Palabras clave:
IA general,
IA generativa,
IA evolucionada,
IA diseñada,
IA distopía.

RICHARD BENJAMINS
Chief Responsible AI Officer en Telefónica y *Head of AI for Society and Environments*

Inteligencia evolucionada versus inteligencia diseñada

EL CAMINO HACIA UNA IA GENERAL SEGURA Y SOSTENIBLE

Evolved versus designed intelligence
THE ROAD TO SAFE AND SUSTAINABLE GENERAL AI

Generative artificial intelligence has performed so well that many think that general AI is close. But is it possible to create a system as intelligent as humans? And what would such a system look like, and is there anything we can do to stop it from destroying humanity?

Keywords: general AI, generative AI, evolved AI, designed AI, dystopia AI.

La inteligencia artificial generativa ha conseguido un rendimiento tan bueno que muchos piensan que la IA general está cerca. Pero ¿es posible crear un sistema igual de inteligente que las personas? Y, ¿cómo sería este sistema? ¿Podemos hacer algo para que no destruya la humanidad?

Es difícil escribir algo sobre la inteligencia artificial que no se haya comentado ya. Las muchísimas aplicaciones de negocio, el carácter transformador y transversal de esta tecnología, sus retos éticos y sociales, la huella de carbón de los modelos fundacionales, las iniciativas internacionales de gobernanza de IA, su regulación, el futuro del trabajo por la creciente automatización y el impacto en la sostenibilidad de las cuentas de los países, la creciente concentración de riqueza, el aumento de la desigualdad, el reto de las noticias falsas, la creciente polarización de la sociedades, el uso en el sector militar (killer robots) y un largo etcétera.

Un ejemplo es este mismo número de TELOS con una riqueza de artículos sobre muchos aspectos de la inteligencia artificial. Y si no se ha escrito aún, en breve lo habrá escrito ChatGPT.

Qué distinta era la situación en 1993 cuando defendí mi tesis doctoral sobre cómo programar un sistema inteligente que simulaba el razonamiento diagnóstico de una persona, y era capaz de reflexionar sobre su propio razonamiento. Salvo algunos expertos académicos, pocas personas sabían qué era la inteligencia artificial y podían imaginarse su utilidad en el futuro, es decir hoy, 30 años más tarde: en 2023.

Y qué equivocado estaba pensando que el hype de la IA era en 2016, cuando escribí un libro blanco titulado Surviving the AI Hype. Fundamental concepts to understand Artificial Intelligence[1] que obtuvo más de 10.000 lecturas en pocas semanas. Sigo pensando que es importante entender los conceptos básicos de la IA, pero el hype de la IA sucede ahora, en 2023. En este artículo sobre la IA del futuro, voy a continuar donde acabó Surviving the AI Hype: la inteligencia artificial general (IAG). ¿Es posible? ¿Estamos casi llegando? ¿Cómo sería?, o quizás, ¿cómo debería ser?

Cuando hablamos de la IA general, es importante ser consciente de que entramos en el espacio de las opiniones. Es decir, nadie puede pretender saber con certeza si la IA general es posible o no y cómo sería. Como explicamos Idoia Salazar y yo en nuestro libro El mito del algoritmo:cuentosycuentasdelainteligencia artificial (Anaya, 2020), hay declaraciones sobre la IA que son ciertas (las cuentas), otras son (semi) falsas (los cuentos) y también hay opiniones (las más frecuentes). Además, la aparición de la IA generativa a finales de 2022 ha disparado la cantidad de opiniones acerca de que la IA general está a un paso.

Cuando leemos sobre la IA general (es decir, la posibilidad de crear un sistema inteligente que iguale a la inteligencia humana), nos pintan máquinas que son como nosotros, las personas, con todos los aspectos, buenos y malos. ¿Pero realmente debe ser así?

En este artículo, voy a debatir tres temas:

- ¿Es posible crear una inteligencia artificial general?
- ¿Cómo sería o debería ser este sistema?
- ¿Podemos hacer algo para evitar una distopía?

¿Es posible crear una inteligencia artificial general?

Para debatir la IA general, me gustaría introducir dos conceptos —quizás— nuevos: la inteligencia evolucionada (evolved intelligence) y la inteligencia diseñada (engineered intelligence).

Está claro que nosotros, los humanos, somos el resultado de un proceso de evolución a lo largo de miles de millones de años, en los que, poco a poco, una simple bacteria mediante mutaciones aleatorias se ha convertido en un ser humano, dando lugar también a millones de otros seres vivos. En algún momento de este proceso, lo que ahora llamamos inteligencia se debe haber iniciado. El filósofo Daniel Dennett explica este proceso en detalle en su libro From Bacteria to Bach and Back: The Evolution of Minds[2] (2017), enfocándose en la apariencia de la consciencia. Podríamos ver la evolución como un proceso de aprendizaje por retroalimentación (reinforcement learning) donde las mutaciones aleatorias son los aprendizajes y el éxito o fracaso de la mutación en la supervivencia la recompensa o castiga. Este proceso, después de miles de millones de años, ha dado lugar a las especies de la tierra y a nosotros, los seres humanos, con su inteligencia superior. De ahí el concepto de una inteligencia evolucionada.

Esta reflexión nos invita a preguntarnos si pudiera ser posible crear una inteligencia diseñada, es decir, no aplicar cambios aleatorios, sino diseñar los cambios en una dirección deseada, igual que el aprendizaje automático por retroalimentación. ¿Sería posible acortar el plazo drásticamente usando un método mucho más dirigido? Como señalé, nadie sabe la respuesta a esta pregunta, pero podemos opinar. Mi opinión es que no sé si es posible, pero creo que no es imposible. Otra cosa es cuántos años nos llevará… igual son miles y como humanidad nos extinguiremos antes de conocer el resultado.

Podemos tener la esperanza de que la inteligencia artificial general, si llega, nos llegará sin intenciones de dominarnos

¿Cómo sería o debería ser una inteligencia artificial general?

Imaginémonos que es posible diseñar un sistema de IA general. De hecho, con la IA generativa hemos dado un paso gigantesco, creando sistemas inteligentes capaces de mantener una conversación sobre cualquier asunto, responder a cualquier pregunta, resumir textos, generar textos (poemas, narrativas, notas, etcétera) según las instrucciones de una persona. Estos sistemas se basan en un modelo fundacional entrenado con una ingente cantidad de datos no estructurados y en un proceso de reentrenamiento donde se adapta el modelo fundacional a un dominio o tarea concreta.

Es importante entender, que, aunque el resultado es impresionante, estos sistemas inteligentes no entienden el contenido que manejan. Simplemente, predicen iterativamente

1 Disponible en: https://www.researchgate.net/publication/311862250_Surviving_the_AI_Hype_-_Fundamental_concepts_to_understand_Artificial_Intelligence

2 Más información en: https://en.wikipedia.org/wiki/From_Bacteria_to_Bach_and_Back

Para resolver los graves problemas de nuestra época necesitamos comportamientos lógicos del neocórtex

les de cerebros que, simultáneamente, están haciendo muchas predicciones para guiar nuestra interacción en el mundo físico. Por ejemplo, si tocamos un vaso, sabemos qué sentir y lo percibimos como normal, a no ser que haya una rotura y nos llama la atención —porque no ha coincidido la predicción con la realidad—.

A nivel conceptual, la teoría se puede entender como aprender un mapa de una ciudad a través del contenido de las cuadrículas: una vez aprendido, desde cada cuadrícula se puede predecir el contenido de las adyacentes. Así, cada concepto o conocimiento —concreto o abstracto— se aprende a través de la creación de estos mapas por cada uno de los miles de cerebros operando en paralelo. Cuanto más sabemos de algo, más fácil es aprender conceptos relacionados porque el mapa ya está medio completo. Y, por ejemplo, por eso cuesta tanto aprender matemáticas, porque hay que empezar desde mapas casi vacíos.

Todo este proceso de aprendizaje ocurre en nuestro neocórtex, la parte más reciente del cerebro y asociada a las capacidades (lenguaje, consciencia, reflexión, sentido común, etcétera) que nos diferencian a los seres humanos de los animales. Según Hawkins, este proceso nos da muchas pistas de cómo llegar a una inteligencia artificial general: esta no emerge simplemente de la mera complejidad de un algoritmo o modelo único, sino de miles de procesos de aprendizaje y predicciones en paralelo.

Si nos movemos por una ciudad (o pensamos en un concepto) con nuestros mapas aprendidos, podemos recordar que unos minutos antes estábamos en otro lugar (o pensando en un concepto relacionado), y este mero recuerdo es una forma sencilla de ser consciente, parecido al experimento con un animal, un espejo[4] y una mancha de pintura en la piel del animal. Si el animal presta atención a la mancha, de algún modo, es consciente de sí mismo. Y así, poco a poco, se construye una consciencia cada vez más compleja a partir de millones de interacciones con el mundo físico y mental.

Si llevamos esta teoría de la inteligencia al mundo artificial, debe ser posible crear o diseñar máquinas conscientes de su propia existencia, a través de millones de interacciones y predicciones con el entorno físico y mental (algoritmos).

Evitar una distopía

Además del neocórtex —la parte nueva de nuestro cerebro, con solo un par de millones de años, que nos permite planificar, imaginar, analizar y emitir juicios—, también tenemos la parte antigua, ubicada hacia el centro de nuestro cerebro, que impulsa nuestras respuestas de supervivencia, emociones e instintos. Muchos de los problemas que tenemos hoy en nuestras sociedades son consecuencias de esta parte vieja que, en el fondo, impulsa la supervivencia de nuestros genes a través de un comportamiento egoísta. Nos ha servido muy bien, a través de la evolución de Darwin, para ser el animal superior del planeta. Pero para resolver los graves problemas de nuestra época, necesitamos comportamientos lógicos del neocórtex y no nos sirven los comportamientos de supervivencia, reproducción, emociones e instintos. Por eso, lo que tenemos que perseguir con la IA general es intentar replicar el neocórtex y no la parte vieja de nuestro cerebro.

Así, podemos tener la esperanza de que la inteligencia artificial general, si llega, lo hará sin intenciones de dominarnos, sino como herramientas o instrumentos para ayudarnos a resolver los problemas difíciles que tenemos que afrontar.

Quedan muchas preguntas por resolver, como, por ejemplo, en qué medida los modelos fundacionales representan la parte vieja o la parte nueva del cerebro. Los problemas de la generación de contenido racista, violencia, odio… suenan a la parte vieja; mientras que el pensamiento lógico —con resultados modestos hasta la fecha—, al neocórtex.

O quizás no tiene sentido hacerse estas preguntas y haya que buscar otro camino hacia los “1000 cerebros artificiales”.

Bibliografía

Benjamins, V. Richard (2016): *Surviving the AI Hype – Fundamental concepts to understand Artificial Intelligence.* Working Paper, December 2016. LUCA Data Driven Decisions.

Dennett, D. C. (2017): *De las bacterias a Bach. La evolución de la mente.* Barcelona, Editorial Pasado y Presente.

Hawkins Jeff (2023): *Mil cerebros. Una nueva teoría de la inteligencia.* Barcelona, Tusquets Editores.

3 Disponible en: https://www.todostuslibros.com/libros/mil-cerebros_978-84-1107-249-6

4 Más información en: https://es.wikipedia.org/wiki/Prueba_del_espejo

la palabra subsiguiente de lo que ya hay (de la consulta o prompt, o de la última generada) según la estadística del lenguaje reflejada —aprendida— en el modelo fundacional.

La pregunta crítica en estos momentos es si podemos llegar a la IA general partiendo de la IA generativa aplicando modelos fundacionales cada vez más grandes, o si son necesarios otros varios descubrimientos científicos. Como he comentado, las posibles respuestas a esta pregunta no son más que opiniones, por bien fundamentadas que estén.

En este sentido, considero que el libro de Jeff Hawkins Mil cerebros. Una nueva teoría de la inteligencia[3] es una revelación para un pensamiento original y científicamente fundado, aunque no probado. No se enfoca tanto en qué partes del cerebro se activan según qué tareas, sino que formula la teoría de que las personas tenemos mi-