OIKONOMICS
Revista de economía, empresa y sociedad

Dosier: «Gestión de la sostenibilidad y la transformación digital» coordinado por Xavier Baraza y August Corrons

IMPACTOS Y PREVISIONES DE FUTURO

Inteligencia artificial generativa y sostenibilidad

Robert Clarisó Viladrosa
Universitat Oberta de Catalunya

RESUMEN En los últimos años, el interés por la inteligencia artificial (IA) ha crecido sustancialmente. La disponibilidad de grandes volúmenes de datos, los adelantos en la capacidad de cálculo y la invención de nuevos métodos y algoritmos han hecho posible entrenar modelos más precisos y fiables que pueden aplicarse con éxito en contextos reales. Este interés creciente se ha disparado todavía más con la aparición de nuevos métodos de IA generativa (IAG), capaces de sintetizar texto, imágenes, vídeo, audio o código a partir de las indicaciones de los usuarios. La popularidad de la IAG, junto con la aparición de nuevas versiones cada vez más sofisticadas que requieren más recursos para su creación y ejecución, ha abierto el debate sobre su escalabilidad a largo plazo. En este artículo analizaremos la IAG desde el punto de vista de la sostenibilidad, considerando el impacto actual y las previsiones de futuro, identificando las tendencias y tecnologías más prometedoras para hacer la IAG más sostenible a largo plazo.

PALABRAS CLAVE inteligencia artificial; aprendizaje computacional; inteligencia artificial generativa; sostenibilidad; huella de carbono

IMPACT AND FUTURE PROJECTIONS

Generative AI and sustainability

ABSTRACT *In recent years, interest in artificial intelligence (AI) has significantly increased. The availability of vast amounts of data, advancements in computing power, and the development of new methods and algorithms have enabled the training of more accurate and reliable models that can be effectively applied in real-world contexts. This growing interest has surged further with the emergence of new generative AI (GenAI) methods, which can synthesize text, images, video, audio, or code based on user input. The popularity of GenAI, along with the introduction of increasingly sophisticated versions that require more resources to create and run, has sparked debate about its long-term scalability. In this article, we will examine GenAI from a sustainability perspective, assessing its current impact and future projections, while identifying the most promising trends and technologies to enhance the long-term sustainability of GenAI.*

KEYWORDS *artificial intelligence; machine learning; generative artificial intelligence; sustainability; carbon footprint*

Introducción

La inteligencia artificial es una rama de las ciencias de la computación que tiene por objetivo entender cómo funciona la inteligencia humana y construir sistemas que puedan emular sus capacidades para resolver problemas (Russell y Norvig, 2020). La IA tiene varias ramas que estudian los diferentes ámbitos de la inteligencia humana: gestionar información (representación del conocimiento), percibir estímulos del entorno (visión por computador), adquirir nuevos conocimientos (aprendizaje computacional), comunicarnos (procesamiento del lenguaje natural), desplazarnos y manipular objetos (robótica), etc.

A pesar de que tiene antecedentes en la filosofía, la lógica y la matemática, la IA como disciplina no surgió hasta la aparición de los ordenadores. Con el ordenador, los humanos disponíamos por primera vez de una herramienta capaz de almacenar y manipular la información necesaria para actuar de forma inteligente. De hecho, el progreso de la IA ha estado muy ligado a las limitaciones técnicas de los ordenadores, como por ejemplo cuánta información se puede almacenar o la velocidad con que se pueden realizar determinados cálculos. Esto ha hecho que ciertas técnicas de IA no hayan sido exitosas hasta que la tecnología informática no ha avanzado lo suficiente para hacerlas viables.

En este artículo, introduciremos fundamentos del aprendizaje computacional, las redes neuronales y el *deep learning*. Esta explicación nos permitirá entender las bases de la IA generativa y los retos que plantea desde el punto de vista de la sostenibilidad.

1. El aprendizaje computacional

Uno de los retos de la IA es cómo proporcionar a los sistemas inteligentes el conocimiento necesario para regular su comportamiento según el contexto y los objetivos a lograr.

Por ejemplo, una de las primeras aproximaciones a la IA fueron los **sistemas expertos**. Un sistema experto emula el conocimiento de expertos humanos sobre una tarea o dominio concreto. Un sistema experto se puede usar como asistente en tareas de clasificación, diagnóstico o predicción. El conocimiento dentro de un sistema experto se puede representar de muchas maneras, por ejemplo, en forma de árboles de decisión o de conjuntos de reglas. Capturar este conocimiento de forma explícita requiere un trabajo intensivo con las personas expertas en forma de entrevistas y análisis de casos, un proceso muy lento y costoso. Además, ciertos tipos de conocimiento pueden ser difíciles de formalizar. Por ejemplo, los humanos podemos diferenciar entre sí los dígitos del 0 al 9, pero es complejo explicar de forma precisa qué elementos nos llevan a decidir que un número es un 8 en vez de un 6 o un 9.

Como alternativa, el aprendizaje computacional (*machine learning*, ML) estudia cómo un sistema puede obtener conocimientos o habilidades a partir de la experiencia y de forma autónoma. Es decir, en vez de recibir unas directrices explícitamente, el sistema analiza los datos recogidos en situaciones previas similares para extraer patrones. Estos patrones permiten ajustar el comportamiento del sistema para conseguir el resultado deseado.

2. Las redes neuronales y el *deep learning*

Uno de los métodos usados en aprendizaje computacional son las **redes neuronales**. Una red neuronal es un modelo matemático que se inspira en la estructura del sistema nervioso: está formada por un conjunto de unidades (**neuronas**) conectadas entre sí, de forma que cada neurona calcula una salida (un valor numérico) y para hacerlo tiene en cuenta una serie de entradas (valores numéricos calculados como salida de neuronas previas). A cada conexión entre neuronas se le asigna un **peso** (también denominado **parámetro**), un valor numérico que representa la importancia de esta conexión en el proceso de cálculo. Así, activar una neurona consiste en hacer un cálculo matemático, donde la salida de la neurona dependerá de las salidas calculadas por las neuronas precedentes y del peso de cada conexión.

Una red neuronal tiene una capa de neuronas de entrada, que reciben los datos del problema codificados en forma de números, y una capa de neuronas de salida, que son las que emiten la respuesta. El resto de capas de neuronas entre la entrada y la salida calcularán resultados intermedios que el usuario no llega a ver, pero que pueden ser pasos relevantes en el cálculo del resultado final.

Para utilizar una red neuronal, primero hay que realizar un proceso de **entrenamiento**: escoger unos valores adecuados para los pesos, de forma que la red proporcione la respuesta más adecuada para cada entrada. El entrenamiento es un proceso de cálculo intensivo. Partiendo de unos valores de pesos aleatorios, en cada paso se proporcionan unos datos de entrada en la red y se observa la salida generada. Según sea la respuesta, adecuada o inadecuada, el algoritmo de entrenamiento ajusta los pesos para favorecer (o evitar) respuestas de este tipo.

Si el conjunto de datos es bastante amplio, este entrenamiento consigue que los pesos de la red capturen los patrones relevantes para definir el comportamiento esperado. Obviamente, el ajuste será mejor cuanto más buenos sean los datos escogidos, por ejemplo, que sean completos, representativos y sin sesgos. Cuantos más datos de entrenamiento tengamos, tanto más mejorará la calidad de la salida. Por lo tanto, el uso de grandes volúmenes de datos (*big data*) ralentiza el entrenamiento, pero garantiza la calidad de la salida.

Hay que destacar que, en una red neuronal, el conocimiento está representado de forma implícita por los valores de cada peso, su relación con el resto de pesos y las conexiones dentro de la red. Es decir, la red neuronal no sigue un árbol de decisión ni un sistema de reglas, solo realiza un cálculo complejo. Esto dificulta considerablemente explicar por qué una red ha generado una determinada salida y no otra. Esto también hace que, para capturar relaciones complejas entre las diferentes dimensiones de un problema, sea bueno utilizar redes neuronales con más neuronas, más conexiones y más capas de neuronas intermedias. Estas son las redes neuronales llamadas **profundas**, en referencia al número de capas, y su diseño y entrenamiento se conoce cómo **aprendizaje profundo** (*deep learning*).

En cuanto una red neuronal ha sido entrenada, se pueden utilizar los pesos de la red para calcular la salida correspondiente a una nueva entrada. Este proceso se denomina **inferencia** y, a medida que la red se va haciendo más grande, puede convertirse en un cálculo muy costoso: almacenar todos los pesos requiere una gran capacidad de memoria y calcular la salida de cada neurona implica hacer un gran número de operaciones matemáticas. Esto puede requerir disponer de *hardware* especializado para hacer inferencia, desde servidores a centros de cálculo dedicados, según las necesidades.

Todo ello ha hecho que el *deep learning* no haya sido posible hasta que el interés por el *big data* ha permitido disponer de grandes volúmenes de datos y la expansión de la computación en la nube (*cloud computing*) ha puesto al alcance de todo el mundo grandes plataformas de computación de altas prestaciones.

3. La inteligencia artificial generativa (IAG)

La inteligencia artificial generativa estudia cómo crear contenidos complejos (texto, código, imágenes, vídeo, ...) teniendo en cuenta unos requisitos definidos como entrada. A pesar de que la investigación en IA generativa tiene una larga trayectoria, este campo ha experimentado un gran crecimiento en la última década gracias a los progresos en el ámbito del *deep learning*. Muchos modelos actuales de IA generativa son redes neuronales profundas que siguen una arquitectura común llamada **transformador** (*transformer*).

Los modelos de IAG a menudo se clasifican según el tipo de contenido que generan. Así, hablamos de **grandes modelos de lenguaje** (*large language models*, LLM) para la generación de texto; modelos de visión para el análisis, transformación o generación de imágenes; o **modelos multimodales** cuando gestionan contenidos de múltiples tipos. El entrenamiento de un modelo de IAG requiere de un gran volumen de información de entrada. Por ejemplo, los LLM se suelen entrenar usando texto escrito por humanos, como por ejemplo libros, noticias, páginas web, comentarios en redes sociales, letras de canciones o código fuente. De manera similar, los modelos de generación de imágenes o vídeo se entrenan con grandes volúmenes de contenidos multimedia.

El *boom* de la IAG ha causado la aparición de nuevas organizaciones centradas en este ámbito, como por ejemplo OpenAI (ChatGPT, Dall-E), Anthropic (Claude) o Mistral AI (Mistral). Por otro lado, las grandes empresas tecnológicas como por ejemplo Google (Gemini), Meta (Llama) o Microsoft (Copilot) han dedicado esfuerzos importantes y están lanzando nuevos modelos, servicios y productos basados en IA generativa.

4. Los retos de una IA generativa sostenible

La competencia por lanzar nuevos modelos más precisos que los existentes ha hecho que cada nuevo modelo sea más grande que el anterior, con redes neuronales más grandes que contienen más parámetros en el contexto de IAG. Como

muestra, la tabla 1 describe la evolución de los modelos de IAG de OpenAI. Para cada modelo, se indica su número de parámetros, el tamaño del conjunto de entrenamiento, la estimación del coste total del proceso de entrenamiento y el tamaño de la ventana de contexto (el número de palabras que se pueden proporcionar como entrada). En esta tendencia se puede observar un crecimiento exponencial: en cada generación el tamaño de los modelos de IA crece 1 o 2 órdenes de magnitud (x10 o x100).

Tabla 1. Evolución de los modelos de lenguaje publicados por OpenAI

Modelo	Fecha de publicación	Parámetros (millones)	Conjunto de entrenamiento (millones de *tokens*)	Coste de entrenamiento (M$, estimación)	Ventana de contexto (*tokens*)
GPT-1	2018	117	0.04	Desconocido	512
GPT-2	2019	1.500	10.000	0,05	1.024
GPT-3	2020	175.000	238.000	4,60	2.048 - 4.096
GPT-4	2023	1.760.000	13.000.000	100,00	32.768 - 128.000

Fuente: elaboración propia

A pesar de que el proceso de entrenamiento de un modelo de IA es mucho más costoso que la inferencia, el entrenamiento solo se tiene que hacer una vez, mientras que el modelo se puede utilizar asiduamente. Estudios de empresas como por ejemplo Amazon o NVIDIA estiman que, a largo plazo, el coste de la inferencia representa el 90 % del coste total de un modelo de IA, mientras que el entrenamiento solo acaba suponiendo un 10 % (Desislavov *et al*., 2023).

Según los datos de la tabla 1, usar GPT-4 para hacer inferencia (es decir, generar texto) requiere cálculos del orden de centenares de teraFLOPS.[1] Otras tareas de IA generativa más complejas, como por ejemplo la generación de imágenes o vídeos tienen un coste superior en uno o más órdenes de magnitud. Esto implica, por ejemplo, que generar una imagen tiene un coste equivalente a cargar un teléfono móvil (Luccioni *et al*., 2024).

Para manejar cálculos de esta magnitud en un tiempo razonable, hay que utilizar un *hardware* específico, como por ejemplo una tarjeta gráfica (GPU) de alta gama en un ordenador personal o consola de última generación, o bien chips específicos diseñados para cálculos de inteligencia artificial. Además del coste de adquisición de estos componentes,[2] hace falta también tener en cuenta su consumo energético y el coste de mantener estos equipos en unas condiciones de temperatura adecuadas.[3] Por ejemplo, un chip de IA como la GPU NVIDIA H100, ofrece 60 teraFLOPS con un consumo eléctrico de 700 W, lo que supondría un coste anual estimado[4] de 10.200 $ solo en concepto de consumo eléctrico. Su fabricante, la empresa NVIDIA, tenía previsto vender 2 millones de estas GPU solo durante el año 2024.

La previsión global sobre la evolución de la IAG es de un incremento sostenido de la demanda, con la aparición de nuevos modelos todavía más potentes y nuevos escenarios de uso. Por ejemplo, la empresa de análisis y visualización de datos Statista prevé que el mercado de la IA generativa, actualmente valorado en 62,7 billones de dólares, crecerá un 41 % anual hasta llegar a los 356 billones de dólares el 2030 (Statista, 2024). Estas previsiones son compartidas por otras consultoras de prospectiva tecnológica, como por ejemplo Gartner, Bloomberg o McKinsey. Por este motivo, tanto gobiernos como empresas están planificando inversiones millonarias en infraestructuras relacionadas con IA, desde la construcción de centros de cálculo de alto rendimiento, redes de comunicaciones de alta capacidad o centrales energéticas dedicadas a cubrir estas necesidades. A escala global, la consultora de prospectiva tecnológica Gartner (Gartner, 2024) prevé que el gasto en tecnologías de la información crecerá un 9,8 %, con un incremento del 23,2 % en el ámbito de los centros de cálculo y un 14,2 % en el ámbito del *software*. Un ejemplo concreto sería la

1. 1 teraFLOPS = 10^12 operaciones matemáticas por segundo usando números de punto flotante.
2. Una GPU de alta gama como por ejemplo la GeForce RTX 5090 tiene un coste superior a los 2.000 €, mientras que un chip dedicado GPU NVIDIA H100 tiene un coste de 25.000 $ en su configuración básica.
3. Se estima que un centro de cálculo puede consumir entre 24,9 y 760 millones de litros de agua cada año.
4. Asumiendo 8.760 horas/año con un coste de 0,1 $/kWh.

iniciativa «Project Stargate», con participación, entre otros, de OpenAI, Oracle y SoftBank, que invertirá 500.000 M$ en infraestructuras de IA en los EE. UU. de aquí a 2029.

Todo esto está disparando el consumo energético de las empresas tecnológicas. Como ejemplo, la Agencia Internacional de la Energía (IEA) prevé que el consumo energético ligado a la inteligencia artificial podría doblarse de aquí al 2026 (IEA, 2024). Este crecimiento está haciendo que muchas empresas estén renunciando a sus objetivos de reducción de la huella ecológica. Por ejemplo, a pesar de tener la neutralidad de carbono como objetivo, Google ha incrementado sus emisiones un 48 % en el periodo 2019-2024. De forma similar, Microsoft ha incrementado sus emisiones un 29 % entre 2020 y 2024 (NPR, 2024).

En resumen, el uso creciente de IAG nos plantea varios retos de sostenibilidad:

- El elevado consumo energético derivado del uso de IA generativa, concentrado en una zona geográfica muy reducida.
- El impacto ecológico de las tecnologías de refrigeración, que pueden representar hasta un 40 % del consumo energético de un centro de cálculo y requerir un consumo de agua significativo (un centro de cálculo puede consumir entre 24,9 y 760 millones de litros al año).
- Los materiales necesarios para construir los chips de IA (por ejemplo, tierras raras), que tienen un impacto significativo en su extracción, procesamiento y reciclaje.

Conclusiones y líneas de futuro

El crecimiento exponencial de los modelos de IAG y la previsión de un incremento continuado de la demanda de nuevos productos y servicios está llevando a gobiernos y empresas a invertir gran cantidad de recursos en infraestructuras tecnológicas. Además de una fuerte inversión inicial, estas infraestructuras tendrán un elevado consumo energético y requerirán de un suministro continuo de chips especializados para el cálculo de IA, que puede ser difícil de mantener o escalar por la escasez de ciertas materias primas como por ejemplo las tierras raras. A medio y largo plazo, la combinación de todos estos factores pone en riesgo la sostenibilidad de todo el paradigma y nos abocará a la necesidad de un cambio de estrategia sobre IA a escala global.

Afortunadamente, algunas tendencias y adelantos técnicos en el campo de la IAG pueden ayudar a reconducir o aliviar los retos de sostenibilidad:

- El impulso de la IAG de código abierto (*open source*) donde las organizaciones hacen públicos los pesos de los modelos que han entrenado, facilitando su reutilización y evitando la necesidad de reentrenar nuevos modelos. El apoyo de instituciones públicas y de empresas, como por ejemplo Meta, puede ayudar a ampliar estas iniciativas.
- El estudio de modelos de lenguaje más pequeños (*small language models*, SLM), menos potentes como herramienta generalista pero muy adaptados a problemas específicos. Un SLM requiere muchos menos recursos para ejecutarse, pudiéndolo hacer en dispositivos con una baja capacidad de cálculo. Algunas técnicas que se están estudiando son la reducción de parámetros innecesarios (*pruning*), de su nivel de precisión (*quantization*) o la transferencia de conocimiento de modelos complejos a modelos más sencillos (*distillation*). El uso de modelos más pequeños reduce considerablemente el consumo energético, evita la concentración del consumo energético en un lugar determinado y la necesidad de construir chips especializados que pueden requerir muchos materiales.
- El descubrimiento de nuevos métodos para entrenar modelos de lenguaje que consiguen una precisión equivalente con un coste inferior. Por ejemplo, la empresa china DeepSeek ha atraído una gran atención al conseguir entrenar un modelo de lenguaje competitivo con modelos de OpenAI utilizando solo una fracción de su coste computacional (5 M$ para DeepSeek en contraste con los 100 M$ estimados para entrenar el modelo GPT-4 de OpenAI).
- La regulación sobre el uso de la IA puede ayudar a evitar ciertos usos o bien asegurar que su implantación tenga en cuenta aspectos de sostenibilidad.

Para acabar, no debemos olvidar que la IAG es una tecnología muy potente que puede ser aplicada a muchos tipos de problemas. En particular, la IAG nos podría ayudar a lidiar con otros retos de sostenibilidad con los que nos enfrentamos (Fraisl *et al*., 2024; Li *et al*., 2024). Al fin y al cabo, la IAG es una herramienta y está en nuestras manos utilizarla de la manera más apropiada.

Referencias bibliográficas

DESISLAVOV, Radosvet; MARTÍNEZ-PLUMED, Fernando; HERNÁNDEZ-ORALLO, José (2023). «Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning». *Sustainable Computing: Informatics and Systems*, vol. 38, 100857. DOI: https://doi.org/10.1016/j.suscom.2023.100857

FRAISL, Dilek; SEE, Linda; FRITZ, Steffen; HAKLAY, Mordechai; McCALLUM, Ian (2024). «Leveraging the collaborative power of AI and citizen science for sustainable development». *Nature Sustainability*, n.º 8, págs.125-132 (2025). DOI: https://doi.org/10.1038/s41893-024-01489-2

GARTNER (2024). *Forecast Analysis: Artificial Intelligence Software, 2023-2027, Worldwide*. Informe [en línea]. Disponible en: https://www.gartner.com/en/documents/4925331

INTERNATIONAL ENERGY AGENCY (2024). *Electricity 2024 – Analysis and forecast to 2026*. Informe [en línea]. Disponible en: https://www.iea.org/reports/electricity-2024

LI, Meng et al. (2024). «Generative AI for Sustainable Design: A Case Study in Design Education Practices». En: Kurosu, M., Hashizume, A. (eds.). *Human-Computer Interaction. HCII 2024. Lecture Notes in Computer Science*, vol. 14687. Cham: Springer. DOI: https://doi.org/10.1007/978-3-031-60441-6_5

LUCCIONI, Sasha; JERNITE, Yacine; STRUBELL, Emma (2024). «Power Hungry Processing: Watts Driving the Cost of AI Deployment?». *Proceedings of the 2024 ACM Conference donde Fairness, Accountability, and Transparency* (FAccT'24), págs. 85-99. Nueva York: Association for Computing Machinery. DOI: https://doi.org/10.1145/3630106.3658542

NPR (2024). «AI brings soaring emissions for Google and Microsoft, a major contributor to climate change». *npr* [en línea]. Disponible en: https://www.npr.org/2024/07/12/g-s1-9545/ai-brings-soaring-emissions-for-google-and-microsoft-a-major-contributor-to-climate-change. [Fecha de consulta: 15-02-2025].

RUSSELL, Stuart; NORVIG, Peter (2020). *Artificial intelligence: a modern approach, 4th edition*. Hoboken: Pearson.

STATISTA RESEARCH DEPARTMENT (2024). «Generative AI - Worldwide». *Statista* [en línea]. Disponible en: https://statista.com/outlook/tmo/artificial-intelligence/generative-ai/worldwide#market-size. [Fecha de consulta: 15-02-2025].

Cita recomendada: CLARISÓ VILADROSA, Robert. «Inteligencia artificial generativa y sostenibilidad». *Oikonomics* [en línea]. Mayo 2025, n.º 24. ISSN 2330-9546. DOI: https://doi.org/10.7238/o.n24.2501

Robert Clarisó Viladrosa
rclariso@uoc.edu
Universitat Oberta de Catalunya

Ingeniero informático (2000) y doctor en Lenguajes y Sistemas Informáticos (2005) por la Universitat Politècnica de Catalunya. Profesor agregado en los Estudios de Informática, Multimedia y Telecomunicación de la Universitat Oberta de Catalunya (2005-actualidad) y profesor asociado en la Universitat Politècnica de Catalunya (2006) y la Universitat Autònoma de Barcelona (2006-2011). Actualmente es investigador principal del grupo de investigación Systems, Software and Models (SOM Research Lab) en el Internet Interdisciplinary Institute (IN3-UOC) y coordinador del grupo de trabajo institucional sobre «IA en educación» en la UOC.

ODS