01/

Inteligencia Artificial Generativa: funcionamiento, posibilidades y riesgos.

Climent Nadeu Camprubí,
profesor Universitat Politècnica de Catalunya (UPC) y catedrático de la Escola d'Enginyeria de Telecomunicació. Barcelona.

Se presenta una breve introducción a la Inteligencia Artificial (IA) centrada en su aspecto de mayor relevancia en la actualidad: la IA generativa. Con una explicación elemental de su funcionamiento, se pretende corregir conceptos erróneos y resaltar las razones de su gran potencial. A partir de esta comprensión básica se muestran las principales limitaciones de las herramientas de IA, los riesgos que conllevan y las cuestiones éticas que plantean, tanto en sí mismas como por la manera de usarlas. Por último, se incluyen algunas consideraciones sobre cómo prepararse para un futuro con esta tecnología.

Palabras clave: Inteligencia Artificial Generativa, Modelos de lenguaje, Alineación, Etica de la tecnología.

A short introduction to Artificial Intelligence (AI) is presented, focusing on its most relevant aspect nowadays: generative AI. By providing a basic explanation of how it works, the aim is to correct misconceptions and highlight the reasons for its strong potential. From this basic understanding, the main limitations of AI tools, the associated risks, and the involved ethical issues, both in themselves and in how they are used, are discussed. Finally, some considerations are included on how to prepare for a future with this technology.

Key words: Generative Artificial Intelligence, Language models, Alignment, Ethics of technology.

La inteligencia y la capacidad técnica son dos rasgos que nos distinguen como especie en nuestro planeta. Si bien la tecnología, es decir, la técnica basada en conocimiento científico, se ha centrado tradicionalmente en el mundo material, las fuerzas, la energía y la información, en las últimas décadas ha comenzado a explorar un nuevo territorio, el de la inteligencia humana.

Este avance es posible gracias a la **inteligencia artificial (IA)**, área tecnológica que desarrolla programas informáticos capaces de realizar tareas cognitivas propias del ser humano, como el manejo del lenguaje escrito y oral, la percepción visual, el aprendizaje, la toma de decisiones, la creación artística o el desarrollo de los propios programas informáticos.

Los programas de IA pueden integrarse en una amplia gama de dispositivos y aplicaciones. Estos programas emplean **algoritmos,** procedimientos matemáticos sofisticados que permiten procesar y analizar vastas cantidades de información. Para ello, se basan en **modelos** de la realidad que actúan como mapas, guiándolos en la obtención de resultados a partir de la información proporcionada por los usuarios.

Los inicios de la IA se pueden situar unos 70 años atrás y su desarrollo a lo largo del tiempo ha sufrido altibajos. Durante mucho tiempo los sistemas de IA más apreciados manipulaban símbolos, usando reglas lógicas para hacer inferencias y deducciones. Actualmente predomina la representación numérica y el **aprendizaje automático** a partir de datos (machine learning). Y dentro de este enfoque destacan las redes neuronales artificiales, llamadas así porque imitan, de forma muy limitada, el funcionamiento del cerebro.

La técnica de las **redes neuronales artificiales** (Artificial Neural Networks, **ANN**), aunque relativamente simple, es tan antigua en sus principios básicos como la propia IA. La reciente eclosión de las tecnologías de IA, especialmente de la IA generativa, que ha captado la atención del público en general con la aparición de herramientas como ChatGPT, se debe más a la convergencia de dos factores clave que a avances científicos disruptivos. Estos factores son el poder de computación cada vez mayor y la explosión -gracias a Internet- de datos digitales disponibles, los cuales proporcionan la materia prima esencial para el entrenamiento de estas redes.

Aquí trataremos de la IA generativa por su potencialidad y actualidad, aunque en el ámbito de la salud se usan también herramientas de IA de otros tipos (ver, por ejemplo, el artículo referenciado de Diego Urgelés).

La IA generativa sirve para crear de forma automática documentos escritos u orales, contenidos audiovisuales (imagen, vídeo, pieza musical), e incluso programas informáticos.

La estrella de la IA generativa es el denominado **modelo de lenguaje de gran tamaño** (Large Language Model, **LLM**), resultado de años de investigación en tratamiento del lenguaje natural (natural en el sentido de humano, en oposición a los lenguajes de programación). Se dice que es un modelo fundamental (**foundation model**) porque sirve de base para el desarrollo de herramientas específicas. Además del LLM existen modelos fundamentales para generar otros tipos de contenidos: imágenes, música, etc.

La generación tiene lugar a partir tanto de la información que el modelo tiene codificada internamente como de la consulta realizada por el usuario, que se denomina prompt. El **prompt,** al igual que el documento generado, puede

El auge de la IA generativa se debe principalmente al aumento en el poder de computación y la disponibilidad masiva de datos digitales

adoptar diferentes modalidades: textual, audiovisual, etc. Cuando se usan conjuntamente varias modalidades los correspondientes modelos multimodales se acercan más a la forma en que los seres humanos interactuamos con el mundo.

1/

¿Cómo funciona la IA? ChatGPT como ejemplo.

Para entrar con cierto detalle en el funcionamiento de las herramientas basadas en IA generativa, nos vamos a centrar en el caso de los chatbots o sistemas conversacionales, que operan con el lenguaje, ya sea escrito u oral. Actualmente existen varios sistemas disponibles, como el pionero y famoso ChatGPT, o sus competidores Gemini, Claude, Llama, etc.

ChatGPT consiste en un LLM, un modelo de lenguaje que engloba varios idiomas a la vez y está formado por estructuras computacionales basadas en redes neuronales, ANNs, las cuales básicamente realizan multiplicaciones y sumas en cada uno de sus nodos.

A veces se supone erróneamente que, para generar su respuesta al prompt de entrada, el sistema LLM realiza búsquedas en bases de datos textuales, ya sean propias o de Internet, y utiliza los resultados para componer las frases. Sin embargo, como veremos, el proceso no consiste en construir frases a partir de resultados de búsqueda, sino en realizar cálculos numéricos usando la representación codificada del lenguaje que posee internamente. La letra T de ChatGPT proviene de la estructura funcional concreta que suelen utilizar actualmente los LLMs: el **Transformador,** la más novedosa aportación de los últimos años en el ámbito del llamado aprendizaje profundo (deep learning).

En la **Figura 1** puede verse un esquema simple de su funcionamiento.

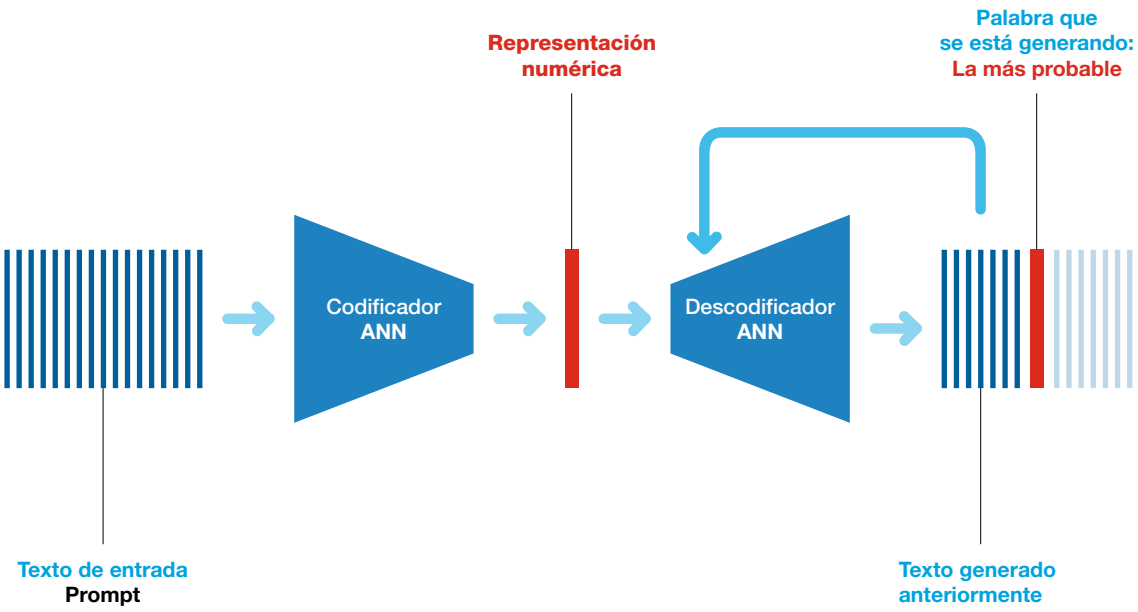

Figura 1. Arquitectura funcional del Transformador, base de LLMs como el de ChatGPT

En la figura, a la izquierda se muestra el prompt o texto de entrada, mientras que a la derecha se observa el texto generado a la salida, que se crea palabra a palabra (la palabra actual está destacada en naranja). El objetivo no es inicialmente generar un texto con sentido o lingüísticamente correcto; esto es un resultado secundario.

En su lugar, el proceso consiste en calcular las probabilidades de todas las palabras posibles y seleccionar la palabra con la probabilidad más alta. Luego, se procede a generar la siguiente palabra, repitiendo el proceso de forma secuencial. En realidad, no se usan palabras del diccionario sino partes de palabras (tokens), porque resultan más eficientes.

Para ilustrarlo con un ejemplo: si el texto de entrada pregunta sobre la información meteorológica, y en la salida ya se ha generado la frase "**hacia media mañana podrá haber...**", la palabra siguiente no puede ser cualquiera, y hay unas pocas palabras que tienen alta probabilidad: nubes, chubascos, ventoleras,... Son estas correlaciones entre palabras y conceptos lo que está contenido de forma codificada en el modelo.

Las probabilidades de todas las posibles palabras (o tokens) se calculan con el Transformador, que consta de dos partes (color verde). El codificador, consistente en una ANN, convierte el texto de entrada en una representación numérica, una secuencia de números, que puede contener miles de elementos. El prompt queda codificado y comprimido en esta secuencia numérica, de manera que han desaparecido las palabras y el texto en sí mismo ya no puede recuperarse.

El descodificador también consiste en una ANN y se encarga de calcular la probabilidad de que cada palabra posible sea la siguiente en el texto que se va generando. Para ello, utiliza no sólo la representación numérica del prompt obtenida por el codificador sinó también el texto generado anteriormente (color azul fuerte); este texto se ha retroalimentado, como se indica en la figura, y numerizado.

Interpretando el funcionamiento del modelo se puede decir que las palabras están representadas como puntos en un espacio multimensional, donde la proximidad entre dos puntos guarda relación con el parecido semántico de las palabras.

De hecho, dado que el lenguaje vehicula el conocimiento y la experiencia de los seres humanos, de alguna forma el modelo, obtenido procesando los textos (o grabaciones orales) existentes, contiene una representación conceptual del mundo, lo que permite generar texto que simule pensamientos, emociones, vivencias, etc.

Esto no significa, sin embargo, que el sistema vincule las palabras con sus referentes en el mundo real ni que les otorgue significado. Por ejemplo, si genera la frase "**la mesa es de madera**", no está asociando como nosotros las palabras "**mesa**" y "**madera**" con algo aprendido por medio de la experiencia, aunque pueda parecerlo.

1/1

Entrenamiento de ChatGPT.

Las redes neuronales se caracterizan por sus nodos y las conexiones entre ellos. Básicamente, en cada nodo se lleva a cabo una suma ponderada de los valores provenientes de otros nodos. Los pesos de la ponderación determinan la importancia relativa de cada conexión entre las neuronas y son parámetros que deben ajustarse durante la etapa de entrenamiento del modelo, la cual se realiza automáticamente utilizando un conjunto de textos.

La letra P de ChatGPT se refiere precisamente a dicho entrenamiento, es decir, a la asignación del valor de los pesos. Significa Pre-entrenado, porque el chatbot se construye partiendo de un LLM de tipo general entrenado previamente, para adaptarlo a la aplicación específica de chatbot, tal como indica la **Figura 2.**

Figura 2. Entrenamiento de ChatGPT

En la Figura 2 puede verse un esquema del proceso de diseño. En primer lugar, se determina un LLM genérico con unos algoritmos de entrenamiento que usan enormes cantidades de textos extraídos básicamente de Internet (páginas web, blogs, redes sociales, artículos científicos, libros, etc.), en varios idiomas. El algoritmo de entrenamiento trabaja de forma autónoma, sin supervisión humana (entrenamiento auto-supervisado); va ocultando palabras o frases de los textos de entrenamiento y trata de predecirlas, una tras otra. Con los errores cometidos en la predicción se van ajustando iterativamente los valores de los pesos a fin de reducir progresivamente el error acumulado sobre el conjunto del corpus de entrenamiento.

Durante esta fase de entrenamiento automático se capta en gran medida la estructura de la lengua (o lenguas, si hay varias en los textos de entrenamiento). El algoritmo consigue extraer patrones, regularidades estadísticas, que corresponden básicamente a relaciones semánticas y morfosintácticas. Dicha información queda codificada numéricamente de forma compacta, lo cual dificulta mucho su observación por parte del diseñador.

Partiendo del modelo pre-entrenado, en un segundo paso se hace el entrenamiento definitivo del LLM chatbot a base de afinar (fine-tuning), de ajustar el modelo genérico partiendo de datos más específicos aportados con intervención humana (por personas que redactan conversaciones, puntúan respuestas que da el sistema,...). Y también usando contenido etiquetado específicamente como no ético, para tratar de evitar que el chatbot genere textos no aceptables.

Nos hemos centrado en la aplicación más conocida, la del chatbot, pero los LLMs pueden estar detrás de otras muchas funcionalidades: traducción automática (que fue la primera aplicación del Transformador), asistentes de voz, respuesta a preguntas, resumen automático, corrección gramatical, generación de software, etc.

2/

Posibilidades y utilidad de la IA generativa.

Al probar ChatGPT, rápidamente se percibe la importancia de redactar buenos prompts. Hay que situar el contexto de lo que se quiere generar, guiar los pasos a realizar, especificar la forma de presentar la información que se solicita, usar expresiones claras y precisas, ...

De hecho, el prompt es una forma de adaptar el procesamiento que se realiza con el LLM a lo que pide el usuario, tanto en forma como en contenido. Equivale en cierto modo a una afinación del modelo, como la que hemos visto que se realiza durante la fase final de entrenamiento, pero en este caso sin necesidad de cambiar el modelo ni de modificar sus parámetros.

Esta capacidad de adaptación al prompt que muestran los LLMs se ha hecho evidente sobre todo cuando se ha pasado a modelos a gran escala (más de 1010 parámetros; por ejemplo, de GPT2 a GPT3). Se llama aprendizaje en contexto (in-context learning) porque, combinando en los prompts descripciones de la tarea con ejemplos de demostración, los LLMs se muestran capaces de obtener buen rendimiento en tareas nuevas, incluso superando en algunos casos los modelos afinados durante el entrenamiento con un conjunto de datos suficientemente grande.

Esta forma de funcionar no había sido prevista y ha sorprendido a los propios investigadores. Y lo mismo ocurre con la posibilidad de dar la respuesta desglosando en pasos la argumentación. Por ello se dice que estos modelos presentan **capacidades** emergentes.

Actualmente ya se dispone de asistentes personales como los chatbots que aceptan prompts de diversas modalidades (texto, voz, imagen, programa informático, etc.) y que generan contenidos también multimodales. También existen sistemas compuestos de IA que combinan distintas componentes (modelos, bases de datos, herramientas externas, etc.) interactuando entre sí para llevar a cabo la tarea deseada.

Además, en poco tiempo aparecerán herramientas todavía más potentes: "**agentes**" de IA que planificaran, gestionaran y dirigiran la ejecución de tareas complejas, los cuales serán pilotados por LLMs y se relacionarán con el usuario a través de ellos.

3/

Limitaciones y riesgos de la IA generativa.

Como todos los productos tecnológicos, las herramientas actuales de IA generativa tienen sus limitaciones e implican ciertos riesgos. En primer lugar, como son modelos basados en redes neuronales, aunque puedan observarse las operaciones matemáticas que realizan, resultan muy opacos en el sentido de que no existen unos criterios y una forma razonada de explicar cómo se ha llegado al resultado a partir de los datos de entrada (el prompt). Esta dificultad, que es especialmente grave en ámbitos como el de la salud o la justicia, es bastante común entre los sistemas basados en aprendizaje automático, pero lo es menos en otros sistemas de IA, como los que se basan en aplicación de reglas.

Es bien conocido que los chatbots actuales producen errores. Preocupan sobre todo las -mallamadas alucinaciones, respuestas que parecen plausibles, pero de hecho son incorrectas.

Recordemos que el objetivo del modelo es maximizar la probabilidad del texto generado, no su adecuación a la realidad. Las versiones más recientes cometen menos errores y este tipo de sistemas irá mejorando continuamente.

Como ocurre con todos los sistemas de IA, las herramientas generativas pueden presentar sesgos que discriminan a individuos o grupos, ya sea por género, etnia, u otros factores. Estos sesgos están estrechamente vinculados al conjunto de datos de entrenamiento utilizado. En el caso de las herramientas generadoras de texto, es importante señalar que no todas las lenguas están representadas en el modelo, y el grado de representación varía considerablemente según el idioma, con el inglés destacando notablemente por encima de los demás.

La IA generativa presenta riesgos como la opacidad, la generación de contenido falso y los sesgos, lo que demanda una regulación urgente

Otra característica que condiciona el desarrollo y uso de la IA generativa es el elevado coste asociado y los problemas de sostenibilidad que esto implica. Se necesitan enormes recursos computacionales, tanto de procesamiento como de almacenamiento, de modo que tan sólo unas pocas empresas e instituciones pueden disponer de ellos. Y, por supuesto, significan un gran coste para el planeta en cuanto a su producción, mantenimiento, etc. y también por su elevado consumo de energía.

Los riesgos mencionados anteriormente están asociados a las limitaciones inherentes de las herramientas de IA generativa. Ahora examinaremos brevemente algunos riesgos específicos y relevantes relacionados con su uso. El primero a destacar es la notable facilidad con la que estas herramientas pueden generar contenido falso o engañoso (fake), lo cual puede ser utilizado para manipular a las personas. No es difícil imaginar un escenario en el que Internet se vea inundada de información falsa, lo que podría resultar en una pérdida generalizada de confianza en los contenidos.

En el ámbito laboral, los principales riesgos son la vulneración de los derechos de autor, especialmente cuando se utilizan imágenes, textos y otros contenidos con propiedad intelectual para entrenar los modelos, y la considerable afectación que la irrupción de estas herramientas probablemente tendrá en el mercado de trabajo.

Además, es difícil detectar plagiarismo porque el texto o imagen generados nunca son iguales a ninguno de los que se han usado para entrenar el modelo. Y, como ocurre con este tipo de herramientas online, la privacidad puede estar comprometida; muchas empresas prohiben a sus trabajadores usar los xatbots de dominio público porque los prompt pueden transportar datos confidenciales.

Dado que conversa fluidamente y escribe con esmero, los usuarios tendemos a interactuar y referirnos al chatbot como si fuera un ser humano (se dice que "**ha aprendido**", "**comprende**","**dialoga**", "**se emociona**", "**tiene valores**"...)porque asociamos capacidad de lenguaje a inteligencia humana. De hecho, como ocurre con muchos artefactos tecnológicos, la IA nos resulta fascinante y al mismo tiempo temible, es decir, tendemos a sacralizarla.

De esta lista no exhaustiva de riesgos asociados a las herramientas de IA generativa, y a las de texto en particular, se desprende la necesidad de regulación, necesidad compartida por muchas tecnologías que incorporan IA. Recientemente, la UE ha aprobado una pionera Ley de Inteligencia Artificial que tiene en cuenta los distintos niveles de riesgo.

4/

Algunas consideraciones sobre nuestro futuro con IA y la forma de enfrentarlo.

Dados los desafíos y oportunidades que presenta la IA, es lógico preguntarse cómo debemos posicionarnos ante estas herramientas tecnológicas. Parece razonable evitar tanto el rechazo como la aceptación acrítica. Es preferible una actitud constructiva, consciente de la naturaleza de estas herramientas y que sea también crítica respecto a cómo las utilizamos.

Puesto que previsiblemente las herramientas de IA se van a integrar cada vez más en nuestra vida cotidiana, es necesario estar bien informados, conocer sus capacidades y limitaciones, para poder usarlas de manera efectiva y ética. Es importante comprender con qué objetivos de optimización funciona la herramienta e interpretar adecuadamente los resultados que proporciona.

En el caso de la IA generativa, es esencial dominar la interacción y el uso de prompts para maximizar su utilidad.

La IA conlleva ciertos problemas y amenazas, pero también ofrece grandes oportunidades. Puede liberarnos de tareas repetitivas o que consumen mucho tiempo, y abrirnos nuevas posibilidades. Un objetivo posiblemente deseable es que las herramientas de IA se encarguen de las tareas habituales que requieren conocimiento, ya sean rutinarias o complejas, permitiendo a las personas concentrarse en procesos más creativos o que precisan del cuidado humano.

Es previsible que la evolución de la tecnología de IA alcance un nivel de desarrollo en el que se considere la conveniencia de otorgar autonomía de decisión a ciertos sistemas de IA en determinadas situaciones.

Ante esta posibilidad, se podría optar por la medida restrictiva de mantener siempre el control en manos humanas. Sin embargo, esta opción podría ser vista como un obstáculo para la eficiencia o resultar inviable debido a las posibles discrepancias entre los usuarios. Además, como se ha señalado, los modelos fundamentales de IA pueden desarrollar capacidades emergentes no previstas por sus diseñadores, y no se puede descartar la posibilidad de que en la fase operativa la IA se comporte de manera distinta a como fue entrenada.

Este conjunto de posibilidades plantea un problema de seguridad en el despliegue de los sistemas de IA. La situación se torna aún más preocupante si consideramos las advertencias de destacados investigadores (como **G. Hinton** y **G. Bengio,** entre otros) sobre el riesgo existencial para la humanidad debido a la IA. Ellos sostienen que el rápido desarrollo de la IA podría alcanzar un nivel de inteligencia tal que los seres humanos queden sin capacidad de control.

Por todo ello, en los ámbitos de desarrollo de la IA crece la preocupación y el interés por trabajar en el problema de la alineación entre el objetivo que persigue realmente el sistema o herramienta IA y el objetivo buscado por el diseñador y el usuario, o más en general, entre los objetivos de la IA y las intenciones y valores humanos. Sin embargo, no es en absoluto inmediato transferir un objetivo al algoritmo de entrenamiento o de operación de los sistemas que triunfan actualmente, especialmente si es de naturaleza cualitativa, debido a la dificultad de expresarlo en términos matemáticos.

Es evidente, por tanto, que la IA representa uno de los grandes desafíos de nuestra época, un desafío que debemos enfrentar desde distintas perspectivas. A continuación, se enumeran tres enfoques complementarios, con la esperanza de que sean compartidos por el lector.

En primer lugar, es necesario desarrollar tecnologías menos opacas y capaces de seguir explícitamente los objetivos que se les asignen. En segundo lugar, como sociedad, debemos establecer mecanismos de regulación y control, incluso antes de que las nuevas herramientas comiencen a aplicarse. Y, en tercer lugar, como personas, debemos cultivar aquellas cualidades que más nos diferencian de la IA y mejor pueden orientar su desarrollo y uso responsable: inteligencia creativa, para adaptarnos a un entorno en constante cambio; sabiduría, para tomar decisiones equilibradas y éticas; y actitud contemplativa, para no someter nuestras funciones mentales al ritmo acelerado propiciado por la IA.

Bibliografía

Christian, B.
The alignment problem. Machine learning and human values,
W.W.Norton&Company, 496 pág., 2020.

Coeckelbergh,
M. Ética de la inteligencia artificia.
Cátedra. 2021.

Libro blanco de la Inteligencia
Artificial Generativa,
Grupo de trabajo IA Generativa,
Transformación Digital, DIGITALES.
Junio 2024.

López de Mántaras Badia,
R. 100 coses que cal saber sobre la intel·ligència artificial,
Cossetània, 319 pág., 2023.

Noble, D. F.
La religión de la tecnología,
Paidós, 298 pág., 1999.

Torres, J.
La inteligencia artificial explicada a los humanos.
Plataforma Actual. 2023

Urgelés Puértolas, D.
"Implicaciones a corto plazo de la inteligencia artificial en la atención sanitaria"
en LABOR HOSPITALARIA n.336-337, pp.36-43, marzo 2023.