Sellars, Kant y Hegel: Reflexiones en torno a los Grandes Modelos de Lenguaje

Xavier Aranda Arredondo
(Universidad de Guanajuato, México)

ABSTRACT.
The pragmatist philosopher W. Sellars proposed in his essay "Philosophy and the scientific image of man" that it would be science, through its theoretical work, that could account for the place of human beings in the world by integrating the conceptual framework of human beings and their relationships into a common ontology. We propose to use large language models (LLMs) such as ChatGPT 3.5, as the clearest attempt at a theory that could integrate such a conceptual framework of human beings. Given Sellars's interest in the philosophies of Kant and Hegel, we will make a critique of the ambitions of LLMs in the Sellarsian terms indicated, and on the possibility of them transitioning from models of language to models of knowledge. Our critique will focus on the possibility of a criterion of unity (such as the use of tools that aid in inferential reasoning) to avoid hallucinations, as well as the main driver behind the possible transition to knowledge models. We will point out that this effort lies in the position that Hegel calls subjective idealism, showing that a principle of explanation cannot be predictive. This will result in the impossibility of integrating the common ontology proposed by Sellars.

KEY WORDS:
Artificial intelligence, Metaphysics, Epistemology, German idealism, Naturalism

RESUMEN.
El filósofo pragmatista W. Sellars propuso en su ensayo "Philosophy and the scientific image of man" que sería la ciencia a través de su labor teórica, la que podría dar cuenta del lugar del ser humano en el mundo a través de integrar el marco conceptual de los seres humano y sus relaciones a una ontología común. Proponemos emplear el ejemplo de los grandes modelos de lenguaje (LLM) como ChatGPT 3.5, como el intento más claro de una teoría que pudiera integrar dicho marco conceptual de los seres humanos. Dado el interés de Sellars en las filosofías de Kant y Hegel, realizaremos una crítica a las ambiciones de los LLM en los términos Sellarsianos señalados, y sobre la posibilidad de que estos transicionen de modelos de lenguaje a modelos de conocimiento. Nuestra crítica se centrará en la posibilidad de un criterio de unidad (como el empleo de herramientas que auxilien en el razonamiento inferencial) con el fin de evitar las alucionaciones, así como el principal motor detrás del posible paso a los modelos de conocimiento. Señalaremos que este esfuerzo se encuentra en la posición que Hegel denomina idealismo subjetivo, mostrando que un principio de explicación no puede ser predictivo. Esto tendrá como consecuencia la imposibilidad integrar esa ontología común propuesta por Sellars.

PALABRAS CLAVE:
Inteligencia artificial, metafísica, epistemología, idealismo alemán, naturalismo

Introducción

Wilfrid Sellars fue un filósofo con gran influencia en el pensamiento analítico norteamericano de la segunda mitad del siglo XX a nuestra fecha, quizás mejor conocido por sus reflexiones pragmatistas en torno a la mente y el lenguaje a la par de un naturalismo epistemológico. Además, ha sido un pensador clave en el acercamiento contemporáneo de la filosofía analítica a la filosofía de Kant y Hegel en particular. Figuras como Richard Rorty, Robert Brandom, y John McDowell han comentado sobre esta influencia[1]. Las lecturas de la filosofía de Hegel, acompañadas de la mano de Sellars, coinciden en una interpretación no metafísica de su sistema. Gran parte de esto se puede adjudicar al naturalismo de Sellars el

[1] Rorty 1997, McDowell 2009, y más recientemente Brandom 2019.

cual plantearemos desde una tesis crucial hallada en el ensayo "Philosophy and the scientific image of man" de 1962. Esta tesis sostiene que la dimensión ontológica de la existencia humana habría de incluirse (si no reducirse) al marco explicativo de la ciencia, aunque manteniendo sus características específicas, logrando conciliar la brecha epistémica entre ambas. En el anterior sentido el enorme interés que despiertan los grandes modelos de lenguaje (y las redes neurales en sus diversas aplicaciones), sostendremos, ilustran una alternativa al acercamiento fisicalista reduccionista habitual a solucionar dicha brecha desde una posición naturalista. El éxito de estos depende de la comprensión no meramente del lenguaje, sino de las características epistémicas e inferenciales de la ontología de los seres humanos y sus relaciones, donde esta discusión a la par del interés de Sellars en Kant y Hegel permite trazar una línea argumentativa pertinente para la epistemología contemporánea. De tal suerte, propondremos entender el dilema epistemológico en el que se encuentran los grandes modelos de lenguaje, remitiéndonos a uno de los más grandes obstáculos que encontrara el sistema crítico Kantiano, de ahí ilustraremos la crítica Hegeliana, esbozando dos puntos cruciales para la posibilidad de dicho análisis.

Wilfrid Sellars y el problema de la imagen del hombre en el mundo

El naturalismo en la epistemología contemporánea es un término comúnmente asociado con W. V. O. Quine, en este sentido, el naturalismo de Sellars no es tan distinto al del primero, quien abogara por abandonar el empleo de términos normativos en la epistemología con la finalidad de adoptar un descriptivismo teórico conducente a una reconstrucción racional de las prácticas científicas (Quine, 1969, p. 78). Quizás, la separación entre dichos pensadores reside en cómo es que ambos llegaron a sus conclusiones, donde Sellars ilustra una diferencia de fondo en cómo la filosofía, y por ende la epistemología, no puede cumplir su meta: encontrar un principio de objetividad que muestre la continuidad con la empresa científica.

La filosofía de Sellars se caracteriza por un tipo de holismo apoyado en un ataque al fundacionalismo epistemológico que depende de algún tipo de contenido dado como lo son "…contenidos sensibles, objetos materiales, universales, proposiciones, conexiones reales, primeros principios, etc." (Sellars, 1991, p.

72)[2]. Nuestro autor concluye, para que dichos contenidos puedan ser dados necesariamente no podrían ser útiles epistemológicamente, mientras que, si son útiles en el anterior sentido, no podrían ser realmente dados. Es una conclusión hegeliana[3]: el resultado del análisis filosófico muestra que todo contenido es mediado.

Sin embargo, Sellars difiere de Hegel en un aspecto clave, ya que apunta a una teoría positiva del conocimiento cuya estructura no sea inferencial (deVries, 2016)[4]. Toma como paradigma las creencias relacionadas con la introspección, la percepción, y la memoria, donde estas creencias además deben de poseer la condición que el sujeto epistémico "sepa" que son confiables, lo cual no sólo significa que los sujetos epistémicos puedan ser capaces de emitir juicios sobre juicios:

> El punto esencial es que, al caracterizar un episodio o un estado como un estado de saber, no estamos dando una descripción empírica de ese episodio o estado, sino que lo estamos colocando en el espacio lógico de las razones, de justificarlo y ser capaz de justificar lo que uno dice. (Sellars, 1991, p. 168)

Operar en el espacio lógico de las razones es estar familiarizado con el discurso normativo, con sus estándares de lo correcto y apropiado. Esta es una caracterización holista muy semejante a la de juegos del lenguaje en el segundo Wittgenstein, y que disuelve la posibilidad de aislar los estados cognitivos. El problema se torna en por dónde comenzar a adquirir semejantes estados cognitivos evitando la circularidad o el regreso. Según deVries (2016), los adquirimos poco a poco mediante hábitos y disposiciones que interrelacionados garantizan la aplicación del lenguaje normativo de lo cognitivo. Para Sellars

> Uno parece forzado a escoger entre la imagen de un elefante que descansa en una tortuga (¿qué sostiene a la tortuga?) y la imagen de una gran serpiente Hegeliana del conocimiento con la cola en su boca (¿dónde comienza?). Ninguna servirá. Para el conocimiento empírico, como para su extensión

[2] De aquí en adelante, la traducción de las citas textuales es de mi autoría salvo cuando se explicite lo contrario.

[3] Hay que recordar que Sellars refiere a su texto como una serie de meditaciones hegelianas evocando tanto a Descartes como las lecciones en Paris de Husserl en la sección no. 20 de *Empiricism and the philosophy of mind* (1997, p. 148).

[4] La epistemología Hegeliana se constituiría en su paso del universal al particular y singular, ejemplificados en el paso del concepto al juicio y silogismo en la *Ciencia de la Lógica*, es decir, bajo la lectura de Sellars, inferencialmente. Sellars atribuye esta idea a Kant "los conceptos son ítems que esencial y no accidentalmente pueden ocurrir sólo en juicios, y los juicios a su vez son ítems que esencial y no accidentalmente pueden ocurrir sólo en argumentos" (Sellars, 2007, p. 4).

> sofisticada, la ciencia, es racional, no porque tenga un fundamento sino porque es una empresa que se auto corrige y que puede poner en peligro cualquier aseveración, pero no todas al mismo tiempo. (Sellars, 1991, p. 128)

Esta idea de la ciencia como una empresa que se autocorrige es la tesis naturalista que mencionamos y aparece en el ensayo "Philosophy and the scientific image of man" (1991), donde su finalidad es mostrar la disyunción entre el conocimiento filosófico y el conocimiento científico.

Sellars argumenta que el marco referencial de la filosofía puede entenderse como la "filosofía perene" cuya intención en términos de la analogía teoría y práctica, es el sentido de la totalidad, esto es, la descripción del mundo y su relación con el ser humano bajo una misma ontología.

En contraste, según Sellars, la imagen científica del hombre-en-el-mundo es una

> [...] concepción que se limita a sí misma a lo que las técnicas correlacionales pueden decirnos acerca de los eventos perceptibles exterior e interiormente y que postula objetos imperceptibles y eventos con el propósito de explicar correlaciones entre perceptibles [...] Nuestro contraste, entonces, es entre dos constructos idealizados: (a) el refinamiento correlacional y categorial de la 'imagen original', a cuyo refinamiento llamo la imagen manifiesta; (b) la imagen derivada de los frutos de la construcción teórica postulacional que llamo la imagen científica. (Sellars, 1991, p.19)

Esta relación entre la imagen manifiesta y científica pone en cierta paradoja a la filosofía, en ya que, en su búsqueda por la imagen de la totalidad, necesita de la objetividad de las diversas disciplinas que abonan a su vez verdades hacia esta empresa. La imagen de esas otras disciplinas, en cambio, según Sellars no busca el todo, sino que está satisfecha dentro de sus confines y coexiste con las otras disciplinas que no se reemplazan entre sí, puesto que incluso un eliminacionismo reduccionista estaría de la mano de una empresa meta-teórica más afín a la propuesta de un epistemólogo (sería a final de cuentas, un compromiso metafísico más allá de las pretensiones de la idealizada imagen científica).

El hombre-en-el-mundo, por tanto, necesita fusionar estas imágenes en una imagen estereoscópica al modo en que los ojos izquierdo y derecho en su conjunto nos permiten observar con profundidad. ¿Cómo realizar esta síntesis? Para Sellars, ambas no pueden coexistir, por ejemplo, proponiendo una imagen intermedia que regule a ambas imágenes, ya que esto preservaría su separación sin solucionar la circularidad que exhibe cómo una presupone a la otra, en este sentido una de las

dos imágenes debería tener primacía sobre la otra.

En el resto del texto ejemplificará los problemas que surgen de dar primacía a la imagen manifiesta o a la imagen científica, en específico resulta interesante la postura en torno a esta última, donde nos señala que incluso asumiendo un tipo de reduccionismo, la complejidad de la realidad de los individuos vistos desde todos sus aspectos como individuos libres, que forman parte de una sociedad, además de poseer la complejidad propia de ser entidades físicas; la imagen científica sería incapaz de dar cuenta del irreducible núcleo del marco conceptual de las 'personas'.

Es en este último sentido que señala la posibilidad de la imagen científica de dar cuenta de la imagen manifiesta, a través de integrar a las personas o a la comunidad

> [...] el marco conceptual de las personas es el marco conceptual con el que pensamos de los demás en tanto que comparten las intenciones de la comunidad que provee del conjunto de principios y estándares (arriba de todos, aquellos con los que el discurso significativo y la racionalidad misma es posible) dentro de los que vivimos nuestras propias vidas individuales. Entonces el marco conceptual de las personas no es algo que debe ser reconciliado con la imagen científica, sino algo que debe ser unido. Entonces, para completar la imagen científica necesitamos enriquecerla no sólo con meros modos de decir lo que es el caso, sino con el lenguaje de las intenciones comunitarias e individuales, de modo que al construir las acciones que pretendemos hacer en términos científicos, directamente las relacionemos con el mundo como es concebido por la teoría científica según nuestros propósitos, y hagamos el mundo ya no un apéndice extraño al mundo en el que realizamos nuestras vidas. Podemos, por supuesto, justo como las cosas se presentan ahora, realizar esta directa incorporación de la imagen científica en nuestro propio modo de vida sólo en la imaginación. Pero hacerlo es, sí sólo imaginariamente, trascender el dualismo de las imágenes científica y manifiesta del hombre-en-el-mundo. (Sellars, 1991, p. 40)

La pregunta que realizamos a continuación se remite básicamente a señalar cómo luciría una teoría científica que pudiera integrar el marco conceptual de los individuos, es decir, un marco conceptual que posibilite el discurso significativo o la racionalidad tal como enuncia Sellars, pero prescindiendo justamente de la idea de totalidad de la imagen manifiesta.

Proponemos que los grandes modelos de lenguaje (large language models o LLM) presentan si no el primer ejemplo de una teoría como esta, quizás el ejemplo

más exitoso dada su prevalencia en el imaginario popular contemporáneo. En este sentido, habrá qué aclarar dos cosas, cómo es que funcionan estos modelos, y segundo, en qué consiste esta ausencia de la idea de totalidad donde mostraremos, posee repercusiones severas contra la postura de Sellars, visibles desde el tratamiento de la filosofía de Hegel.

Los grandes modelos de lenguaje

Sostenemos que los grandes modelos de lenguaje (LLM de aquí en adelante) representan adecuadamente la idea de Sellars en tanto que estos buscan integrar el marco conceptual de los individuos, es decir, el lenguaje natural justo como se practica, y no bajo la idea de una necesaria depuración lógica que sustraiga la estructura veritativa, que en consecuencia evite las confusiones entre teorías, como lo plantearía el positivismo lógico[5].

Los LLM pretenden hablar justo como los individuos hablamos, no existe pues, inicialmente, una pretensión de alterar el lenguaje, o encontrar un lenguaje mejor al lenguaje natural con el fin de adoptarlo en nuestra empresa epistemológica. A esto nos referimos con integración.

La cuestión se torna en ¿cómo es que funcionan los LLM? Primero es necesario esclarecer a qué se refieren por "lenguaje", mientras que los LLM tratan de dar cuenta cómo hablamos, no lo hacen aprendiendo las reglas gramaticales, es decir, la comprensión del lenguaje no es precisamente lingüística. Lo aprenden de manera no muy diferente a la que un niño lo haría, observando múltiples instancias del uso hasta sustraer una generalización. En este sentido un LLM como ChatGPT 3.5 no tiene un conocimiento explícito de las reglas gramaticales, pero de algún modo, durante su alimentación o entrenamiento, las descubre de manera

[5] Un ejemplo de esto lo encontramos en R. Carnap quien intentó proponer un principio de tolerancia que redirigía el nivel de análisis del sintáctico al semántico, señalando la necesidad de que un lenguaje específico esté conformado por reglas explícitas, elaborando una distinción entre el lenguaje elegido y sus posibles interpretaciones (Awodey & Carus 2007, p. 191). La verdad de una premisa es el resultado de las reglas o elementos para la evaluación de la verdad misma según se presuponen en el lenguaje sobre el que se postula la teoría (Hylton, 2004, p. 128-131). De tal suerte, los errores de las teorías son propiamente los errores del lenguaje, es decir, de la ineficiencia o carencia de un lenguaje -que bien podría ser artificial- para poder proveer de una evaluación satisfactoria de la verdad de una teoría (Hylton, 2004, p. 128-131). Las proposiciones analíticas no podrían ser modificadas sin modificar el lenguaje de base, mientras que las sintéticas no necesitarían modificar ese lenguaje de base para ser cambiadas (Cerath, 2004, p. 56-58).

implícita (Wolfram, 2023). La aproximación, por lo tanto, no es superponer una estructura, sino asumir que esta existe y capturarala, cumpliendo con la meta de integrar el marco conceptual de los individuos, bajo una ontología común que mostraría el lugar del ser humano en el mundo.

¿Por qué los grandes modelos de lenguaje no se entrenan con las reglas gramaticales? La razón muestra la ontología de fondo en el modelado conceptual. Los LLM son una implementación avanzada de las redes neurales, que como su nombre lo indica, son modelos de cómo se supone que funcionan las neuronas, entendiéndolas como nodos interconectados cuya función es el procesamiento de datos. Es decir, son neuronas idealizadas y no representaciones fidedignas de las neuronas biológicas[6].

En este sentido no hay una comprensión completa de cómo es que funcionan las neuronas reales, se trata de una aproximación estrictamente empírica-observacional que pretende retratar como funcionarían neuronas teóricas, delimitadas de un modo manejable para un modelo matemático que se aproxime al tipo de interacción que existiría en los cerebros reales. La aproximación al lenguaje funciona del mismo modo, no es comprender cómo funciona el lenguaje, sino emplear herramientas propias a una estadística de una lengua, ej. ¿cuál es la palabra más repetida en el idioma inglés? ¿qué vocal es la más común? Etc.

Los LLM dividen las palabras en "tokens", que son partículas carentes de significado en sí mismas pero que componen palabras, algunos tokens son sufijos, otro tipo de token puede ser la letra "s" al final de una palabra indicando el plural, etc. Estas partículas no son estrictamente significativas para una gramática, aunque algunas pueden coincidir. La función de los tokens es descomponer textos en partículas analizables en términos estadísticos: ¿qué token es más probable que siga inmediatamente a otro? ¿cuáles tokens suelen emplearse conjuntamente y en qué porcentaje? ChatGPT 3.5 y otros LLM semejantes buscan producir una continuación razonable a cualquier indicación textual que se les proporcione por medio de "prompts". La palabra clave aquí es razonable, que significa grosso modo, que sea coherente o plausible con el contexto proporcionado. Lo que es llamativo de los LLM es la capacidad de producir textos que lucen justo como otros que podríamos encontrar en libros, revistas, internet y demás fuentes. Los LLM pueden decir cosas que suenan bien gracias a que fueron entrenados con cosas

[6] Por ejemplo, Nagyfi (2018) argumenta que la idea general de los predecesores a las neuronas artificiales es imitar ciertas partes de las neuronas como las dendritas, los cuerpos celulares y los axones, a través de modelos simplificados que no imitan ni la creación y destrucción de conexiones entre neuronas además de omitir la duración de las señales.

que suenan bien (Wolfram, 2023). Por ejemplo, Stephen Wolfram reflexiona lo siguiente:

> Es sorprendente cuán humanamente parecen los resultados arrojados... lo que sugiere que el lenguaje humano (y los patrones del pensamiento detrás de este) son de alguna manera más simples y con "forma de leyes" en su estructura de lo que pensábamos. ChatGPT lo ha descubierto de manera implícita. Pero podemos exponerlo de manera explícita con la gramática semántica, lenguajes computacionales, etc. (Wolfram, 2023)

Una crítica común a los LLM es que representan modelos de lenguaje y no de conocimiento. Pero es sencillo observar cómo el modelado exitoso del lenguaje ambiciona ahora el modelado del conocimiento. La presuposición consistiría en que aprender a hablar (dominar el lenguaje natural) es aprender a pensar (el pensamiento se puede asumir incluso qua subjetividad), donde este modelo que equipara el lenguaje natural con la consciencia nos parece conducir a: 1) una reconstrucción de la conciencia donde los tokens (y sus transformadores) constituyen tantas "simulaciones" o "proxys" de estados mentales como sean posibles de constituir, llevando a una conciencia fragmentaria. Las "alucinaciones", que son estos ejemplos donde los LLM se salen del camino predefinido e inventan información, dicen disparates, o empiezan a repetir discursos de odio y de discriminación, son ejemplos de esta falta de unidad[7]. 2) La intención de solucionar esto implementando un filtro con formas inferenciales al análisis de prompts con el afán de caer en línea con los criterios del conocimiento, es decir, cumplir con un criterio de unidad de la razón (conocimiento) visible en los recientemente propuestos modelos grandes de lenguaje aumentados[8].

Este último punto es relevante para el pensamiento de Kant, quien buscó por una parte responder a la crítica a la causalidad de Hume, y por otra, al problema de la unidad de la consciencia. Como resultado de la deducción trascendental, el conjunto de reglas que producen las 3 síntesis (figurativa, de la reproducción, y de la apercepción), y que son justificadas más no deducidas (podría decirse de

[7] Huang, L. et. al. (2023, p. 5-7) dividen las alucinaciones en dos conjuntos: 1) alucinaciones relativas a hechos (inconsistencia y fabricación de hechos), y 2) alucinaciones relativas a la fidelidad hacia las fuentes (inconsistencia relativas a las instrucciones, inconsistencia relativa al contexto, inconsistencia lógica).

[8] Como su nombre lo indican, buscan introducir criterios de razonamiento (mayoritariamente informal), a partir de perfeccionar diversos elementos del modelado. Para una lectura del estado de la cuestión en los LLM referimos a Huang, J., & Chang, K. C. C. (2022) y Mialon, G. et. al. (2023).

modo más contemporáneo pragmáticamente inferidas),[9] obtenemos la unidad del entendimiento; la síntesis de la apercepción muestra que fue una sola conciencia unida la que percibió un mismo objeto primero como aparición (Erscheinung) para determinarlo como fenómeno (Phänomen).

La síntesis de la apercepción, representativa de la facultad del entendimiento, traza las condiciones del sujeto trascendental (subjetividad) pero no es un criterio suficiente para esclarecer la unidad de la consciencia entendida como la razón en general (el conjunto de las facultades). El problema que encuentra Kant, y apunta directo a la metafísica, es que nada impide el empleo de los principios constitutivos del entendimiento fuera del ámbito de la experiencia (esto es, buscar la síntesis entre conceptos puros y no entre intuiciones y conceptos meramente). Kant implica que el mal empleo de los conceptos trascendentales responde a la ambición de encontrar principios absolutamente generales (que apliquen tanto a la razón como al mundo fenoménico), buscando la razón última[10] de las cosas en la forma de principios incondicionados (en A 331/ B 338).

Aunque los principios incondicionados que busca la razón en su facultad especulativa no pueden constituir conocimiento, Kant encuentra que su utilidad yace en la empresa de buscar la unidad de la razón, incluso si los principios se proponen aplicar a las cosas mismas, pues evitarían el problema del eterno regreso de la razón en tanto que no sería posible encontrar un fundamento para su actividad especulativa, regulándola. Los dos tipos de principios (constitutivos y regulativos) son cruciales para capturar la unidad de la razón en general, y como podemos observar, al menos en esta pretensión de los LLM de la igualdad entre hablar y pensar, el papel de los principios regulativos queda subestimado

En general no argumentamos sobre la importancia del papel de los principios regulativos, sino a la problemática que estos conducen en la denominada intuición

[9] Aunque tradicionalmente se apunta a estos conceptos con la diferencia entre quid juris y quid facti en KpR A 87/ B 119, la tradición pragmatista norteamericana ha revisado una y otra vez la herencia Kantiana, por ejemplo, en el pragmatismo clásico de C. S. Peirce, en la noción del a priori analítico de C. I. Lewis, la intención de Sellars de llevar la filosofía analítica a su fase Kantiana, y el más reciente acercamiento a Hegel con autores como R. Brandom y J. McDowell entre otros, quienes parten de una lectura Sellarsiana/ Kantiana del sistema Hegeliano en su mayoría.

[10] Razón aquí refiere a explicación, ya que la idea de causa primera representaría el uso del principio de causalidad que es inmanente a la experiencia (en A 299-301/ B 356-8) como un tipo de principio incondicionado, esto es, de razón de ser de las cosas pero que no necesariamente es la única explicación posible si de establecer principios incondicionados se trata, dado que los principios incondicionados no generan conocimiento cualquier forma de explicación que emplee este proceder cae en el mismo problema.

intelectual. Es decir, que la implementación de principios regulativos, como un siguiente paso para los LLM aumentados, conduciría a la serie de problemáticas que el idealismo alemán se dedicó a solucionar. Por ejemplo, dado que los LLM son modelos de carácter predictivo, la concepción que poseen del lenguaje es causal (esto es por la identificación entre poder causal y poder explicativo a nivel de las teorías científicas[11]). Los modelos estocásticos no cumplen, sin embargo, con un criterio de causalidad fuerte (ontológico), sino de causalidad débil (epistemológico). Aplicado a este modelo que presupone la igualdad entre el lenguaje y la consciencia, se presupone la causalidad ya no a un nivel ontológico (puesto que el modelo estocástico sólo nos provee de causalidad débil), sino a un nivel metafísico, como un compromiso teorético-descriptivo donde lo que no puede explicar la teoría se plantea en términos de una falta de información acerca de los sistemas que pretende modelar (la consciencia en este caso), o más concretamente, una falta de poder computacional[12] de los LLM.

La causalidad como como principio metafísico cuya función es teórico-descriptiva termina siendo una suposición necesaria para un gran modelo de conocimiento, pensando en un sucesor posible de ChatGPT (llamémosle ChatGPT K de aquí en adelante) que tendría que asumir que aquello que modela actúa de acuerdo con leyes, para poder generar sus propias leyes no coincidentes con la naturaleza en sí misma de aquello que modela. Este supuesto no sería del LLM, del que no necesitaríamos que fuese autoconsciente ya que esta suposición estaría a nivel del modelado de la red neural de conocimiento, entrenada para identificar formas inferenciales, y demás criterios explícitos o implícitos de la unidad de la razón. R. König recientemente advirtió esto:

> Se podría decir que la realidad se construye como un área problemática con propiedades muy específicas, ya que los problemas individuales habrán de resolverse mediante el uso de los mismos algoritmos en los que se basan inicialmente, por lo que repito esto, un problema debe ser resuelto por los mismos algoritmos que contribuyeron al surgimiento de los problemas que luego deberían ser resueltos por ellos. (König, 2023)

[11] Dicho de otro modo, de nada sirve postular entidades teóricas que no posean poder causal, ya que no poseerían poder explicativo (en términos de la necesaria predicción a ser realizada por la teoría).

[12] La idea de que hay procesos donde el modelado requiere realizar el trabajo de cómputo paso por paso, sin atajos de por medio, y que implica una cantidad de tiempo inviable por el proceso (computational irreducibility en Wolfram, 2023)

Esta circularidad, califica König, es propia a un punto de vista trascendental. También es sujeto a la crítica que Hegel realiza al idealismo subjetivo en la *Fenomenología del espíritu* (PG), en el apartado de "Certeza y verdad de la razón" (p. 178). Existe un juego de doble ceguera en este respecto, por una parte, se puede sostener que ChatGPT K no estaría asumiendo explícitamente estas suposiciones, pero estarían contenidas necesariamente en la información con la que se entrena, por otra parte se podría señalar que tampoco habría una suposición explícita de este criterio en la información que selecciona para entrenarle, pero la misma idea de seleccionar información relevante y no relevante para que ChatGPT K encuentre la unidad de la razón, conlleva ya esta suposición.

La crítica hegeliana al idealismo subjetivo y al rendimiento causal de un principio explicativo

Hay dos aspectos generales que desde el pensamiento Hegeliano permiten caracterizar la problemática:

1) La idea de causalidad como una suposición metafísica (teórico-descriptiva) implicada en la predicción.

Esta es una problemática que surge no sólo en la predicción, sino que atraviesa temas tan diversos como lo son ¿qué es una explicación científica? Y ¿qué es una explicación filosófica? El reduccionismo y el fisicalismo, y la definición de leyes naturales diferenciándolas de meras generalizaciones y/o disposiciones.
Identificada así, la suposición metafísica causal (teórica-descriptiva) plantea que, si la predicción produce resultados exitosos, entonces se trata de una confirmación de esta suposición que se identifica a su vez con una explicación (científica-causal). El problema es que tenemos numerosos ejemplos donde la predicción es altamente exitosa pero no hay una explicación en el sentido de una descripción narrativa, en el caso concreto de los LLM observamos que:

> ¿Podemos decir "matemáticamente" cómo la red hace sus distinciones? La verdad es que no. Simplemente "hace lo que la red neuronal hace". Pero resulta que eso está bien en tanto que es equiparable con las distinciones que hacemos los humanos.
> [...] Ese no es un hecho que podamos "derivar de primeros principios". Es algo que se ha encontrado empíricamente como cierto, al menos en algunos

> dominios. Pero es una razón clave por la cual las redes neuronales son útiles: de alguna manera capturan una forma "similar a la humana" de hacer las cosas. (en Wolfram, 2023)

Hegel objetaría a esto desde la introducción a la *Fenomenología del espíritu* (PG) y en sus primeras secciones. En la introducción pone de cabeza la noción kantiana de experiencia, esto es, la experiencia de lo sensible, señalando que no se trata de la mera relación entre la consciencia y el objeto (activa y constitutiva de conocimiento), sino del proceso en que la consciencia da cuenta de sí como consciencia y en relación con un objeto (PG p. 78). Dicho de otro modo, lo que la crítica Kantiana muestra es que el nivel de análisis metateórico, el que realiza teoría de la teoría orientada a delimitar las condiciones de constitución del conocimiento, no se encuentra en el nivel en que la consciencia experimenta sensiblemente el objeto, sino en el nivel en que la consciencia da cuenta de esta relación, la abstrae, y se permite universalizarla (en forma de generalidades, disposiciones o leyes, por ejemplo, del buen pensar); el posicionarse desde este nivel de análisis permite la evaluación necesaria para demarcar el conocimiento.
La distinción entre fenómenos y conceptos se revitaliza en esta noción de la experiencia En específico en "Fuerza y entendimiento" (PG p. 107) señalará que el modo en que se vinculan las leyes entre sí no es el modo en que los fenómenos se vinculan entre fenómenos; las leyes se estructuran lógicamente, aunque su contenido sea empírico, de tal suerte que su legalidad (la propiedad misma de ser legales) es la forma inferencial de la metateoría qua filosofía. Esto es, la ley que buscamos es el entendimiento mismo[13].

De lo anterior, podemos apuntar que un principio causal (epistemológico) no necesariamente es explicativo ya que en sí mismo carece aún del elemento metateorético[14], y al menos como una hipótesis temprana (y que esperamos se materialice en futuros proyectos de investigación), queremos plantear que cuando pensamos que una explicación consiste en una descripción causal (que apunta a leyes de modo deductivo), coinciden dos cosas diferentes, por una parte, la suposición metafísica causal (teórico-descriptiva), y por otra, la ontología de la teoría.

---

[13] A esto refiere el pasaje final "Se ha levantado, pues, este telón que había delante de lo interior, y lo que tenemos es el mirar de lo interior hacia dentro de lo interior puro" (en la traducción de Gómez, 2010, p. 243, o en PG, p. 135-6)

[14] Retomando la cita anterior de Wolfram (2023), el capturar una forma similar a la humana de hacer las cosas, no es explicar en qué consiste esa forma de hacerlas sino tratarla como epifenomenal a la predicción.

2) El otro aspecto de la problemática es que la idea de causalidad como una suposición metafísica implicada en la explicación no es posible como un principio a priori.

Para esclarecer este punto no necesitamos enfocarnos en la causalidad misma como sucede al querer delimitarla en el problema de la inducción, más bien planteamos que la función de la causalidad aquí es la de ser un principio explicativo con rendimiento a priori (nótese la sustitución de poder causal por poder explicativo que mencionamos anteriormente).

La idea general es que un principio explicativo presupone aquello que explica[15], ¿de qué otro modo podría explicarlo? Esta pregunta parecería absurda para la epistemología naturalista puesto que describir el poder causal de un fenómeno es explicarlo, pero separando explicación de causación, esto es, separando el nivel de los fenómenos del nivel de los conceptos, o la experiencia sensible de la metateoría como demarcadora de conocimiento, la pregunta pierde su ingenuidad. Localizamos esta crítica en el comentario sobre el idealismo subjetivo en la sección de "Razón" en la *Fenomenología del espíritu* (p. 178-184), donde se plantea que, si la síntesis originaria del Yo es la primera categoría, presupone ya las otras categorías presentes en la consciencia y el mundo[16], en esto consiste el problema de presuponer un principio causal/ predictivo/ explicativo como un principio de la unidad de la razón en un modelo de conocimiento.

Lo que Hegel critica no es en estricto sentido la idea del Yo, sino que la idea del Yo como síntesis originaria cae en esta misma trampa, de asumirse como un fundamento carente de mediación, pero esto no es posible de sostener porque la idea de una primera categoría implica ya el conocimiento y domino de las categorías en general[17].

Transferido a la idea de la causalidad como explicación, evidenciamos el siguiente dilema: El modo en que los LLM predicen el lenguaje posee un estatuto causal epistémico en tanto que nos provee de un tipo de predicción inductiva[18],

[15] Si revertimos el empleo de poder explicativo al poder causal obtendríamos que el efecto provoca la causa, como en el caso de la paradójica retrocausalidad, en su variante circular como la paradoja del abuelo.

[16] A esto se refiere con "…hay diferencias o especies en la categoría." (PG, p. 182)

[17] Hegel va más allá indicando que la cosa a explicar y la explicación son constitutivas una de otra en tanto que sólo vistas en su actualidad (en sí y para sí) pueden constituir conocimiento (ser fundamento de conocimiento). Este es el trasfondo de la crítica a la cosa en sí y el eventual paso de la realidad (Realität) a actualidad (Wirlichkeit) en L I, p. 119.

[18] Donde los aspectos concernientes a las generalidades propuestas por la teoría que

sin embargo, dado el nivel de coherencia aparente de los resultados de los LLM se asume que mediante el añadido de criterios de racionalidad (en el caso de los LLM aumentados) estos modelos darán cuenta exitosamente del pensamiento humano como un tipo de sistema a predecir, pero que al mismo tiempo no plantea una explicación de qué es el pensamiento y mucho menos del lenguaje, sino sólo una explicación de los criterios mismos que ya han sido empleados para entrenar los LLM sobre lo que es el pensamiento y el lenguaje (definición circular).

Es decir, los LLM no son capaces de caracterizar la ontología común que imaginaba Sellars, debido a que sus principios presuponen las mismas condiciones iniciales que deben de predecir. La crítica Hegeliana se dirige a que este tipo de acercamientos omite el papel reflexivo de los principios explicativos al ser demarcativos.

Para Hegel, el conocimiento requiere de dos aspectos esenciales al funcionamiento de una categoría, el determinar un interior (an sich) y un exterior (für sich). Para poder evaluar, todo criterio necesita explicar tanto qué incluye, tanto qué deja fuera (demarcación). Lo subsumido dentro de la categoría y lo excluido de esta, son en su conjunto, constitutivos de conocimiento, lo cual implica que la objetividad de un criterio se muestra en su actualidad (en sí y para sí, an und Für sich) y no sólo en uno de estos aspectos vistos a solas. Esto último impide que la objetividad de los criterios sea puesta a priori o trascendentalmente, ya que es necesario el desarrollo del criterio en conjunto con aquello que norma, de otro modo sólo permanece como una mera abstracción.

Dado que los principios son reflexivos, no pueden ser puestos a priori y no pueden ser ni predictivos o causales. Lo cual muestra cómo la relación entre poder causal y poder explicativo depende de la causalidad como el compromiso metafísico teórico-descriptivo que hemos mencionado, y que es el obstáculo más grande para la comprensión del conocimiento y de la ontología común que mencionaba Sellars, y según Hegel, no podría alcanzarse bajo un criterio a priori de unidad.

En conclusión, podemos señalar que la tesis Sellarsiana representada en los LLM es difícil de sostener debido a la circularidad inherente al mecanismo de alimentación de los modelos. Del mismo modo podemos observar que no se trata meramente de los LLM sino de las suposiciones meta-teóricas de los cuales los LLM participan como el ejemplo más reciente, y que estarían ya depositadas

---

les permitirían vincularse con una teoría basal (es decir las denominadas leyes puente), son tratados como epifienomenales (Kim, 2008, p. 135-6).

desde la tesis naturalista Sellarsiana. La crítica Hegeliana pone en evidencia que el modo de romper dicha circularidad sería a través de la reflexión entendida como el proceso de demarcación, pero que implicaría abandonar la idea de que los principios pudieran poseer un rendimiento a priori o trascendental. Sin embargo, para comprender en su justa medida la aportación Hegeliana a este debate, será necesario abrir un proyecto que vincule la solución al problema del fundamento, de la demarcación del conocimiento, y de la metateoría[19] al sistema Hegeliano. Hemos esbozado algunos puntos de interés esperando despertar interés en esta relación, no obstante, el proyecto queda abierto hacia el futuro.

Referencias

Awodey, Steve & Carus A. C. 2007. "The Turning Point and the Revolution: Philosophy of Mathematics in Logical Empiricism from Tractatus to Logical Sintax". *The Cambridge Companion to Logical Empiricism*. (Alan Richardson & Thomas Uebel, Ed.) Cambridge University Press. USA.

Brandom, R. 2019. *A Spirit of Trust: A Reading of Hegel's Phenomenology*. Harvard University Press. UK.

Cerath, Richard. 2004. "Quine on the Intelligibility and Relevance of Analyticity". *The Cambridge companion to Quine*. (Roger F. Gibson, Ed.) Cambridge University Press. USA

Gomez Ramos, Antonio (traductor). 2010. *Fenomenología del Espíritu*. G. W. F. Hegel. Abada Editores. UAM. Madrid

Hegel, G. W. F. 1986a. *Werke* (Bd. 3, Phänomenologie des Geistes). (Moldenhauer & Michel, Ed.). Surhkamp Taschenbuch Verlag. Frankfurt am Main.

Hegel, G. W. F. 1986b. *Werke* (Bd. 5, Wissenschaft der Logik I, L I). (Moldenhauer & Michel, Ed.) Surhkamp Taschenbuch Verlag. Frankfurt am Main.

Hegel, G. W. F. 1986c. *Werke* (Bd. 6, Wissenschaft der Logik II, L II). (Moldenhauer & Michel, Ed.). Surhkamp Taschenbuch Verlag. Frankfurt am Main.

[19] Distinta de una ciencia de ciencias como una aplicación de los métodos predominantemente cuantitativos de la ciencia a su propio quehacer.

Hegel, G. W. F. 1986d. *Werke* (Bd. 9, Enzyklopädie der philosophischen Wissenschaften im Grundrisse, Zweiter Teil Die Naturphilosophie Mit den mündlichen Zusätzen, EN). (Moldenhauer & Michel, Ed.). Surhkamp Taschenbuch Verlag. Frankfurt am Main.

Huang, J., & Chang, K. C. C. (2022). "Towards reasoning in large language models: A survey". arXiv preprint arXiv:2212.10403.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2023). "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions". arXiv preprint arXiv:2311.05232.

Hylton, Peter. 2004. "Quine on reference and Ontology". *The Cambridge companion to Quine*. (Roger F. Gibson, Ed.) Cambridge University Press. UK.

Kant, I. 2009. *Crítica de la Razón Pura*. (trad. Mario Caimi) FCE. UNAM. Biblioteca Inmanuel Kant. México.

Kim, Jaegwon. 2008. "Making Sense of Emergence". *Emergence: Contemporary Readings in Philosophy and Science*. (Mark A. Bedau & Paul Humphreys, Ed.) MIT Press. USA

König, Robert. 2023, Octubre 3. Trascendental idealism meets AI: challenges and opportunities [Video]. YouTube. https://www.youtube.com/watch?v=2LdXNCzFJRY&t=825s&ab_channel=RobertK%C3%B6nig

McDowell, J. 2009. *Having the World in View: Essays on Kant, Hegel, and Sellars*. Harvard University Press. UK.

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., ... & Scialom, T. 2023. "Augmented language models: a survey". arXiv preprint arXiv:2302.07842.

Nagyfi, Richard. 2018, Septiembre 4. "The differences between artificial and biological neural networks". *Towards data science*. https://towardsdatascience.com/the-differences-between-artificial-and-biological-neural-networks-a8b46db828b7

Quine, W. V. 1951. "Two dogmas of empiricism". *The Philosophical Review*. Vol. 60, No. 1 (Enero). Duke University Press. URL = <http://www.jstor.org/stable/2181906>.

Quine, W. V. 1969. "Epistemology naturalized". *Ontological relativity and other essays*. Harvard University Press. USA.

Rorty, R. 1997. "Introduction". *Empiricism and the philosophy of mind*. W. Sellars. Harvard University Press. UK.

Sellars, W. 1974. *Essays in Philosophy and its History*. D. Reidel Publishing Company. USA.

Sellars, W. 1991. *Science Perception and Reality*. Ridgeview Publishing Company. USA.

Sellars, W. 2007. *In the Space of Reasons: Selected Essays of Wilfrid Sellars*. Harvard University Press. USA.

Wolfram, Stephen. 2023, febrero 14. "What is ChatGPT doing... and why does it work?". Stephen Wolfram Writings. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/