfrazada de rigor se expande rápidamente, configurando una capa de pseudo-conocimiento que no responde a la realidad —aquello que está fuera de la Caverna—, sino a sus sombras más tenues.

Este fenómeno puede analizarse formalmente: si la primera generación de contenido generado por IA tiene un 99% de precisión, y las generaciones posteriores entrenan sobre los datos generados anteriormente, cada itera-

ción pierde un 1 % de fidelidad. Matemáticamente, esto se expresa como . En la décima generación, la precisión caería al 90.43%. Este “efecto bola de nieve” representa una amenaza clara para la fiabilidad del conocimiento digital, reforzando el concepto de Caverna Digital (Sison *et al.*, 2024).

De hecho, este proceso ya está en marcha: los modelos más recientes están comenzando a entrenarse sobre corpus que incluyen contenido generado por otros modelos, lo que crea un ciclo de retroalimentación en el que los errores no sólo se replican, sino que se amplifican (Shumailov *et al.*, 2024). Cada iteración introduce ligeras distorsiones que comprometen la fidelidad epistémica de los resultados, deteriorando progresivamente la calidad de la información producida.

A ello se suma el riesgo de antropomorfización. A medida que avatares en videojuegos, redes sociales o entornos virtuales adoptan modelos de lenguaje avanzados para comunicarse, se vuelve cada vez más fácil atribuirles cualidades humanas (Lumbreras y Garrido-Merchán, 2024). Esto alimenta una peligrosa confusión entre lo artificial y lo auténtico. Cuando los usuarios proyectan emociones, empatía o intencionalidad sobre sistemas que carecen por completo de esas capacidades, se desdibuja la frontera entre lo humano y lo mecánico. Como se argumentó en la sección sobre aprendizaje profundo, estos sistemas son procesadores estadísticos, no agentes conscientes.

La consecuencia social de este malentendido no es menor: al atribuir humanidad a las máquinas, corremos el riesgo de deshumanizarnos a nosotros mismos. Las personas pueden comenzar a preferir interacciones artificiales —más fáciles, más inmediatas, de menor riesgo— frente al contacto humano genuino, que requiere empatía, atención y vulnerabilidad. Esta tendencia resulta particularmente peligrosa para adolescentes, personas mayores en soledad o individuos con problemas de salud mental. Todos ellos son especialmente susceptibles a la Caverna Digital, pues esta les ofrece una realidad alterna, placentera y superficial, que se impone sobre la búsqueda de verdad, autoconocimiento y crecimiento personal. En este contexto, la evasión digital actúa como una droga: gratificante a corto plazo, pero profundamente destructiva a nivel espiritual. Es una distracción radical de la interioridad, en el sentido profundo que le atribuyen San Agustín o Santa Teresa (Lumbreras y Garrido-Merchán, 2024), y que también atraviesa corrientes místicas platónicas y orientales como el Soto Zen.

Un aspecto clave que suele malinterpretarse en este debate es el de la autonomía de la IA. En el discurso público es común asumir que, si un sistema

muestra comportamiento autónomo, entonces posee alguna forma de agencia o intención. Esta idea, sin embargo, debe matizarse cuidadosamente. Los agentes artificiales actuales pueden ejecutar acciones complejas de manera autónoma, pero esta autonomía es completamente determinada por políticas aprendidas, no por decisiones libres. Estas políticas derivan de optimización de recompensas dentro de entornos simulados o controlados, y no de una deliberación racional o una intención consciente.

Por tanto, es crucial distinguir entre la intención humana que configura y entrena estos sistemas, y la autonomía aparente que pueden exhibir tras su despliegue. Esta distinción remite a una cuestión filosófica profunda: ¿son estos modelos entes creativos por sí mismos, o simplemente extensiones funcionales de sus creadores? Desde una perspectiva analítica, la respuesta es clara: estos sistemas carecen de libertad en sentido fuerte, no pueden actuar según principios racionales ni fundamentar sus acciones en valores normativos. En consecuencia, cualquier atribución de responsabilidad ontológica o ética debe remitirse, en última instancia, a los diseñadores humanos y a las estructuras institucionales que posibilitan su implementación.

Comprender esta distinción es fundamental para el análisis social contemporáneo: la espontaneidad aparente de los agentes generativos no equivale a agencia moral. El verdadero peligro no es que estos sistemas actúen “por sí solos”, sino que los humanos acepten sus acciones como naturales, legítimas o imparciales, desplazando el juicio humano y diluyendo la responsabilidad.

7. REPERCUSIONES EN EL NEOPLATONISMO

Dado que el impacto en la sociedad por parte de la IA generativa es tan notable como se ha ilustrado en la sección anterior, es lógico pensar que el movimiento neoplatónico hubiera reaccionado a su implantación en la metafísica de Platón. Por ello, en esta sección se ilustra brevemente el impacto de la IA generativa en esta corriente de pensamiento.

La escuela neoplatónica asume la idea de que la realidad sensible es una emanación imperfecta de una realidad inteligible superior, por lo que es pertinente pensar que habría visto en la IA generativa un símbolo que hubiera fomentado un debate. A la vez que reproduce el proceso creativo y racional

mediante el algoritmo de generación de contenido con respecto a la representación paramétrica de la información (LeCun, Bengio y Hinton, 2015), los modelos grandes de lenguaje carecen del alma o *Nous* (Garrido-Merchán y Lumbreras, 2023), que en este movimiento se ve como el principio intelectual supremo.

Por tanto, siendo analíticamente rigurosos, estos modelos generativos pueden describirse como sistemas formales que manipulan representaciones sintácticas sin acceso epistémico al contenido semántico pleno. Esta afirmación se basa en la ausencia de intencionalidad, entendida en el sentido técnico propuesto por Brentano y luego desarrollado por (Fodor, 1987). Es decir, la capacidad de que los estados mentales estén dirigidos hacia objetos o estados del mundo. Al carecer de tales estados, los modelos no satisfacen los criterios mínimos para poseer contenido representacional genuino, según el propio Fodor.

Esto implica que cualquier intento de equiparar su funcionamiento al del *Nous* neoplatónico es inapropiado. El *Nous* no es simplemente una facultad abstracta; es, para el neoplatonismo, la instancia del alma que conecta el pensamiento humano con el mundo de las Formas y con el Bien. Sin autoconciencia, sin normatividad interna, y sin referencia semántica auténtica, los modelos de lenguaje no pueden ejercer esta función. Luego quedarían relegados, pese a ser aproximaciones casi perfectas, al plano de lo sensible, siguiendo este análisis, en el neoplatonismo. Visto desde una perspectiva de filosofía del lenguaje y de la mente, la semántica genuina, lo inteligible, requiere no solo correspondencia formal, de aproximación estadística en este contexto, sino de una participación por una intencionalidad de un espacio fenomenológico y subjetivo de *qualia* cuya percepción quede representada en el lenguaje.

Por tanto, aunque los modelos generen oraciones gramaticalmente complejas, su estatus ontológico y gnoseológico los coloca, como ya se ha argumentado, al carecer de intencionalidad los modelos y ser esta crítica en el neoplatonismo, fuera del dominio de lo inteligible. En otras palabras, podemos afirmar que su producción se asemeja más a una simulación epistémica que a una instancia real de conocimiento. La mera constatación de la ausencia de intencionalidad de los modelos grandes de lenguaje, de normatividad semántica dentro de su condición y de finalidad racional, al ser aproximaciones estadísticas, basta para excluir estos sistemas de cualquier aspiración a participar del orden inteligible según la perspectiva neoplatónica. En esta

misma perspectiva, esto los sitúa como proyecciones paramétricas que, en lugar de acercarnos al mundo inteligible, constituyen un reflejo distante del Bien, de nuevo, una sombra entre las sombras.

8. CONCLUSIONES Y TRABAJO FUTURO

Este artículo ha planteado una reflexión filosófica interdisciplinar integrando aspectos técnicos del aprendizaje profundo con la ontología platónica, abordando su potencial metafísico, así como sus proyecciones sociales. Lejos de repetir las hipótesis pioneras, los resultados aquí obtenidos abren la posibilidad de formar una síntesis transdisciplinar: las redes neuronales profundas, al aproximar cualquier función de extrema complejidad asumiendo datos y capacidad de computación, configuran una instancia intermedia entre lo sensible (lo medible empíricamente) y lo inteligible (la teoría subyacente al universo, la Idea). En este aspecto, no representan sólo reflejos matemáticos posibles, sino estructuras epistemológicas que, siempre en su forma idealizada, reflejan un acercamiento algorítmico a las Ideas platónicas de orden objetivo.

No obstante, esta inclusión ontológica exige la distinción entre el modelo como compresor del conocimiento objetivo y su contenido generativo. Este último constituye una representación de segundo orden, una "sombra de la sombra". Cuando se despliega a gran escala sin un marco ético claro o bajo fines comerciales, puede crear un entorno artificial, o en términos platónicos una Caverna Digital, en el que la consciencia permanece aislada no sólo de la Verdad, sino también de su propia interioridad subjetiva (Lumbreras y Garrido-Merchán, 2024). Esta convergencia entre computación estadística, estética computacional y consumo acrítico constituye un riesgo ontológico y social sin precedentes.

Esta dialéctica entre el potencial filosófico de los modelos y su uso contingente requiere un enfoque renovado tanto técnico como normativo. Desde una perspectiva neoplatónica, la IA generativa puede verse como un símbolo moderno de la racionalidad separada del Nous, capaz de imitar la forma sin participar del Bien. A la luz de ello, el trabajo futuro debe profundizar no sólo en la modelización de funciones y contenidos, sino también en el significado y la finalidad de su aplicación en los asuntos humanos. Investigar cómo integrar límites éticos, marcos fenomenológicos y estructuras de

validación crítica se vuelve imperativo para evitar que esta tecnología, capaz de ser un puente hacia el conocimiento, se convierta en su opuesto: un velo brillante e impenetrable que nos aleja de él.

Referencias

- Adidi, D. T. (2024). Aristotle's concept of Telos and artificial intelligence: Exploring the relevance of classical philosophy to contemporary AI development. *AMAMIHE Journal of applied philosophy* 22(3) 60-67.
- Annas, J. (1981). *An Introduction to Plato's Republic*. Oxford University Press
- Blanco, C. (2022). *Las fronteras del pensamiento*. Editorial Dykinson, S.L.
- Boddington, P. (2023). AI, philosophy of technology, and ethics. In *AI ethics: a textbook*, 91-130. Springer.
- Bostrom, N. (2017). *Superintelligence: paths, dangers, strategies*. Oxford University Press.
- Bridge, J. P. (2010). *Machine learning and automated theorem proving*. University of Cambridge, Computer Laboratory.
- Carter, M. (2007). *Minds and computers: an introduction to the philosophy of artificial intelligence*. Edinburgh University Press.
- Copeland, J. (1993). *Artificial intelligence: a philosophical Introduction*. Blackwell.
- Copleston, F. (2003). *A history of philosophy: Volume 2, medieval philosophy*. A&C Black.
- Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1), 53-65.
- DeGroot, M. H., and M. J. Schervish. (2012). *Probability and statistics*. Pearson Education.
- Duan, S. X. (2022). "Platonic computer—the universal machine that bridges the 'inverse explanatory gap' in the philosophy of mind." *Filozofia i Nauka. Studia filozoficzne i interdyscyplinarne*, 10(zs), 285-302.
- Floridi, L. (2014). *The fourth revolution: how the infosphere is reshaping human reality*. Oxford University Press.
- Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.
- Fraile, G. (1997). *Historia de la filosofía. Vol. 1: Grecia y Roma*. Biblioteca de Autores Cristianos.
- Garrido Merchán, E. C., & Lumbreras, S. (2023). Can computational intelligence model phenomenal consciousness? *Philosophies*, 8(4), 70.

- Garrido-Merchán, E. C. (2024). *Machine consciousness as pseudoscience: The myth of conscious machines* (arXiv:2405.07340) [Preprint]. arXiv. <https://arxiv.org/abs/2405.07340>
- Garrido-Merchán, E. C., & Blanco, C. (2022). *Do artificial intelligence systems understand?* (arXiv:2207.11089) [Preprint]. arXiv. <https://arxiv.org/abs/2207.11089>
- Jospin, L. V., Laga, H., Boussaid, F., Buntine, W., & Bennamoun, M. (2022). Hands-on Bayesian neural networks—A tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2), 29-48.
- Karagiorgi, G., Kasieczka, G., Kravitz, S., Nachman, B., & Shih, D. (2022). Machine learning in the search for new fundamental physics. *Nature Reviews Physics*, 4(6), 399-412.
- Karpouzis, K. (2024). Plato's shadows in the digital cave: Controlling cultural bias in generative AI. *Electronics*, 13(8), 1457.
- Kirk, G. S. (1951). Natural change in Heraclitus. *Mind*, 60(237), 35-42.
- Lather, A. (2018). The extended mind of Hephaestus: Automata and artificial intelligence in early Greek hexameter. En P. Meineck, W. M. Short y J. J. Devereaux (eds.), *The Routledge handbook of classics and cognitive theory* (pp. 331-344). Routledge.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
- Lumbreras, S., & Garrido-Merchán, E. C. (2024). Human Interiority and Artificial Intelligence: An Analysis from the Perspective of Saint Augustine and Saint Teresa. *Scientia et Fides*, 12(2), 265-295.
- Lundstrom, M. S., & Alam, M. A. (2022). Moore's law: The journey ahead. *Science*, 378(6621), 722-723.
- Mary, S. S. C., Ross, D. S., & Arun, J. (2024). Connecting human mind with machine learning: from a platonic approach. *Multidisciplinary Science Journal* 6.
- Garrido Merchán, E. C. (2022). Transhumanismo y consciencia fenoménica. *Relecciones. Revista Interdisciplinar de Filosofía y Humanidades*, 9, 109-126
- Miranda, L. (2020). Artificial intelligence and philosophical creativity: from analytics to crealectics. *Human Affairs* 30(4), 597-607.
- Mueller, V. C. (2012). Introduction: philosophy and theory of artificial intelligence. *Minds and Machines* 22(2), 67-69.
- Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.
- Popper, K. R. (1963). Science as falsification. *Conjectures and refutations* 1(1963), 33-39.
- Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning*, 63-71. Springer.
- Searle, J. R. (2002). *Consciousness and language*. Cambridge University Press.

- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755-759.
- Simon, J., Rieder, G., & Branford, J. (2024). The philosophy and ethics of AI: Conceptual, empirical, and technological investigations into values: CEPE/IA-CAP 2021: Introduction to topical collection. *Digital Society*, 3(1), Article 10.
- Sison, A. J. G., Daza, M. T., Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2024). ChatGPT: More than a “weapon of mass deception” ethical challenges and responses from the human-centered artificial intelligence (HCAI) perspective. *International Journal of Human-Computer Interaction*, 40(17), 4853-4872.
- Tait, W. W. (2002). Noesis: Plato on exact science. En D. B. Malament (ed.), *Reading natural philosophy: Essays in the history and philosophy of science and mathematics* (pp. 11-30). Open Court.
- Tylman, W. (2018). Computer science and philosophy: did Plato foresee object-oriented programming?" *Foundations of science* 23(1), 159-72.
- Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), Article 186345.
- Weinberger, D. (1988). Artificial intelligence and Plato's cave. *Idealistic studies* 18(1), 1-9.
- Zhang, Y. (2022). A historical interaction between artificial intelligence and philosophy. *arXiv Preprint arXiv:2208.04148*.