10.52712/issn.1850-0013-700

Las entrañas de la inteligencia artificial y lo entrañable de su uso *

As entranhas da inteligência artificial e a natureza entranhável de seu uso

The Entrails of Artificial Intelligence and the Engaging Nature of its Use

Fernando Broncano **

Este artículo examina dos cuestiones. La primera es si las modernas inteligencias artificiales generativas y de propósito general pueden ser consideradas tecnologías entrañables; la respuesta es ambivalente, pues depende de cómo se insertarán en las vidas de los usuarios. La segunda cuestión es si la promesa de la singularidad, que abocaría a la imposibilidad de que sean tecnologías entrañables, puede ser realizada o no; la respuesta es negativa por varios argumentos que se exponen y que afectan tanto a lo general de su propósito como a la artificialidad. La conclusión es que deberíamos considerarlas como herramientas en un marco de agencia híbrida.

Palabras clave: inteligencias artificiales; inteligencias generativas; tecnologías entrañables; agentes híbridos

Este artigo examina duas questões. A primeira é se as modernas inteligências artificiais generativas e de propósito geral podem ser consideradas tecnologias entranháveis; a resposta é ambivalente, pois depende de como elas serão incorporadas à vida dos usuários. A segunda pergunta é se a promessa de exclusividade, que levaria à impossibilidade de serem tecnologias entranháveis, pode ou não ser cumprida; a resposta é negativa para vários argumentos apresentados que abordam tanto a finalidade geral quanto a artificialidade. A conclusão é que devemos considerá-las como ferramentas em uma estrutura de agência híbrida.

Palavras-chave: inteligências artificiais; inteligências generativas; tecnologias entranháveis; agentes híbridos

* Recepción del artículo: 05/06/2024. Entrega del dictamen: 14/10/2024. Recepción del artículo final: 15/01/2025.

** Profesor emérito, Departamento de Humanidades: Filosofía, Lenguaje y Literatura, Universidad Carlos III de Madrid, España. Correo electrónico: fernando.broncano@gmail.com. ORCID: https://orcid.org/0000-0002-7705-2947.

10.52712/issn.1850-0013-700

This article examines two questions. The first is whether modern generative and general-purpose artificial intelligences can be considered engaging technologies; the answer is ambivalent, as it depends on how they will be embedded in the lives of users. The second question is whether the promise of singularity, which would lead to the impossibility of their being engaging technologies, can be realized or not: the answer is negative for several arguments that are put forward and affect both the generality of their purpose and their artificiality. The conclusion is that we should consider them as tools in a framework of hybrid agency.

Keywords: *artificial intelligences; generative intelligences; engaging technologies; hybrid agents*

10.52712/issn.1850-0013-700

Introducción

La irrupción de los *large linguistic models* (LLM), como Bard, Gemini, Copilot o el más popular GPT 3, 3.5 y 4, los modelos fundacionales (FM) -junto a otros muchos programas de inteligencia artificial generativa aplicada a imágenes, traducción lingüística, música, juegos y videojuegos, a gestión empresarial y a una multitud de otras aplicaciones- ha producido una conmoción inusitada en muchos ámbitos de la sociedad, incluyendo quizás, sobre todo, el sistema educativo en todos los niveles. Vivimos estos últimos tiempos una fiebre competitiva entre las grandes empresas por ofrecer resultados espectaculares que aumenten el debate social y, desde luego, el valor de las acciones de aquellas. La inteligencia artificial (generativa) se ha convertido ya en el gran tema de controversia social y en uno de los catalizadores de las varias ansiedades tecnológicas y ecológicas que aquejan al mundo contemporáneo. Hay múltiples y profundas cuestiones que abren estas tecnologías, pero en este artículo me centraré en la relación entre la inteligencia humana y la artificial, y en si podemos considerar que esta tecnología cumple los requisitos que Miguel Ángel Quintanilla establece para evaluar tecnologías bajo el término de "tecnologías entrañables" (Quintanilla, 2015; Parselis, 2016).

Una tecnología entrañable (*engaging*, en la traducción inglesa del propio autor del término), según Quintanilla (2015), debería cumplir estos criterios: i) abiertas: accesibles y apropiables; ii) polivalentes: susceptibles de usos alternativos; iii) dóciles: controlables por el usuario; iv) limitadas: consecuencias previsibles, o aplicar el principio de precaución; v) reversibles: tenemos que poder volver hacia atrás; vi) recuperables: mantenimiento activo y de recuperación de residuos; vii) comprensibles: diseño manifiesto, transparente, no opaco; viii) participativas: para facilitar la cooperación humana; ix) sostenibles: que permitan el ahorro, el reciclado de energías y recursos; y x) socialmente responsables: que la implantación de una nueva tecnología no contribuya a empeorar la situación de los colectivos más desfavorecidos.

¿Cumple la inteligencia artificial estos requisitos? Como ocurre con muchas otras técnicas y tecnologías, la evaluación dependerá del alcance y el contenido normativo de lo requerido. Uno a uno, la respuesta dependerá de si el punto de vista es el del promotor entusiasmado o el del crítico pesimista. i) No son abiertas en un sentido, pues ninguna arquitectura de redes neuronales puede serlo por su propio diseño y funcionamiento, pero sí pueden serlo (si se logran imponer las normas que muchas agencias están concibiendo) sus modos de entrenamiento y actualización. ii) Al pretender ser generales y generativas, el uso alternativo está garantizado, pero, al depender de las bases de datos que las han alimentado, lo alternativo del uso es más discutible. iii) ¿Son dóciles? En la medida en que las respuestas (en los LLM) son dependientes de los prompts, podría decirse que tienen cierta docilidad; en tanto que generativas y por ello dotadas de espontaneidad, cabe negar tal docilidad o al menos hay que pensar en qué significa este requisito. De hecho, este será el punto central de lo que sigue en el artículo. iv) Respecto a la previsibilidad de consecuencias, todo indica que es una tecnología intersticial que está reingenierizando el entorno técnico de la civilización, por lo que hay que pensar que, aun si los resultados de las inteligencias artificiales puedan ser previsibles, tomados uno a uno, el contexto general es más bien poco anticipable. v) Todo es reversible y las IA lo son en

cuanto dependen de los enormes almacenes de datos y de los complejos sociales de entrenamiento. Destruyendo estos, la reversibilidad está garantizada, pero en un sentido cognitivo, el saber cómo se produce la inteligencia generativa no creo que sea ya reversible. Es una conquista de la humanidad, como lo es el conocimiento de que la Tierra gira alrededor del Sol. vi) Si son sostenibles es algo sobre lo cual también puede disputarse. Por un lado, sus productos tienen un cierto grado de inmaterialidad, pero, por otro lado, las infraestructuras necesarias para acoger estos modelos son enormes en costos ambientales. (Esta respuesta sirve de paso para responder anticipadamente al criterio ix) de sostenibilidad.) vii) La comprensibilidad depende de la formación de los usuarios. En un cierto sentido, la base general es bastante comprensible pues las redes neuronales y los procesos bayesianos de realimentación solo tienen una dificultad relativa; en otro sentido, las grandes empresas esconden a través de sistemas de secreto las estructuras de transformers y otros artilugios de la arquitectónica de los modelos. viii) Es obvio que son participativas pues los datos que las alimentan provienen de la participación humana, pero que esta participación lo sea de buen grado o su funcionamiento sea, por el contrario, fruto de una inmensa expropiación de datos, es algo que debe ser considerado con cuidado. x) Que las IA contribuyan o no a mejorar la situación de los más desfavorecidos, y en general de toda la humanidad, es la promesa que guía la investigación y crea un horizonte de liberación del trabajo más estúpido, que, por el contrario, sea un nuevo desarrollo del capitalismo que permita la explotación de nuevos aspectos del ser humano es algo que también está en el horizonte de expectativas.

Este somero y poco cuidadoso recorrido no tiene por objetivo evaluar las IA generativas, sino recordar que las evaluaciones son de hecho producto de reflexiones y, sobre todo, de instituciones jurídicas internacionales que expresen en sistemas de control los miedos justificados que producen estas tecnologías al tiempo que permitan sus desarrollos más positivos. Como en otros campos, todo dependerá de las correlaciones de fuerzas entre los intereses más egoístas y los más humanistas y ecológicos. Diversas instituciones, como por ejemplo la Unión Europea,[1] están elaborando marcos regulatorios para el desarrollo de la inteligencia artificial. Mientras ocurren estos cambios, y de modo paralelo, desde la ontología y la epistemología podemos y debemos centrarnos en las características peculiares de estas formas artificiales de inteligencia, y en reflexionar sobre las promesas y amenazas que se ciernen en su progresiva extensión sobre todos los aspectos económicos, políticos y culturales. ¿Es posible lo que los propagandistas llaman la “singularidad”?

Una ambigua controversia

Cuando el futurólogo Ray Kurzweil logró en 2005 una popularidad internacional con su amenaza “La singularidad está cerca” (2045, según su predicción), tras varios textos donde anunciaba el crecimiento exponencial e inminente de la inteligencia artificial, toda la cultura mediática quedó infectada en adelante de esta profecía y, tras cada

1. Más información en: https://digital-strategy.ec.europa.eu/es/policies/regulatory-framework-ai.

anuncio de un nuevo producto más inteligente que el anterior, se ha ido extendiendo un miedo generalizado a que pueda tener razón y que las viejas películas ochenteras describan fielmente el futuro (Kurzweil, 2005, 2016). Para comenzar, hay que reconocer que parte de su éxito está en la atracción de la narrativa de la "singularidad", que evoca transformaciones radicales y puntos de ruptura, algo así como una revolución de las máquinas que alcanzan el nivel de superinteligencia.

El rótulo de la singularidad debe también su éxito a que plantea de hecho interesantes preguntas filosóficas que se sostienen incluso si el horizonte del evento se ve tan lejano como la conversión del sol en un agujero negro.

¿Qué es la inteligencia artificial general? Es una pregunta no sencilla de responder. En principio, habría tres grados:

1. Que supere sin restricciones el test de Turing.
2. Que resuelva problemas que solamente pueden resolver los humanos.
3. Que tenga las mismas funciones cognitivas de los seres humanos.

Los tres criterios acerca de qué pueda ser una inteligencia artificial general son los más populares y extendidos, pero descansan sobre términos o afirmaciones controvertibles. ¿Qué es superar sin restricciones el test de Turing?, ¿qué son problemas?, ¿cuáles son las funciones cognitivas de los humanos? Precisamente por la confianza en lo intuitivo de estos criterios el término "inteligencia artificial general" ha alcanzado tanta popularidad sin demasiadas preguntas sobre su significado.

¿Qué es la singularidad? La definición kurzweiliana señala el evento histórico en el que la inteligencia de las máquinas supera la inteligencia colectiva humana. Que las máquinas resuelven problemas que los humanos no resuelven es algo poco discutible en el mundo contemporáneo. Suponiendo que la inteligencia pueda medirse numéricamente (por ejemplo, el test IQ, que compara la capacidad de resolución de algunos problemas-índice respecto a la edad y a la curva normal poblacional), una posible interpretación sería que una máquina procesadora superase el máximo de la campana de Gauss humana. Una segunda forma de entenderla sería que rompiese los límites definidos por la tarea encomendada y comenzase a entremezclar dominios de razonamiento de acción, algo así como parece que logró la especie humana a través del lenguaje y lo conceptual. En este sentido, podríamos definir el punto de ruptura en tres niveles:

1. Límites de las capacidades cognitivas personales (incluyendo a los miembros más favorecidos intelectualmente de la sociedad).
2. Límites de la inteligencia colectiva, que incluye la deliberación, la crítica y las potencialidades de los recursos compartidos intelectuales de la humanidad.
3. Límites de la inteligencia aumentada y expandida por la integración virtuosa de humanos y máquinas.

La idea de una superinteligencia artificial general en su grado más alto sería aquella que lograse romper los techos de capacidades de solución de problemas en los niveles crecientes desde el (1) al (3). El máximo sería aquel en el que las máquinas no solamente fueran capaces de rediseñar su estructura algorítmica interna por una suerte de meta-aprendizaje, e incluso de rediseñar su *hardware* y producirlo, sino también (ese es el escenario SkyNet de *Terminator*) de producir su propio espacio de problemas y una agenda propia de transformaciones, ajena, e incluso contraria, a los intereses humanos.

Paralela a la línea de profecías desastrológicas, está la de las promesas mesiánicas que anuncian la posibilidad técnica de inmortalidad digital y paraísos pseudoteológicos similares. Ciertamente, no habría que descartar demasiado rápido el componente utópico de la singularidad en algunos aspectos, especialmente en lo que significaría la posibilidad del fin del trabajo asalariado (no del trabajo como arte de la transformación de lo real), en la medida en que las máquinas fueran sustituyendo a las personas en las tareas más tediosas, repetitivas y mal pagadas, creando un escenario poscapitalista. Pero estos escenarios críticos no parece que sean los que dirigen la atracción de numerosos capitales de riesgo hacia las empresas implicadas en la creación de inteligencias artificiales generales.

La mezcla de los dos temas ("¿Qué es la inteligencia general artificial?" y "¿Qué es la singularidad?") hace que la controversia que ha generado este bombo publicitario de la singularidad se haya convertido en un territorio pantanoso, lleno de adjetivos épicos sostenidos por la incertidumbre y pseudolenguajes de posibilidad. Porque los espacios de posibilidad siempre son territorios minados donde la escala tiene mucha importancia. La posibilidad lógica, la física, la realmente técnica, la moral y política, todas ellas intersecan y pueden ser contempladas tanto desde la perspectiva y escala humanas como desde alguna perspectiva cósmica o cósmico-técnica.

Lo general de la inteligencia artificial general

No habrá singularidad mientras las inteligencias artificiales no alcancen un grado de generalidad similar al que tiene la especie humana. El paleontólogo Stephen Mithen (Mithen, 1996) explicó gráficamente en qué podría consistir la generalidad de la mente humana con la metáfora de una catedral gótica en la que la nave central está rodeada de capillas. Mithen escribía sobre el origen de la mente en los grandes simios y observaba cómo homínidos y homininos desarrollaban inteligencias específicas orientadas a ámbitos separados de problemas ecológicos: comunicación y relaciones sociales, clasificación de especies animales y vegetales de interés en la supervivencia, y culturas técnicas orientadas a la fabricación de herramientas. En las especies de simios más cercanos a la especie humana hay algunas formas de solapamiento entre estas inteligencias de dominio específico, pero las interacciones no han conducido a lo que propiamente podría considerarse una inteligencia general. Mithen explicaba el papel causal predominante que el lenguaje tuvo en la intersección e integración de todas las habilidades particulares. El lenguaje rehizo completamente todos los conocimientos específicos al permitir que se conectaran entre sí mediante las redes

conceptuales e inferenciales. El efecto de generalidad fue muy discutido en varios contextos de la filosofía analítica de los años 80 y 90:

> "*It seems to me that there must be a sense in which thoughts are structured. The thought that John is happy has something in common with the thought that Harry is happy, and... something in common with the thought that John is sad... Thus, someone who thinks that John is happy and that Harry is happy exercises on two occasions the conceptual ability which we call 'possessing the concept of happiness*'" (Evans, 1982, p. 100).

Gareth Evans proponía este requisito como característica de la posesión de conceptos, y por ello del pensamiento conceptual, y lo formulaba mediante un conocido requisito:

> "*If a subject can be credited with the thought that a is F, then he must have the conceptual resources for entertaining the thought that a is G, for every property of being G of which he has a conception*" (Evans, 1982, p. 104).

Teniendo en cuenta este criterio de generalidad, y la imagen de Mithen, el proyecto de alcanzar una inteligencia artificial general se ordena en varios grados de logro.

Grado 1. El *criterio de generalidad* de Evans podría considerarse un grado uno de generalidad inducida por el lenguaje en tanto que productor de conceptos, una de cuyas funciones principales es la recognoscitiva. No sabemos si los actuales modelos lingüísticos grandes satisfacen este criterio de generalidad. En apariencia, sí, al menos dependiendo de la interacción conversacional que tenemos con ellos usando *prompts* adecuados, pero tendríamos que probar su habilidad en contextos recognoscitivos más abiertos, en particular en los reconocimientos de imágenes y la posible interacción física con objetos reales en un entorno tecnológico de robots.

Grado 2. La inteligencia general no se limita, sin embargo, a las cadenas conceptuales. La inteligencia humana contiene además la integración de lo conceptual y lo no conceptual: habilidades, emociones. En este nivel es conveniente recordar la controversia que se desarrolló en la década de los 80 como resultado de la propuesta de Jerry Fodor de una arquitectura para la mente humana dividida en módulos especializados e inteligencia general. Aunque el debate se centró fundamentalmente en la naturaleza de los módulos, es muy relevante recordar que Fodor consideraba como rasgo fundamental de la inteligencia general no simplemente la ausencia de dominio específico, sino lo que denominaba "holismo quineano"; es decir, la posibilidad de conectar contenidos muy lejanos en la red de conceptos y creencias. Más allá, incluso, de las fronteras que consideraría Fodor aceptables, la inteligencia general conecta lejanías improbables entre lo conceptual y lo no conceptual, entre habilidades sensoriomotoras y habilidades conceptuales, entre creencias y emociones. La inteligencia general, en este grado, relaciona lo epistémico con los espacios prácticos y los desiderativos y normativos.

Son rasgos centrales en la inteligencia humana, por un lado, la potencia inferencial abductiva, en la que está implicada la hiperconectividad del holismo. Es cierto que un modelo lingüístico generativo como GPT4 contiene una asombrosa conectividad basada en casi dos millardos de parámetros, que le permite producir resultados tan sorprendentes. Pero la conectividad humana no solo es mucho mayor. El cerebro humano, compuesto por aproximadamente 100 millardos de neuronas, cada una de ellas con 7000 sinapsis, alcanza a un orden de casi 800 millardos de parámetros. Pero no se trata de la comparación puramente cuantitativa, que los partidarios de la Ley de Moore podrían afirmar que es técnicamente alcanzable, sino de la estructura de esta conectividad, que no solamente se produce entre las neuronas, en tanto que transmisores y procesadores de información, sino también y sobre todo en la interacción entre la información soportada por la conectividad eléctrica y la transportada por la energía química de las hormonas y neurotransmisores que componen el entorno neuronal y conectan las glándulas y los efectores fisiológicos. La complejidad de la integración humana no es solamente entre neuronas, sino entre células, tejidos y órganos en todo el esplendoroso espectro de la arquitectura de lo vivo. El pensamiento abductivo humano produce inferencias en las que todos estos componentes están implicados y no solamente las propiedades lógicas inscritas en la parte conceptual, sino que, cuando alguien afirma algo así como: "Lo único que tiene sentido, dados estos datos, es…", está apelando a un registro de trasfondo de sentido que incluye toda la complejidad fisiológica.

Supongamos en aras del argumento que alguna máquina llega a este estadio de organización en el horizonte de eventos. Habrá sido un avance asombroso en la historia de la técnica y seguramente pasará versiones más rigurosas del test de Turing que las de los chats actuales, pero no será la singularidad, incluso si procesa todo el caudal del conocimiento humano y resuelve problemas prohibidos a los limitados cerebros humanos. La biología y la historia han producido otros niveles de organización e integración que tienen una dimensión vertical, ortogonal a la del holismo de contenidos.

Grado 3. Me refiero a la integración vertical en la constitución del sujeto en sus dimensiones experiencial, cognitiva y performativa. Para comenzar, es más que dudoso que dispongan de habilidades de metacognición. La metacognición es una función mental, compartida por casi todas las especies de mamíferos y quizá de aves, que consiste en la capacidad para evaluar la dificultad de una tarea en relación con las posibilidades de éxito y, en caso negativo, abandonar el proyecto antes de emprenderlo.

En un segundo nivel, la integración y constitución del sujeto exige la determinación de un orden de valores, entendido como una jerarquía de lo que importa. Más allá de la metacognición, la ordenación de valores regulativos de la existencia es lo que diferencia un sujeto agente de un mecanismo instrumental. No sabemos el grado en que algunos animales logran tener un orden de valores propios (pensamos en un orden natural instintivo, pero muchas especies logran formas de ordenación de valores mucho más sofisticados; por ejemplo, los que establecen sus vínculos emocionales y apegos). La constitución de un sujeto entraña la definición de límites y barreras

que incluyen aquello con lo que se puede vivir y dejan fuera lo que hace de una vida indigna de ser vivida.

Queda aún un último nivel de integración que también pertenece a nuestra naturaleza animal, por más que la cultura humana lo haya sofisticado y llenado de complejidad y matices.

Grado 4. El enactivismo radical se ha alejado de las arquitecturas cognitivistas de la mente, tan inspiradas ellas en los modos de procesamiento combinatorio y secuencial de los ordenadores clásicos. En un cierto aspecto ofrece un modelo de la mente que se asemeja mucho a los sistemas artificiales en lo que respecta al estilo conductista de aprendizaje, también en lo que respecta a la dependencia de instructores y de entrenamiento social. La semejanza, sin embargo, solamente llega hasta aquí. El nivel más complejo de integración agente de los sujetos humanos es el de la capacidad de constituir la experiencia como algo más que como una interacción informacional sobre el medio. La mente humana no se limita a la evidencia, sino que integra los aspectos fenoménicos de la corporeidad con el sentido que adscribe a las intenciones de los otros con los que interactúa. La experiencia, en este sentido, es la constitución de un relato que integra lo vivido en un orden y proyecto de existencia, donde las voces y el discurso de los otros es una parte constitutiva esencial. La integración experiencial y de la alteridad contiene ya una forma muy particular de topología del tiempo implicado en lo narrativo de la experiencia: el conjunto de asimetrías pasado-presente-futuro, sin las que no podría constituirse la memoria y la imaginación presentes en el modo de asimilación de la integración práctica.

En resumen, el proyecto de generalidad de la inteligencia se estructura en grados y niveles que exigen algo más que independencia de dominio y funcionalidad específicos. La inteligencia general exige además integración mente-cuerpo, agente-entorno y sujeto-cultura, que, por lo demás, coinciden con los requisitos de las tecnologías entrañables.

Lo artificial de la inteligencia artificial general

La historia de la tecnología nos muestra muchas cosas. Una de ellas, no menor, es que las trayectorias de las "innovaciones disruptivas" y la transformación real de la historia por la extensión de una cierta innovación tecnológica no son la misma, ni siquiera son paralelas, sino que se entrecruzan y divergen de las formas más extrañas. Un ejemplo clásico es el de la tecnología de la energía de vapor, desarrollada ya en Alejandría, pero cuyo impacto no fue notorio en la humanidad hasta el siglo XIX avanzado. Mucho más reciente, y relacionado con nuestro tema, es la historia de las redes neuronales, ya pensadas en los mismos inicios conceptuales de la informática, pero no tomadas en serio técnicamente hasta finales de los años 90 del siglo pasado.

El problema histórico, cultural, sociológico y filosófico relevante es por qué se producen estas desigualdades en el desarrollo de la tecnología y su integración en la sociedad, y sobre todo por qué no son percibidas por los discursos escatológicos de

la tecnología. Probablemente hay muchas causas y razones sociales que convergen en estas contingencias, pero una de las más significativas es que los inventos, como los descubrimientos científicos, no pasan directamente a formar parte del entorno social. En primer lugar, debe percibirse su funcionalidad; en segundo lugar, debe percibirse la potencialidad de beneficio económico, social o militar que tendría su producción en masa y, en tercer lugar, pero no menos importante, están las derivas y transformaciones, muchas veces creativas, que tienen en el ámbito del uso y de la incorporación a la cultura material.

Ningún invento o innovación es por sí mismo disruptivo o transformador sin una modificación de la cultura material en su entorno que hace posible su extensión social, y, sobre todo, sin una creación de líneas o trayectorias de uso que modifican su funcionalidad. La tecnología es una mediación que transforma identidades y culturas, pero al mismo tiempo es transformada por las pautas que componen esas identidades y culturas. Pensemos, por ejemplo, en la "invención" del motor de combustión interna patentado en 1886 por Carl Benz -después de una larguísima serie de prototipos que nos llevan a centurias anteriores-. Nada estaba escrito en esa patente ni en su diseño que el mundo y las ciudades se llenaran de carreteras de asfalto por la que circulasen millones de automóviles privados. La trayectoria tecnológica que indujo a la generalización del automóvil privado fue tanto un producto de la tecnología como del entorno socioeconómico que creó el nuevo urbanismo y el espacio del transporte privado. Tampoco estaba escrito en el invento del ordenador que iba a ser la informática personal la que reingenierizase el mundo creando el entorno digital. La distancia entre un invento y su impacto social es la que media entre cualquier hecho histórico notable; por ejemplo, las revoluciones americana y francesa y las transformaciones que genera en las sendas ulteriores de la historia.

El determinismo tecnológico suele ser un compañero de viaje tan inevitable como molesto en los discursos sobre la inteligencia artificial general y la singularidad. Puede que "determinismo tecnológico" sea un oxímoron: si es tecnológico no puede ser determinista, pues depende contingentemente de las sendas erráticas por las que se constituye un entorno técnico favorable, y de las mucho más aleatorias de los usos y el entrenamiento humano y las caprichosas de los intereses institucionales que sostienen la tecnología en su difusión (Aibar, 2023).

Lo generativo de la inteligencia artificial generativa

La fuente mayor de asombro que producen los nuevos dispositivos que comenzaron a hacerse públicos en la segunda década del presente siglo es la novedad de sus respuestas, el que los textos que escribían o las imágenes, sartas de códigos o músicas que producían eran sorprendentemente correctos o al menos tenían la apariencia de serlo. Eran una nueva generación de algoritmos de la larga tradición de aprendizaje automático que estaba en el origen de la historia de la inteligencia artificial. Estas nuevas generaciones habían abandonado la vieja programación para adoptar las redes neuronales recurrentes (RNN), las redes generativas adversativas (de términos en competencia, GAN) y, últimamente, lo que ha fundamentado el éxito de OpenAI, la arquitectura transformadora basada en los *transformers*. GPT significa *generative*

pre-trained transformer. Contiene la alianza de un LLM o modelo lingüístico grande con un *transformer*; es decir, un conector de frecuencias de aparición de términos en frases en enormes cantidades de texto con un transformador que produce frases con la probabilidad más alta de que encajen con la respuesta a la pregunta *prompt* hecha por el usuario. El punto de ventaja que ha convertido estos modelos en los héroes del año que acaba mientras escribo estas líneas, 2023, es la técnica que imita la atención humana, la llamada *multi-head attention*, que combina resultados de trabajos estadísticos en paralelo para esta producción de resultados estadísticos.

Necesitan una innumerable secuencia de sesiones de entrenamiento, puesta a prueba y, de hecho, son entrenadas por cada usuario que interactúa con estos *transformers*. Estímulo-respuesta, sofisticada arquitectura de conexiones, enormes cantidades de datos y enormes cantidades de tiempo en paralelo entrenándolos. Esa es la base de la admirable capacidad generativa de las inteligencias artificiales de nueva generación. Tal vez se pudiera aducir, no sin razones, que esta mezcla es lo que hace que las inteligencias artificiales sean tan cercanas al cerebro humano. Ahora que ya no creemos tanto en el innatismo tipo Chomsky, y se ha reivindicado por la corriente enactivista algo similar a un neoconductismo, podría afirmarse que el cerebro de un niño tiene una admirable capacidad generativa por la arquitectura de sus conexiones y el cuidadoso trabajo de sus cuidadores entrenándole.

No. Las analogías entre las máquinas generativas y el cuerpo humano llegan solamente hasta aquí. Aunque la comprensión humana se basase, como en los modelos lingüísticos, en proximidades estadísticas almacenadas en los pesos de las conexiones neuronales, un cerebro humano, y los de muchas especies animales, está dotado de espontaneidad, de una capacidad de autopreguntarse, de reaccionar y activar redes neuronales en la imaginación, el sueño y, sobre todo, la capacidad de elaborar proyectos que entrañan la producción de autoproblematizaciones.

La generatividad automática está conducida y controlada por el entorno de interacciones de los usuarios. La máquina responde muchas veces a ellas afirmando sus propios límites (es lo mejor de los últimos resultados de los entrenamientos, debido a las quejas de tantos usuarios) o (todavía, desgraciadamente) haciendo aparecer respuestas que nacen únicamente de los pesos de conexión, muchas veces equivocadas para desesperación de quienes las reciben, y mucho más de quienes las usan confiando en ellas.

No voy a responder aquí a la cuestión filosóficamente profunda y difícil de qué es lo auténticamente novedoso en la creatividad humana. La filosofía de la ciencia y de la mente se ha cuestionado este tema desde hace décadas. Lakatos, por ejemplo, basó en el criterio de producción de novedades relevantes su filosofía del cambio científico progresivo y no estancado. Hay algo misterioso en la espontaneidad del cerebro, si pensamos en creadores reconocidos; por ejemplo, Monet, un pintor que nunca pintó nada inspirado en narrativas religiosas o míticas, ni siquiera un desnudo, ni se inspiró en la historia del arte, sino que construyó un mundo propio de colores que está en la base de la pintura abstracta. Los cuadros de Monet nacen de un mundo propio autocorregido, en permanente cambio dentro de un mismo proyecto. La creatividad de gente como Monet puede que sea combinatoria, puede que el almacén de recursos

esté ya en el ambiente y haya permeado al cerebro creador, pero la novedad está en el acto de juzgar qué combinatoria es relevante para un modelo de mundo interno generado espontáneamente.

Seguramente la generatividad de las nuevas máquinas es creativa en un cierto sentido. Admirable, sin duda, pero aún demasiado dependiente de las preguntas y correcciones de los usuarios y entrenadores. La creatividad de estas máquinas sigue siendo en gran medida humana.

La IA como tecnología entrañable: la posibilidad de hibridación

Volvamos ahora a nuestra pregunta inicial sobre la "entrañabilidad" de la inteligencia artificial. Que llegue a ser en el futuro una tecnología entrañable o no dependerá en buena medida no solamente de sus aplicaciones, sino también del modo en que se articule la relación de lo humano con las nuevas máquinas. Que supere o no la inteligencia humana es de poco interés, además de poco probable. "Inteligencia" es un término controvertido que refiere a múltiples aspectos y a ejercicios extremadamente diversos. Lo que ha importado para la humanidad no ha sido tanto la inteligencia individual como la colectiva, ejercida en redes de colaboración y acumulada por la memoria oral, práctica y escrita. Hasta el momento, lo que están haciendo las inteligencias artificiales generativas es, primero, apropiarse de la inmensa cantidad de datos y conocimiento que producen los humanos y entrenarse en producir patrones estadísticos de proximidad. No es poco, es un avance impresionante, por cuanto el trabajo de encontrar patrones excede muchas veces las capacidades humanas o exige mucho tiempo de trabajo. Por ello, tiene mucha más funcionalidad e interés la búsqueda de formas aceptables de coordinación entre humanos e inteligencias artificiales. En este camino, aunque son herramientas inteligentes y relativamente autónomas, siguen siendo herramientas que se acoplan con mejores o peores resultados a las necesidades humanas. Es conveniente investigar las causas de por qué se producen los malos acoplamientos. Desde mi punto de vista, habría que considerar dos grupos de factores que explican la persistencia de la poca "entrañabilidad" o, si se quiere, de la persistencia de la alienación humana en relación con la máquina.

En primer lugar, no pueden evaluarse las tecnologías como algo separado de la propia realidad de los humanos y, en este caso, de su inteligencia. La propuesta, respecto a este campo de problemas, debe ser la de centrarse en el carácter de co-construcción y mediación que se produce entre mentes (personales, colectivas) y los artefactos inteligentes. Una nueva conceptualización, en la forma de hibridaciones o composiciones que se evalúen de forma situada y concreta, ayudaría a mejorar las relaciones con las inteligencias artificiales. En segundo lugar, hay otros factores de peor solución que tienen que ver con las trayectorias tecnológicas de producción de inteligencias artificiales no orientadas a resolver problemas humanos, sino a introducir en el mercado nuevos productos que, bajo el título de inteligentes, generan una carrera por el monopolio de las grandes proveedoras de inteligencia artificial y no trayectorias más modestas y diversas, acopladas a los múltiples problemas en que pueden ser usadas de modo positivo las inteligencias artificiales.

La inteligencia humana situada es una inteligencia encastrada, encarnada, enactiva, emocional, extendida técnicamente y social y culturalmente distribuida. Estas características definen la escala y el punto de vista desde el que cabe juzgar y comparar inteligencias. Tomadas una a una, las inteligencias personales habrían sido poco más inteligentes que las de cualquier miembro de otra especie cercana sin la potencia de la constitución de redes de mentes, artefactos, sentidos y valores que ordenan la cultura humana. Lo característico de la inteligencia humana es su increíble capacidad de hibridación en composiciones ontológicamente heterogéneas de formas vivas, artefactos materiales, afectos y memorias compartidas y procesos del entorno materiales, energéticos e informacionales. Por ahora, las máquinas seguirán desarrollándose como mediaciones mediadas por seres híbridos, *cyborgs*, en un mundo mucho más complejo que ellas, como lo es una simple célula o un organismo unicelular.

Las IA generativas avanzadas, construidas con una gigantesca cantidad de conexiones, alimentadas por una descomunal masa de datos, entrenadas por una multitud de auxiliares, son una gran conquista de la humanidad y también, claro, para las empresas que las crean y explotan. No cabe duda sobre su futura utilidad. Serán aportaciones valiosas siempre que haya un entorno adecuado de usuarios que se apropien de sus posibilidades y las empleen como herramientas por las que circulan muchas de otras conquistas anteriores de la cultura humana. Pues son procesadores generativos, pero dependen del reservorio común de producciones humanas depositadas en los almacenes digitales de datos y contenidos.

No cabe duda tampoco de que son instrumentos poderosos y que, por ello, formarán parte de un entorno técnico material que mediará en la formación y el desarrollo de identidades culturales futuras. Sí hay razones para dudar, por el contrario, que su futuro ya esté escrito en alguna ley determinista de explosión computacional. Su vida futura dependerá de trayectorias contingentes que escribirán y describirán las vidas de los usuarios, de otras máquinas y de nuevos diseños.

No cabe duda de que los modelos lingüísticos son impresionantes y causan asombro cuando chateamos con ellos pidiéndoles respuestas a preguntas sofisticadas. Ahora bien, al menos de momento, dependen de los contenidos que hemos producido los humanos y, también por el momento, solo son capaces de moverse en un entorno semiabierto de imágenes y textos digitales bien controlado por los instructores. Han sido creados para fascinar y cumplen bien su función. Está por ver, y eso sí será asombroso, si se producirán IA que tengan capacidades multimodales, multisensoriales, y se muevan en entornos abiertos físicos y sociales, sin mapas y con la tarea de resolver problemas absolutamente nuevos en ilimitados niveles de profundidad. Quizás, desgraciadamente, sean las guerras actuales y futuras los laboratorios donde se pongan a prueba esas nuevas habilidades que, entonces, sí, les harían más cercanos a los animales, también a los humanos.

En un futuro previsible, lo único singular de la singularidad va a ser la capacidad social para domesticar los sistemas complejos de diseño, entrenamiento, difusión comercial o institucional y uso inteligente de un entorno técnico en la idea de constituir

espacios de posibilidad de un mundo más justo y un tiempo liberado progresivamente de la sumisión al trabajo asalariado en un planeta que preserve la vida.

Una parte de este ambiguo futuro dependerá de cómo se resuelva la disputa por la transparencia de las IA generativas. El diseño de sistemas con tal nivel de conectividad seguramente seguirá siendo confidencias. También seguirá siendo opaca la estructura interna de las conexiones pues es una característica de la espontaneidad autoorganizativa de las redes neuronales, algo que comparten con las biológicas. Pero la transparencia exigible legal, ética y políticamente debe alcanzar cuanto antes a los modos de entrenamiento, a la expropiación de datos y contenidos, y a los usos discapacitadores de estos dispositivos.

El control del ciberespacio por parte de los humanos que lo producen y mantienen: eso sí será una verdadera singularidad en la historia. De manera análoga al proyecto de descolonización del espacio físico, el control democrático del digital será uno de los grandes proyectos y retos futuros, mucho más apasionante que el juego de guerra entre humanos y máquinas. Pues no es a las inteligencias artificiales a las que hay que temer, sino a las mucho más peligrosas, irracionales, prejuiciosas y crueles de los poderes oscuros humanos, demasiado humanos.

Las cuestiones seguirán siendo filosóficas, por ello epistemológicas, ontológicas, morales, políticas: cómo conquistar, preservar, cuidar la autonomía humana bajo condiciones de dependencia internas (psicológicas, sociales, culturales) y externas (ecológicas, técnicas). Restan por responder muchas preguntas y desarrollar temas abiertos, pero las líneas esenciales deberían poderse dibujar ya en una agenda filosóficamente informada:

- Nuevas virtudes epistémicas híbridas
- Nuevas racionalidades distribuidas
- Nuevas agencias extendidas
- Viejos valores y lealtades a la vida y a la humanidad preservados

La entrañabilidad de las inteligencias artificiales va a depender de transformaciones sociales profundas que permitan explotar las posibilidades que abren. Por ahora, la situación es ambigua. En muchos casos, lo que se explota no son las posibilidades de libertad, sino las de una mayor sumisión en el trabajo y el consumo, en una loca carrera por el control del mercado de las inteligencias, que solo produce más costos ambientales y más ansiedad por las posibles pérdidas de trabajo y la constante vigilancia.

Bibliografía

Aibar, E. (2023). El culto a la innovación. Barcelona: NED Ediciones.

Bostrom, N. (2022). Superinteligencia. Caminos, peligros, estrategias. Madrid: Teell Editorial.

Callahan, V., Miller, J., Yampolskiy, R. & Armstrong, S. (2017). The Technological Singularity. Managing the Journey. Dordrecht: Springer.

Coeckelbergh, M. (2022). The Political Philosophy of AI. Londres: Polity.

Dee, C. (2023). Large language models (LLMs) vs generative AI: what's the difference?. Algolia. Recuperado de: https://www.algolia.com/blog/ai/large-language-models-llms-vs-generative-ai-whats-the-difference/.

Evans, G. (1982). The Varieties of Reference. Oxford: Clarendon Press.

Kurzweil, R. (2005). La Singularity está cerca: Cuando los humanos trascenderán la biología. Barcelona: Editorial Planeta.

Kurzweil, R. (2016). Hacia una era de inteligencia superhumana. The New York Times, 24 de enero.

Lee, D. (2023). The Singularity: Artificial General Intelligence (AGI) and ChatGPT, A Sky Curation.

Mithen, S. (1996). Arqueología de la mente: Orígenes del arte, de la religión y de la ciencia. Barcelona: Crítica.

Morozov, E. (2023). The True Threat of Artificial Intelligence. The New York Times, 30 de junio. Recuperado de: https://www.nytimes.com/2023/06/30/opinion/artificial-intelligence-danger.html.

Parselis, M. (2016). Tecnologías entrañables como marco para la evaluación tecnológica [Tesis doctoral]. Salamanca: Universidad de Salamanca.

Quintanilla, M.A. (2015). Engaging Technologies: Criteria for an Alternative Model of Technological Development. En B. Laspra & J. A. López-Cerezo (Eds), Spanish Philosophy of Technology. Contemporary Work from the Spanish Speaking Community (103-123). Dordrecht: Springer.

Russell, S. & Norvig, P (2021). Artificial Intelligence. A Modern Approach. Hoboken: Pearson.

Tegmark, M. (2017). Life 3.0. Being Human in the Age of Artificial Intelligence. Nueva York: A. Knopf.

Wang, P., Liu, K. & Dougherthy, Q. (2018). Conceptions of Artificial Intelligence and Singularity. Information (MDPI), 4(9), 79, 1-15. DOI: https://doi.org/10.3390/info9040079.