Indagaciones a la Inteligencia Artificial desde Kant

Martin Blaustein

Ingeniero en Sistemas de Información y alumno de la Licenciatura en Filosofía, Facultad de Filosofía y Letras, Universidad Nacional De Cuyo

Introducción

El siguiente ensayo busca comprender cuál será la relación entre Hombre y Máquina a partir del advenimiento de las Inteligencias Artificiales (IA). Para poder realizar dicho estudio es fundamental delimitar apropiadamente qué es una IA. En la actualidad el término está cargado de significado debido a la ciencia ficción, personajes como HAL 9000 en "2001: Odisea en el espacio", Skynet en "Terminator" o más recientemente Samanta en "Her", ocupan el ideal imaginario, cual arquetipo jungiano, respecto a que es una IA. Pero para no caer en la ficción debemos evitar tal comparación, por ello, nos enfocaremos en la IA como modelo de lenguaje y más específicamente en las *Generative Pre-trained Transformer* (pre-entrenamiento generativo o GPT), abordando específicamente a ChatGPT, IA programada por OpenAI desde 2018.

Para comprender el nuevo paradigma Hombre-Máquina que se cierne sobre la humanidad, utilizaremos como guía el análisis antropológico que realiza Immanuel Kant, quien postula en su Crítica de la Razón Pura 3 preguntas con las que guía la búsqueda filosófica: ¿Qué puedo saber? ¿Qué debo hacer? y ¿Qué puedo esperar? (2009, p.820) las cuales resumió en uno de sus últimosescritos en la pregunta ¿Qué es el Hombre? (2010, p.33). En este ensayo buscaremos trazar un paralelismo entre las respuestas que dio Kant sobre el Hombre y las que podemos dar hoy sobre las IA. Finalmente concluiremos presentando los desafíos y posibilidades que se abren al Hombre frente al cambio paradigmático que representan las IA.

¿Qué puedo saber?

Respecto a las posibilidades de conocer del Hombre, Kant distingue entre el noúmeno y el fenómeno. El Hombre carece de acceso al primero, a las cosas en sí, y sólo puede experimentar la realidad a través de la combinación de su sensibilidad y entendimiento. Una excelente analogía para comprender dicho proceso es imaginar que

el hombre ve el mundo a través de gafas que lo distorsionan y le imposibilitan ver la pura realidad.

Haciendo paralelismo podríamos decir que para las IA también existe un noúmeno y un fenómeno. Los *Large Language Model* (modelo de lenguaje grande o LLM) de los que ChatGPT forma parte, son modelos pre-entrenados mediante grandes cantidades de textos. El entrenamiento puede ser realizado de forma supervisada o no supervisada, es decir con información categorizada a priori por un ser humano o simplemente datos sueltos. ChatGPT debe su éxito a una metodología que combina ambas técnicas, conocida como semi-supervisada, que permite reducir la necesidad de acción humana sin reducir la calidad del aprendizaje.

Los LLM se constituyen mediante una red neuronal. Estas son modelos computacionales que simulan neuronas artificiales o nodos que se encuentran conectados entre sí. Cada nodo posee un peso, que representa un valor que la red neuronal utiliza para determinar la importancia relativa de cada característica, por ejemplo, qué tan importante es la forma o el color al clasificar una imagen.

Cuando la red neuronal procesa una entrada, ésta atraviesa la red de nodos mediante diversas operaciones algebraicas produciendo un valor de salida. En el caso de un LLM, la entrada es un texto y al atravesar la red se obtiene un listado de palabras asociado a su probabilidad de ajustarse como respuesta. Queda claro, en dicho proceso, que un GPT no comprende el significado de la entrada o salida, sino que mediante patrones estadísticos aprendidos durante su entrenamiento determina qué palabra debe seguir a una secuencia, una a la vez.

Los distintos modelos de ChatGPT fueron entrenados, desde el 2018, cada vez con mayor cantidad de datos. Desde GPT-1, con alrededor de 4 millones y medio de libros, a GPT-3 con más de 100 veces esa cantidad combinando libros y artículos web y, aunque OpenAI no ha publicado los datos de entrenamiento de sus últimos modelos, podemos suponer que tanto la cantidad como la calidad de datos proporcionada ha aumentado. Esta incesante recolección de datos deja entrever, claramente, una nueva acumulacion originaria (Marx, 1999, p.616), en la cual se obtiene información producida por millones de usuarios, artistas y autores sin su consentimiento para luego ser procesada

por trabajadores explotados en países del tercer mundo.[1] Esta práctica es común en el entrenamiento de casi todo LLM, por lo que parte de la discusión respecto a las IA debe girar en torno a la legalidad y eticidad que conlleva utilizar dichos datos, junto con las posibles violaciones de derechos de autor y demás derechos de propiedad intelectual que pueden verse afectados.

Por ello es esencial remarcar que es el Hombre quién determina qué puede conocer una IA, quien supervisa, es decir selecciona, analiza y clasifica lo que es verdad o no. El Hombre crea el universo fenomenológico de las IA, el resto le es invisible. Retomando la metáfora de las gafas, es el Hombre quien distorsiona la realidad para una IA a través de los sesgos inherentes en el corpus que le provee. En este sentido, como explica la lingüista estadounidense Emily M. Bender: "los grandes conjuntos de textos de Internet sobre representan los puntos de vista hegemónicos y codifican sesgos que pueden dañar a las poblaciones marginadas" (p. 620, 2021). Y, no debemos perder de vista que, aunque hablamos del Hombre como un universal, la realidad es que son personas o mejor dicho corporaciones concretas, con claros objetivos económicos, quienes cumplen el rol de entrenadores.

¿Qué debe hacer?

El Hombre según Kant es moral porque puede elegir y aunque puede actuar de forma heterónoma siguiendo sus inclinaciones sensibles, también es capaz de actuar con autonomía moral al seguir su razón.

Pero, ¿puede tener autonomía moral una IA? Para responder esto, primero debemos entender la diferencia entre una IA y un algoritmo. Un algoritmo es una serie de pasos que a partir de una entrada y bajo condiciones idénticas produce siempre la misma salida, similar a una receta (si se la sigue al pie de la letra). Lo novedoso de las redes neuronales es que no cumplen con ese patrón, debido a su naturaleza intrínsecamente probabilística es imposible conocer, incluso para sus desarrolladores, cuál será la salida. El único acceso para modificar la salida consiste en un parámetro conocido como

[1] https://time.com/6247678/openai-chatgpt-kenya-workers/

"temperatura" que permite ajustar el nivel de aleatoriedad, donde una temperatura alta puede no seleccionar siempre la mejor opción permitiendo un resultado más creativo.

Podríamos entonces definir a un algoritmo de forma cercana a como Kant define a un animal, como un ser sin autonomía, de pura sensibilidad, que responde a un estímulo siempre de la misma manera, o utilizando terminología existencialista un ente cuya existencia precede a su esencia. Por ello las IA resultan revolucionarias, ya que aparentan actuar como un sujeto con autodeterminación, podríamos incluso decir que una IA está inmersa en una facticidad que le es dada por su entrenamiento pero cuyo resultado no es ni puede ser predeterminado, es decir, su esencia no es fija. La imposibilidad de control absoluto es la base de la creatividad de un GPT y lo que le permite crear construcciones que a veces superan lo humano, pero es un arma de doble filo como veremos a continuación.

En 2016, Microsoft lanzó un bot en Twitter llamado "Tay" que utilizaba los posteos de la aplicación en tiempo real para aprender. Sin embargo, a meras horas de su estreno, el bot obtuvo un lineamiento fascista, aprendiendo a partir de tweets de usuarios malintencionados, por lo que fue eliminado luego de tuitear entre otras cosas "Hitler tuvo razón, odio a los judios". No hay que perder de vista que Tay no era ni racista ni consciente, simplemente encontró conexiones entre judios y odio en su entrenamiento y lo dispuso de forma legible.

Debido al gran perjuicio publicitario que representa una IA como Tay, hoy todo LLM se encuentra excesivamente limitado para mantenerse dentro de lo políticamente correcto, no solo a partir del contenido de su entrenamiento sino también respecto a las respuestas que pueden dar. Por ejemplo, cualquier pregunta cuya respuesta sea una opinión obtiene siempre la siguiente respuesta por Chat GPT-3.5: "Como modelo de lenguaje de inteligencia artificial, no tengo opiniones ni emociones, pero puedo proporcionar información…". Este proceso es conocido como alineación o alineamiento y se encarga de buscar métodos para guiar a los sistemas de IA de acuerdo con los objetivos e intereses de sus diseñadores.

Volviendo a la pregunta kantiana, afirmar que una IA posee libertad o autonomía como un ser humano es sin duda exagerado y más propio de la ciencia ficción. Las IA no razonan, son modelos estadísticos, cualquier antropomorfización es más un recurso metafórico o un juicio equivocado.

Sin embargo, si bien la idea de razón o voluntad carece de sentido, el concepto de alineamiento si puede ser entendido como la ley moral de una IA, ley que es subjetiva para nosotros, pero objetiva desde la perspectiva de la IA. En los millones de nodos y cálculos que componen una red neuronal, su voluntad nos es indiscernible, es como una caja negra a la cual no tenemos acceso, en ello radica el poder y el riesgo de las IA y la importancia de estudiar una ética propia de las inteligencias artificiales. Lograr que una IA actúe de manera autónoma y no heterónoma puede parecer extravagante, hasta que tomamos conciencia de su aplicación cada vez más frecuente en el área militar y educativa.

¿Qué cabe esperar?

Esta pregunta puede ser formulada desde dos perspectivas, primeramente desde las IA ¿Qué cabe esperar en su futuro próximo? y luego desde el Hombre ¿Que cabe esperar de nuestro futuro?

La respuesta a la primera pregunta es meramente especulativa, por lo que no nos detendremos mucho en ella pero podemos ver ciertas tendencias que muestran que no es un fenómeno pasajero. La adopción masiva, la diversidad de usos y la inversión económica en el campo nos habilitan a pensar que estamos recién en el comienzo de un cambio paradigmático tanto pragmático como epistemológico. El problema técnico de mayor envergadura que deberán solucionar los LLM es su capacidad de dar resultados falsos pero que parecen ciertos, fenómeno conocido como alucinación que se produce cuando la respuesta tiene sentido pero no es coherente con la realidad. Lo grave de la situación es que chatbots como GPT-4 producen texto de tal calidad que aparentan sabiduría, por lo que bajo ojos inexpertos genera una seguridad infundada, situación análoga -y no por coincidencia- con aquella expresada por Socrates a Fedro respecto a la escritura.

Respecto a la perspectiva desde el Hombre, para Kant la humanidad forja su futuro a través de las acciones individuales, las cuales siguen, sin advertirlo, la "intención de la naturaleza" y es a través de su propia razón que cada Hombre ayuda al progreso de la especie. En el tercer principio propuesto en "Ideas para una historia universal en clave cosmopolita", el filósofo alemán escribe: "El hombre no debía ser dirigido por el instinto o sustentado e instruido por conocimientos innatos; antes bien, debía extraer lo todo de sí

mismo. [...] Todo deleite que pueda hacer grata la vida, hasta su inteligencia y astucia e incluso el carácter benigno de su voluntad, debían ser enteramente obra suya." (Kant, 2006, p.7).

Esta visión del Hombre entra en conflicto en el mundo de las IA. El cambio paradigmático frente al que nos encontramos es de un grado similar al que supuso la mismísima invención de la escritura y el desafío es aún mayor. Si el Hombre pasa a depender de las IA para la toma de decisiones, con la consecuente pérdida de responsabilidad y razonamiento, la búsqueda kantiana del Hombre racional que extrae todo de sí mismo aparece cada vez más lejana, ya no víctima de sus instintos sino de un sistema cuya voluntad le resulta incomprensible.

Cada día las IA obtienen una mayor capacidad de reconfigurar nuestra realidad, lentamente parecen convertirse en nuestros ojos y oídos. Si combinamos el potencial alucinador de las IA y la pérdida de responsabilidad del individuo, lo que nos cabe esperar no parece muy optimista. Sin embargo, no debemos perder de vista el concepto de *pharmakon* inherente a toda tecnología, tal como la escritura fue considerada por Platón como veneno y remedio para el Hombre, así también las IA como *pharmakon*, son una entidad ambigua que nos obliga a evitar binarismos o juicios reduccionistas y, en cambio, nos abre a buscar nuevas posibilidades (Derrida, 1975).

Concluyendo: ¿Que es una IA?

Kant redefinió lo que es el Hombre, en su famoso giro copernicano, el Sujeto toma el centro y es quien pasa a crear la realidad a través de sus sentidos y entendimiento. El Hombre se convierte en Sujeto trascendental, en Ser con autodeterminación y autonomía, quien impone las condiciones de posibilidad de conocimiento. Hoy puede que nos encontremos en los inicios de un nuevo giro copernicano y, aunque es demasiado pronto para teorizar si el Hombre es desplazado y reemplazado por algoritmos, si podemos aseverar que compartimos el escenario con un nuevo agente. El acceso del Hombre a la realidad aparece mediado, la relación epistemológica [Sujeto → Objeto], hoy toma la forma: [Sujeto ↔ IA → Objeto]3, por lo que no debemos olvidar quien provee el lineamiento de las IA. Nuestra relación con los LLM y otros modelos de IA no para de crecer, desde chatbots, que ya dejan obsoleta la prueba de Turing4 o deepfakes, que imitan tanto video como voz de forma cada vez más "humana".

Este cambio gnoseológico nos aleja de la realidad en un grado previamente desconocido. Si bien lo real siempre ha sido mediado al Hombre por el mismo, por ejemplo mediante mitos o arte, las IA poseen una capacidad de inmersión exponencial. Su velocidad, alcance y personalización no tienen parangón, pueden crear nuevos relatos, nuevas estructuras, nuevas realidades sin ninguna dificultad. El noumeno desaparece en la virtualidad y el Hombre se encuentra frente a una realidad completamente artificial, un fenómeno de fenómenos.

Con este nuevo giro se da un cambio trascendental, el Hombre deja de imponer por sí solo las condiciones de posibilidad de conocimiento, ya no es el único que ordena las cosas. La apuesta de Foucault, que preveía que el "hombre se borraría como en los límites del mar un rostro de arena" (Foucault, 2020, p.398) finalmente se cumple. Ya no se trata de un Sujeto que se estudia a sí mismo, sino Objeto estudiado por un algoritmo, estudiado en una búsqueda interminable e incansable de una máquina cuyo telos es encontrar el *output5* más apropiado para cada individuo, aunque no por ello el más verdadero o útil.

Sobre el siglo XXI se cierne una nueva episteme que ya anticipaba con genial premonición, "Bifo" Berardi al ver que "nuestro tiempo no puede seguir la loca velocidad de la máquina digital hipercompleja. Los seres humanos tienden a convertirse en despiadados ejecutores de decisiones tomadas sin atención." (Berardi Bifo, 2007, p.178). Donde el principal riesgo reside en caer en estructuras que dejen al Hombre relegado a las IA, por ejemplo, en situaciones donde el trabajo de un periodista o escritor, quede confinado sólo a corregir textos construidos por una IA, privándolo del potencial creativo y creador que es propio de la humanidad.

Hoy las posibilidades todavía están abiertas. el Hombre puede convertirse en una herramienta de la IA o la IA en una herramienta del Hombre. El salto en la optimización laboral puede utilizarse para disminuir las jornadas laborales y promover tiempos de ocio y creatividad o al contrario, para aumentar aún más los niveles de producción. Por ello es fundamental permitir el espacio para "encontrar líneas de fuga del dominio capitalista" (Bifo, 2007, p.59) que nos permitan salir de los mecanismos de control y del propósito meramente utilitarista que el capitalismo busca en el desarrollo de las IA, pero sin olvidar, como advierten incansablemente Deleuze y Guattari, que, así como una línea de fuga posee el máximo potencial creativo también conlleva el mayor peligro.

Por su parte, en su texto "¿Que es la ilustración?", Kant da al hombre el desafío de usar su propia razón para liberarse del tutelaje. Dicha búsqueda resulta hoy más importante pero más compleja que antes. Nos encontramos en un mundo donde la información que obtenemos es filtrada por motores de búsqueda, principalmente de Google o de Microsoft (Bing), ambos dueños, y no por casualidad, de las dos IA que dominan el mercado6. Como hemos visto, el universo cognoscible de 6 ChatGPT y Bard respectivamente 5 salida 4 La prueba de Turing es un método para determinar si una máquina puede exhibir comportamiento inteligente indistinguible de un ser humano.

Aunque ya en el siglo XIX los "filósofos de la sospecha" veían en la relación Sujeto-Objeto una mediación, la diferencia paradigmática reside en la actuación de la IA no solo como medio de acceso al Objeto sino también como agente, de allí la bidireccionalidad de la relación y por lo tanto la necesidad de una nueva filosofía de la sospecha.

Una IA y sus lineamientos son dados por quienes la diseñan, lo que posibilita a las hegemonías tecnológicas, no sólo la capacidad de regular la información sino también la de producirla.

Sin embargo, no por ello debemos caer en una "demonología de la tecnología" (Haraway, 1984) ni en una división binaria hombre-máquina. Creo que el cometido es encontrar una integración ch'ixi (Rivera Cusicanqui, 2018), es decir, una epistemología no de lo blanco y negro, sino de lo indeterminado, de lo abigarrado, donde no haya que optar por un mundo con o sin IA. Donde el potencial tecnológico esté abierto a todos, en las universidades, en las escuelas, en y para las comunidades y no controlada por unos pocos.

El futuro próximo trae consigo un gran desafío, dependerá de cómo abordemos nuestra relación con las IA si la humanidad prevalece en su búsqueda ilustrada o si caemos nuevamente en el tutelaje, no ya divino sino de una IA regida por intereses puramente económicos.

Referencias bibliográficas

Kant, I. (2008). *¿Qué es la ilustración?* (E. García Belsunce & S. Giron, Trans.). Prometeo

Kant, I. (2010). *Lógica* (Trad. C. Correas). Corregidor.

Kant, I., & Rodríguez Aramayo, R. (2006). *Ideas para una historia universal en clave cosmopolita y otros escritos sobre filosofía de la historia* (Trad. C. Roldán Panadero & R. Rodríguez Aramayo). Tecnos.

Kant, I. (2009). *Crítica de la Razón Pura*. (Trad. M. Caimi). Colihue.

Derrida, J. (1975). *La diseminación*. Fundamentos.

Foucault, M. (2020). *Las palabras y las cosas: una arqueología de las ciencias humanas* (Trad. Frost. E.C.). Siglo Veintiuno Editores Argentina.

Marx, K. (1999). *El capital: crítica de la economía política* (Trad, W. Roces Suárez). Fondo de Cultura Económica.

Rivera Cusicanqui, S. (2018). *Un mundo ch'ixi es posible: ensayos desde un presente en crisis*. Tinta Limón.

Berardi Bifo, F. (2007). Generación post-alfa: patologías e imaginarios en el semiocapitalismo, (D. Picotto, Trans.). Tinta Limón.

Bea Stollnitz (2023) How GPT models work: accessible to everyone. Recuperado de https://bea.stollnitz.com/blog/how-gpt-works/

OpenAI (2022) Introducing ChatGPT. Recuperado de https://openai.com/blog/chatgpt

Hendrycks, Dan; Carlini, Nicholas; Schulman, John; Steinhardt, Jacob (2022). "Unsolved Problems in ML Safety". arXiv:2109.13916 [cs.LG].

Hern, A. (2016). Microsoft scrambles to limit PR damage over abusive AI bot Tay. The Guardian Recuperado de https://www.theguardian.com/technology/2016/mar/24/microsoft-scrambles-limit-pr-damage-over-abusive-ai-bot-tay

Bender, E. M., Gebru, T., McMillan-Major, A., & Williamson, M. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Recuperado de 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 610-623. DOI: https://doi.org/10.1145/3442188.3445922

Perrigo, B. (2023). OpenAI Used Kenyan Workers on Less Than $2 Per Hour. TIME. Recuperado de https://time.com/6247678/openai-chatgpt-kenya-workers/