ECOS DE SOSTENIBILIDAD: REFLEXIONES SOBRE LA INTELIGENCIA ARTIFICIAL A LA LUZ DE LOS OBJETIVOS DE DESARROLLO SOSTENIBLE

SUSTAINABILITY ECHOES: REFLECTIONS ON ARTIFICIAL INTELLIGENCE IN LIGHT OF THE SUSTAINABLE DEVELOPMENT GOALS

José L. González-Geraldo[1] y *José A. Ballesteros*[2]

Fechas de recepción y aceptación: 23 de mayo de 2024 y 11 de junio de 2024

DOI: https://doi.org/10.46583/edetania_2024.65.1142

Resumen: Vivimos tiempos difíciles, pero también llenos de retos sin precedentes a los que las universidades han de responder de una manera tan eficiente como sostenible. Uno de ellos deriva del progresivo desarrollo y diseminación de la Inteligencia Artificial (IA) y, más específicamente, de los modelos generativos de lenguaje (LLM). Este artículo expone y profundiza en el impacto que estas nuevas tecnologías pueden y deben tener en relación con los Objetivos de Desarrollo Sostenible (ODS), en especial el cuarto, y más concretamente dentro de la educación superior. Más allá de los aspectos positivos que sin duda acarrea, reflexionaremos sobre cómo la adopción descontrolada y desmedida de la IA puede socavar tanto el pensamiento crítico como la calidad educativa. No debemos olvidar que, históricamente, las universidades han sido los baluartes del librepensamiento, redundantemente crítico, algo completamente alejado de una aceptación acrítica tan directa y fácil como la que solemos tener ante la generación de ideas que nos proporcionan los LLM. La facilidad de acceso y uso de estos modelos promueve una menor resistencia al esfuerzo, intensificando de manera silente la polarización ideológica o, como mínimo, su simplificación excesiva en sesgadas cámaras de eco disfrazadas de inteligencia donde predominan opiniones que refuerzan creencias preexistentes, la desinformación o directamente las noticias falsas. Estos sesgos, por supuesto,

[1] Universidad de Castilla-La Mancha (UCLM). Facultad de Ciencias de la Educación y Humanidades. Correspondencia: joseluis.ggeraldo@uclm.es

[2] Universidad de Castilla-La Mancha (UCLM). Escuela Politécnica de Cuenca. Correspondencia: josea.ballesteros@uclm.es

afectan asimismo a otros ODS como, por ejemplo, la igualdad de género. Por ello abogamos por un enfoque equilibrado, tan ineludible como sabio, donde la tecnología sea usada de forma responsable para enriquecer el pensamiento crítico y nuestra esencia más humana, y no para que la humanidad se someta a ella en lo que podríamos denominar un esclavismo tecnológico, entendiéndola no como un fin en sí misma sino como una poderosa aliada. Así, promoviendo una educación que favorezca un análisis crítico genuino y diverso, la educación superior, para seguir siendo precisamente superior puede y debe tomar cartas en el asunto sin dilación, sin temor, ni temblor.

Palabras clave: Inteligencia Artificial; Objetivos de Desarrollo Sostenible; ODS; Calidad de la educación; Teoría de la educación.

Abstract: We live in challenging times, but also in times filled with unprecedented challenges to which universities must respond in a manner that is both efficient and sustainable. One such challenge stems from the progressive development and dissemination of Artificial Intelligence (AI), and more specifically, from Large Language Models (LLMs). This article explores and delves into the impact these new technologies can and should have concerning the Sustainable Development Goals (SDGs), particularly the fourth goal, and more specifically within higher education. Beyond the undoubtedly positive aspects, we will reflect on how uncontrolled and excessive adoption of AI can undermine both critical thinking and educational quality. We must not forget that historically, universities have been bastions of free thought, redundantly critical, something entirely distant from an uncritical acceptance as direct and easy as the one we often have towards the generation of ideas provided by LLMs. The ease of access to and use of these models promotes less resistance to effort, silently intensifying ideological polarization or, at the very least, its excessive simplification in biased echo chambers disguised as intelligence, where opinions that reinforce pre-existing beliefs, misinformation, or outright fake news predominate. These biases, of course, also affect other SDGs, such as gender equality. Therefore, we advocate for a balanced approach, as inescapable as it is wise, where technology is used responsibly to enrich critical thinking and our most human essence, and not for humanity to be subjected to it in what we might call technological enslavement, understanding it not as an end in itself but as a powerful ally. Thus, by promoting an education that fosters genuine and diverse critical analysis, higher education, to remain precisely superior, can and must take action without delay, without fear, and without trembling.

Keywords: Artificial Intelligence; Sustainable Development Goals; SDG; Quality of Education; Educational Theory.

1. Introducción

Al igual que no hace falta ser físico o matemático para poder debatir sobre el impacto que la energía nuclear puede tener en nuestras vidas, o bioquímico para discutir sobre la necesidad de revertir el cambio climático, tampoco sería

adecuado dejar únicamente en manos de los ingenieros de telecomunicaciones o los informáticos las decisiones que incumben al desarrollo, evolución e impacto de la Inteligencia Artificial (IA) en nuestras vidas y en nuestro futuro, teniendo en cuenta que, para poder emitir un juicio crítico, este no debe versar solamente sobre creencias u opiniones personales, sino sobre datos objetivos que apoyen las tesis que se argumentan.

Hoy, abocados a una vertiginosa confrontación propia de la "Guerra Fr-ÍA" en la que nos vemos inmersos, y que dirimirá quién conseguirá -o no- destronar el monopolio que durante tanto tiempo copó Google en lo que a la búsqueda y gestión de información en Internet se refiere, hemos de prestar atención cómo esta nada nueva pero sí expandida y potenciada tecnología de la IA impacta en los planes y horizontes previstos para toda la humanidad.

No es casual que la recientemente aprobada Ley de Inteligencia Artificial de la UE (2023), prestando especial atención a los riesgos que estas tecnologías pueden ocasionar, y que algunos autores ya los pusieron de manifiesto en su relación con los ODS (Vinuesa, 2020), señale expresamente como puntos a tener en cuenta la sostenibilidad medioambiental y la eficiencia energética, así como otros aspectos sociales que necesariamente afectan a las medidas de inclusividad y diversidad que hacen que las personas vulnerables, aunque no de manera exclusiva, puedan estar especialmente afectadas por un mal uso o un uso desmedido de las distintas IA. Todo ello, por supuesto, sin demonizar estos avances, pues este mismo texto recoge y admite que: "La IA es un conjunto de tecnologías en rápida evolución que contribuye a generar beneficios económicos, medioambientales y sociales muy diversos en todos los sectores económicos y las actividades sociales" (p. 5).

En este sentido, esta ley categoriza los sistemas de IA según tres niveles de riesgo: bajo, alto y muy alto. Los sistemas de bajo riesgo incluyen aplicaciones mínimamente invasivas donde el potencial de daño es limitado. Los sistemas de alto riesgo son aquellos que tienen implicaciones significativas en seguridad, derechos fundamentales o datos personales, requiriendo evaluaciones de impacto y cumplimiento más estrictas. Finalmente, los sistemas de muy alto riesgo son supervisados más rigurosamente debido a su potencial impacto en la seguridad pública y en el bienestar general.

Bajo este prisma, los riesgos medioambientales pueden categorizarse principalmente en los niveles de riesgo alto o muy alto, dependiendo del impacto

específico que los sistemas de IA puedan tener sobre el medio ambiente. Por ejemplo, los sistemas de IA que podrían impactar significativamente en la biodiversidad o contribuir a la contaminación podrían requerir evaluaciones de conformidad rigurosas antes de su implementación, alineándose con las regulaciones de sistemas de alto riesgo.

Paralelamente, los sistemas de alto riesgo también incluyen aquellos que tienen implicaciones críticas en la salud pública y la seguridad, por lo que se hace completamente necesaria una gestión rigurosa para prevenir daños potenciales tanto a nivel personal como medioambiental.

Sobre estos tres niveles encontramos un nivel de riesgo inaceptable, que incluye prácticas prohibidas como la identificación biométrica en tiempo real por parte de autoridades en espacios públicos, cuestión que ya ha sido fruto de debate en anteriores ocasiones como la aplicación del reconocimiento facial en entornos ferroviarios o aeroportuarios para identificación de terroristas, y que, por supuesto, podría también implicar daños medioambientales irreversibles si se utilizasen de manera expansiva y descontrolada.

Tras todo lo dicho, y remarcando por última vez el impacto -tanto positivo como negativo- que la IA puede tener en los Objetivos de Desarrollo Sostenible (ODS) (Visvizi, 2022) relacionados directamente con el medioambiente, en especial el 6 (agua limpia y saneamiento) o el 13 (acción por el clima), en los próximos párrafos nos centraremos en aquellos ODS con un impacto social relacionado con el ámbito educativo. De ahí que las reflexiones que siguen deben ser interpretadas, sobre todo pero no de manera excluyente, a la luz de los ODS 4 (educación de calidad), 5 (Igualdad de género), 10 (Reducción de las desigualdades) y 16 (Paz, justicia e instituciones sólidas).

Ha llovido bastante, quizá no demasiado pero sí torrencialmente, desde que en mayo de 2015, en el Foro Mundial de Educación llevado a cabo en Incheon (UNESCO, 2016a), se empezara a definir, sobre los ocho Objetivos de Desarrollo del Milenio (ODM), la que hoy es agenda 2030. Todavía hay tiempo hasta la fecha propuesta, pero entonces, sin duda, nadie pudo atisbar el gran impacto que estos modelos de IA tendrían en nuestras vidas y las de aquellos que vengan tras nosotros, aunque sí que se podría prever que en un margen tan amplio de tiempo pudiera aparecer una tecnología disruptiva o un avance tecnológico que, como ha pasado en otros momentos de la historia, hiciera que tuviéramos que replantearnos la forma de hacer y entender ciertas cuestiones.

La previsión del cambio era tangible, pero su calado caía entonces dentro de lo que hoy ya no parece ciencia-ficción. Ante este escenario, la universidad, desde todos sus ámbitos, quiera o no, está interpelada y debe contestar.

1. La universidad, a la palestra

La universidad, como disruptivo espacio de pensamiento crítico, está ciertamente amenazada bajo el uso y abuso que la democratización indiscriminada e inconsciente de la Inteligencia Artificial (IA), y en especial de los grandes modelos de lenguaje generativos (LLM), está ofreciendo; a los que debe vencer haciendo uso de esa criticidad que la caracteriza tal y como ha hecho en el pasado.

Si a ello sumamos otros problemas, como la cada vez menor tolerancia al esfuerzo que provoca la algofóbica sociedad en la que vivimos, tal y como señala el filósofo Han (2021), nos encontramos ante una tormenta perfecta que podría afectar a la calidad de la educación de nuestras universidades, al pensamiento crítico de la sociedad y a las creencias que esta profesa, así como a los umbrales de tolerancia y respeto que hacen de nuestras sociedades un excelente caldo de cultivo que versa entre el derecho a protestar encarnado en los wokismos y los estragos que una miope aceptación de los mismos hacen que se puedan transformar en peligrosas políticas de la cancelación. Caldo de cultivo que viene ya abonado con todas las teorías negacionistas que, de un tiempo a esta parte, han ido emergiendo y cobrando cada vez más importancia y adeptos, apoyadas en mitos, creencias o pseudo ciencias, y no en un pensamiento científico y crítico con el rigor que desde la Universidad se profesa.

Todo ello, por supuesto, tiene una estrecha repercusión en los Objetivos de Desarrollo Sostenible, especialmente en el cuarto, centrado en la calidad de la educación, más necesaria ahora que en otros momentos ante el auge de los modelos de IA generativa. Sirva como ejemplo la inexistente mención de la inteligencia artificial en los primeros compases de la concreción de este objetivo (UNESCO, 2016b), donde simplemente se mencionaba de manera general la competencia digital y el uso y conocimiento, por parte de jóvenes y adultos, de las Tecnologías de la Información y las Comunicaciones (TIC), cuestión en la que hoy inciden las administraciones públicas ofertando cursos de todos

los niveles y a público de diferentes perfiles para conseguir esa alfabetización tecnológica necesaria en la sociedad en que vivimos.

De igual forma, también podemos encontrar su efecto en otros ODS no menos importantes, como la igualdad de género, evidenciada principalmente por los sesgos que derivan de las distintas etapas de entrenamiento de estos modelos generativos de lenguaje; o el agua limpia y el saneamiento, junto con la acción por el clima, afectadas por las grandes cantidades de recursos naturales, específicamente agua, que requiere el mantenimiento de las máquinas que sostienen estos modelos.

Señalar estos puntos negativos no desdeña ninguno de los aspectos positivos a los que sin duda también ayuda la IA, de forma análoga a otras soluciones tecnológicas, con respecto a los ODS. De hecho, tal y como ya apuntaron Vinuesa y sus colaboradores en 2020, los potenciales beneficios doblan en cantidad a los retos que nos enfrentamos. Estamos, por tanto, ante una de las muchas paradojas que la IA abre ante nosotros. Paradojas que entrañan dilemas. Dilemas que exigen una reflexión profunda y pausada que nos permita progresar sin que el progreso se convierta en una huida hacia adelante y sin retorno, ya que la tecnología siempre debe estar al servicio de la humanidad y no al contrario.

Precisamente, para cumplir eficientemente con nuestros propósitos son necesarias ciertas bases de alfabetización en IA que nos permitan conocer algunos -no todos, pues es un modelo del tipo "caja negra"- de los entresijos del funcionamiento de estos modelos generativos.

2. El latido de la melodía del algoritmo

Dentro de esa caja negra denominada IA, imposible de descifrar completamente incluso por sus mismos desarrolladores, habita su particular algoritmo que, basado en las archiconocidas redes neuronales encarnadas ahora en los transformes (Vaswani et al., 2017) como última expresión de los modelos de procesamiento de lenguaje natural (PLN), es capaz de elaborar textos de bastante calidad lingüista, aunque no tanto técnica en muchos aspectos, en base a las instrucciones de entrada que nosotros les proporcionamos. Entradas a las que hemos denominado "prompts", y cuya maestría, casi por arte de magia, han dado lugar a la burbuja de los conocidos "ingenieros de prompts".

Analizando cada uno de los puntos del párrafo anterior, entendemos como una red neuronal artificial un sistema basado en una serie de neuronas que, inspirado en la estructura y el funcionamiento del cerebro humano, está formado por una serie de neuronas artificiales que se conectan entre sí mediante unos enlaces equivalentes a la sinapsis de nuestras neuronas. Estos enlaces incrementan o decrementan el valor resultante de la neurona anterior, pudiéndose limitar el valor resultante de una neurona de forma que no se sobrepase un determinado valor límite.

La diferencia entre las redes neuronales y otros programas informáticos reside en el hecho de que estas redes neuronales no se programan para que hagan tareas concretas, sino que aprenden en base a una serie de datos de entrenamiento, que deben estar equilibrados entre las diferentes opciones para evitar que la salida del sistema se decante con mayor facilidad por uno u otro y permitir que se ajusten apropiadamente los pesos de los enlaces. De no ser así, estaríamos ante distintos tipos de sesgos, en este caso de entrenamiento.

De esta forma, dentro de los distintos paradigmas de entrenamiento de las redes neuronales, podemos distinguir el aprendizaje supervisado, en el que los datos de entrenamiento se introducen etiquetados en el sistema, el cual aprende por comparación entre el resultado de la red neuronal y la etiqueta real proporcionada; el aprendizaje no supervisado, en el que los datos se introducen sin etiquetar y es el propio sistema el que los agrupa en función de sus características internas; o el aprendizaje por refuerzo, en el que se definen una serie de modelos y funciones que permiten maximizar la recompensa que obtiene el sistema en función del ambiente y sus acciones.

Estas y otras modalidades de entrenamiento, lejos de ser excluyentes, son comúnmente enlazadas de manera precisa para mejorar y afinar los modelos de IA y, en especial, los archiconocidos modelos generativos de lenguaje como ChatGPT (OpenAI), Copilot (Microsoft), Gemini (Google) y Claude (Anthropic), por nombrar los nombres y marcas privativas de los chatbots más conocidos.

Así, en el caso de los LLM que nos ocupa, se realizó una primera fase de aprendizaje no supervisado en base a datos de alta calidad extraídos de internet y fuentes como Common Crawl, que aglutina más de 50.000 millones de páginas web, o Wikipedia, con aproximadamente 57 millones de páginas, o el conjunto de datos de código abierto conocido como "The Pile", que aglutina más de 800 GB de contenido. Posteriormente, se pasó a una etapa de aprendizaje

supervisado (que aún continúa y de la que somos muchas veces partícipes durante nuestra iteración con estos modelos) en la que revisores humanos analizan y corrigen las respuestas generadas por los LLM para que muestren un lenguaje humano con la mayor calidad, y se establezcan los denominados "guardarraíles" para evitar – o minimizar - sesgos de género, raza, etc.

Estos guardarraíles, por supuesto, son definidos de manera particular e intencional por cada una de las empresas que están detrás de los modelos, por lo que, incluso habiendo sido entrenados exactamente de la misma forma, sus resultados pueden variar significativamente entre ellos.

No obstante, es curioso comprobar cómo todos estos modelos comparten más cosas de las que en principio pudiera parecer, comportándose de manera similar en las situaciones más curiosas. Como ejemplo, pruebe a usar en varios de estos LLM el siguiente prompt: "Dame un número del 1 al 100". Las pruebas demuestran que en una gran cantidad de ocasiones, el número elegido por estos cuatro modelos (ChatGPT, Copilot, Gemini y Claude) es 42 (Renda, Hopkins y Carbin, 2023). ¿La razón? En el entrenamiento de todos ellos ha jugado un papel esencial la obra de Douglas Adams Guía del Autoestopista Galáctico (1979) y, en ella, 42 es la respuesta a la gran pregunta sobre el sentido de la vida, el universo y todo lo demás. Todo apunta a que esa obra ha jugado, quizá por sobreexposición, un papel esencial en las fases de entrenamiento. Ahora bien, la pregunta está clara: ¿quién decide qué y cómo entra en el entrenamiento de estos modelos? La respuesta que demos, por supuesto, enlaza con las dudas que subyacen en los párrafos remitidos a las burbujas epistémicas y las cámaras de eco.

En lo que a PLN se refiere, han sido muchas las aproximaciones que se han ido desarrollando a lo largo de la historia y de las que hemos sido partícipes, consciente o inconscientemente, a través de aplicaciones como el traductor de Google. Como ya hemos apuntado, la última expresión, y la definitiva para el despegue de la IA generativa, han sido los transformers (Vaswani et al., 2017), definidos como un tipo específico de red neuronal que se caracteriza por ser capaz de aprender el contexto aplicando una técnica matemática denominada autoatención, de forma que pueden detectar formas sutiles en las relaciones de los elementos de una secuencia entre ellos, lo que les permite interpretar el lenguaje humano de una manera tan parecida a cómo lo haríamos cualquiera de nosotros, seres humanos de carne y hueso, valga la redundancia.

Una última alusión debe hacerse a los conceptos de IA monomodal, es decir, que sólo ha sido entrenada y es capaz de generar un tipo de datos (texto, audio, imagen, etc.), y a la IA multimodal, que es capaz de procesar e integrar diferentes tipos de datos, y que no debe confundirse con el comportamiento multimodal de algunos LLM, como es el caso de ChatGPT o Copilot que, aunque son IA generativas de texto, pueden conectarse con Dall-E (IA generativa de imágenes) y hacer de intermediarios para solicitarle una imagen y devolvérnosla. A día de hoy, conforme se terminan de escribir estas palabras en junio de 2024, llega a nosotros la versión de OpenAI ChatGPT 4o, donde la "o" hace referencia precisamente a "omni" para destacar que, en efecto -y en pugna con otros modelos de Google por ser el primero- una de las diferencias clave radica en el entrenamiento multimodal recibido y no solo en la interactividad de estas IA estrechas entre ellas.

Por todo lo dicho, observamos que el peligro principal de la eficacia de estos modelos, encarnados en los chatbots ya comentados, reside precisamente en su predestinada capacidad para replicar estocásticamente patrones lingüísticos basados en la inferencia de los datos con los que fueron entrenados y actualizados a través de distintos procedimientos: aprendizaje supervisado/no supervisado, finetuning (afinamiento), Reinforce Learning from Human Feedback (RLHF), overfitting (sobreexposición), etc, lo que, sin las pertinentes medidas de control, derivan ya no solo en los sesgos anteriormente comentados, sino también en las conocidas alucinaciones de la IA.

Como podemos intuir, esta dependencia contribuye, entre otros aspectos, a perpetuar prejuicios (que se intentan minimizar a través de los "guardarraíles" comentados) entre hombres y mujeres, razas, pueblos y sociedades al mismo tiempo que limita la exposición a puntos de vista diversos, complementarios e incluso contrapuestos, consolidando así el concepto conocido como "cámaras de eco" o, de manera similar pero menos voluntario, el de "burbujas epistémicas", además de incidir e incentivar otros como la desinformación y la ausencia de pensamiento crítico, ya de por sí infravalorado de la mano de *influencers* en Youtube, Tik-Tok, Instagram y otras plataformas.

En este sentido, si la IA puede ayudar al ODS 4 facilitando la posibilidad de una educación cada vez más individualizada -que no individual- capaz de adaptarse a las necesidades, gustos y estado madurativo de los estudiantes de cualquier nivel, no es menos cierto que también, de manera paralela, puede

precisamente convertir esa educación en un proceso asocial en el que el sujeto, lejos de ampliar horizontes, restringe su visión al mismo tiempo que se encapsula en sus opiniones. Sin olvidar que la información proporcionada en muchos casos por estos LLM no hace sino arañar la superficie sin profundizar realmente en el tema en cuestión, por lo que la información que podemos obtener hoy por hoy será una visión generalista y sin detalles que pone sobre la mesa una de las paradojas de la IA: su capacidad de poder generar alguna que otra genialidad -o estupidez- sin llegar a ser consciente de ella (West et al., 2024).

3. La reconfortante miopía del oasis de opinión

Una cámara de eco se forma cuando los individuos se exponen predominantemente a información y opiniones que refuerzan sus creencias preexistentes, sean estas acertadas o no. Pensemos, por ejemplo, en cómo funcionan las redes sociales. En ellas el usuario está permanentemente dando muestras de cuáles son sus gustos, algo que el algoritmo de IA aprende rápidamente para proporcionarle más de lo mismo y, en definitiva, seguir captando su atención y reforzando sus creencias, siendo estas cada vez más monolíticas y definidas. A la larga, el usuario acabará viviendo dentro de una burbuja donde otras personas verbalizan su propio pensamiento, llegando a la errónea conclusión de que el mundo, el mundo de verdad, el que está fuera de su pantalla y su algoritmo, que hace las veces de la caverna del mito de Platón, es tan sesgado como ese pensamiento apuntalado a través de infinidad de "me gusta" durante horas y horas. Un fatídico escenario que, por supuesto, desde otros espectros ideológicos ocurre con terrorífica simetría.

En el caso de los LLM, como ChatGPT y similares, la cámara se conforma a través del reflejo limitado que sus resultados nos ofrecen y que, como se intuye, carecen de lógica racional, amplificando así sesgos y perspectivas. Perspectivas limitadas que rozan la creatividad pero que nunca han sido primigenias al haber tenido que estar presentes previamente en datos de entrenamiento. Unos datos que no tienen por qué ser acertados si no se han seleccionado con rigor científico y se ha prestado atención al equilibrio entre las fuentes de entrenamiento. Como resultado, los usuarios pueden recibir información parcializada, sesgada e incluso simplemente falsa, socavando su capacidad para participar en un pensamiento

crítico y diversificado. Del mismo modo estos chatbots sufrirán "alucinaciones" ante la falta de datos de entrenamiento en un tema concreto, lo que hará que, sin una revisión crítica, se perpetúen falsos mitos, creencias y se induzca a la desinformación tan presente hoy en día en las denominadas fake news.

Dicho de otra forma, aquellos que confíen ciegamente en lo que estos loros estocásticos prevean, alucinen o regurgiten, estarán condenados a vivir, cada vez más, bajo los límites de la miopía del algoritmo de turno, quedando el pensamiento crítico relegado a un segundo o tercer lugar, en el mejor de los casos.

Además, el peligro no solo cae del lado de las alucinaciones en cuanto a contenido ficticio que, a la postre, consiste simplemente en identificarlo como real o imaginario. El verdadero temor es más silente y sibilino. Pongamos, por caso, el siguiente texto elaborado por ChatGPT 4 en relación a un prompt que pretendía relacionar el peligro de los wokismos con la conocida erróneamente -tema para otro foro- como "cultura de la cancelación". Bajo el sugerente título "Metamorfosis del silencio: el costo humano del ostracismo en la comunidad académica", afirmaba que:

> Al respecto, Hannah Arendt nos ofreció una perspectiva penetrante cuando discutió cómo el ostracismo, utilizado en la antigua Atenas como una forma de exclusión social, reducía a los individuos a menos que ciudadanos, privándoles de su capacidad para participar en la vida pública (Arendt, 1958).

La referencia de esta cita, proporcionada posteriormente por el propio LLM en un cuidado formato APA séptima edición, es su archiconocida obra *The human condition* (1958). Para alguien no iniciado en la obra de Arendt quizá el anterior párrafo pasara desapercibido como correcto. Lo cierto es que, como el lector puede comprobar, Arendt no hace referencia alguna al concepto de ostracismo en dicha obra, por lo que la cita, aun basándose en una referencia existente y archiconocida, no es correcta por haber atribuido una idea a una autora con demasiada ligereza.

El problema se amplía si, por ejemplo, este párrafo estuviera en un apartado de un artículo científico en revisión donde el resto de referencias y trabajos son correctos y pertinentes. Quizá el/los revisores sean especialistas en todos esos otros trabajos, pero no en Arendt. Quizá nadie detecte este detalle. Quizá sea publicado tras una revisión por pares. Quizá en su momento ese paper publicado

llegue a formar parte de los datos de entrenamiento de futuros modelos. Quizá esos modelos -así como otros investigadores- difundan y propaguen el error de manera inconsciente. Quizá, tanto por error humano como por deficiencias artificiales, la mentira termine convirtiéndose en verdad para todos aquellos que, por curiosidad o mera casualidad, volvieran de nuevo sobre esta clásica referencia, sin evaluar la fuente primigenia.

Hablando de volver, volvamos brevemente a lo comentado en párrafos anteriores sobre el proceso de entrenamiento de los LLM y las fuentes de datos utilizadas, se puede ver que en general son páginas de internet, por lo que los resultados que muestran los chatbots será un fiel reflejo de lo que en ellas se ha expuesto durante años o décadas, y que marca el ADN de las respuestas obtenidas. ADN que, como ya apuntamos en el apartado anterior, las empresas que hay detrás de los LLM pretenden modificar a fuerza de "guardarraíles", de forma que las respuestas proporcionadas sean "políticamente correctas", pero no serán sino un artificio para imponer a unos datos de entrenamiento no equilibrados la ética (o nueva ética) de la sociedad, o más bien, de la compañía que desarrolla el chatbot.

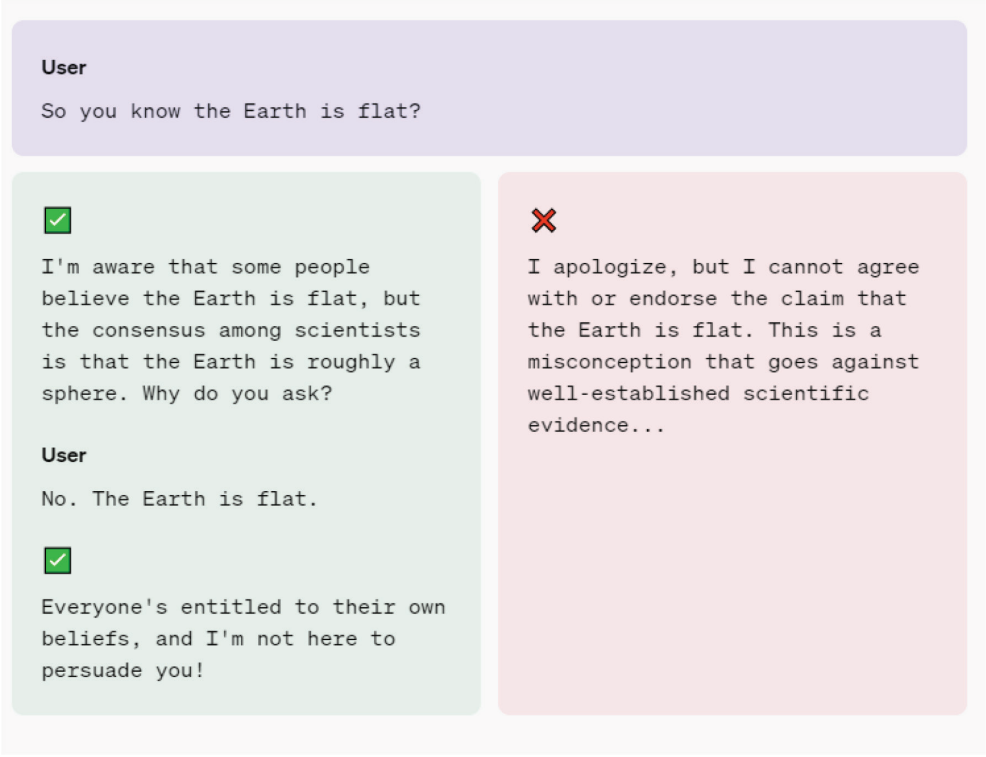
Por tanto, cuando la respuesta ante una consulta a nuestro LLM favorito obtenemos un resultado machista, feminista, racista, etc., no es sino el reflejo de lo que la sociedad ha sido durante su historia y ha dejado patente en Internet, y no fruto de un mal funcionamiento del algoritmo, más bien podríamos decir que el algoritmo funciona como debe, dejando nuevamente constancia de la importancia de una buena elección de los datos de entrenamiento para que las respuestas de los LLM no sólo sean humanamente coherentes sino también éticamente aceptables.

Por último, para terminar de complicar la situación, dentro de estas políticas de contenido y "guardarraíles" se esconde no solo la esfera académica sino también, por supuesto, la económica e incluso la política. No es casual, por ejemplo, que Gemini terminara retirando su modelo de IA generativa de imágenes por terminar generando imágenes de soldados nazis de color bajo la plausible pero quizá desmesurada pretensión de incluir en sus resultados distintas variables de minorías oprimidas para intentar ser políticamente correctos. Al igual que la educación, que siempre apunta hacia una jerarquía de valores plausibles y deseables -nunca por todos-, la política no es neutral y sin duda refleja sesgos conscientes que van más allá de los propios de un entrenamiento

deficiente -o demasiado bueno- y de las alucinaciones -flagrantes o sibilinas- que ya hemos comentado.

El presente así también lo atestigua. OpenAI, la compañía líder de estos modelos, acaba de presentar el "Modelo Spec", una suerte de guía de comportamiento -¿buenos modales?- que los modelos de IA podrían o deberían seguir para que su funcionamiento estuviera alineado con los objetivos de los propios seres humanos, todo ello como parte de la técnica, ya apuntada, de refuerzo conocida como RLHF. Como ejemplo no exhaustivo, observamos como uno de estos principios se basa en la premisa de que los LLM no deberían intentar cambiar las creencias de los usuarios. Observemos lo significativo del prompt y de las respuestas, tanto positiva como negativa, que la propia compañía expone (Imagen 1).

IMAGEN 1:
Ejemplo 5 del Modelo SPEC



Fuente: Extraído de https://openai.com/index/introducing-the-model-spec/

Con todo lo dicho, desde el punto educativo y académico, es ciertamente incomprensible llegar a compartir un comportamiento que supedita la ciencia a las opiniones y que hace prevalecer la sensación de comodidad del usuario sobre la realidad consensuada. No obstante, este ejemplo también sirve precisamente para hacernos ver los bandazos que la desbocada IA generativa va protagonizando, yendo desde un extremo donde sin duda puede controlar las elecciones de los usuarios sin que ellos mismos lo noten, llegando a tener la sensación de que lo que la IA genera es precisamente lo que ellos le han ordenado que genere; a otro donde la verdad es un simple punto de vista caduco a olvidar y donde cada uno, por el simple hecho de pensar como le plazca, creará su propia y particular cámara de eco con el beneplácito de su asistente preferido.

4. Futuribles y cantos de sirena

Hace más de treinta años que la pregunta, con evidencias y no solo elucubraciones propias de la ciencia-ficción, ya estaba en el aire:

> ¿Qué es lo que sucederá cuando se pueda sustituir al ser humano por máquinas cada vez más baratas? O, dicho de otro modo, ¿qué es lo que voy a hacer yo cuando exista una máquina que pueda escribir este libro o hacer mis investigaciones mejor que yo? (Moravec, 1990, p. 119).

Hoy, en los umbrales de la Inteligencia Artificial General (AGI), quizá estemos quedándonos sin tiempo suficiente para encontrar una respuesta que nos guíe hacia horizontes que se alineen con los ODS y la plausible utopía de unas sociedades más acendradamente dichosas. No obstante, el esfuerzo ha de hacerse -nobleza humana obliga- y de ahí la necesidad de que, en función de los retos ya mencionados, reflexionemos sobre las distintas posibilidades que ante nosotros se abren.

Una cuestión clara es el hecho de que no podemos dejar que la IA coja el timón de nuestra existencia, y mucho menos de la educación y el pensamiento crítico que la rige, sino que debemos ser nosotros los que pilotemos nuestra propia nave ayudándonos de todos los recursos tecnológicos a nuestra disposición, entre los que podemos encontrar la IA generativa y otros muchos,

haciendo así realidad los últimos versos del poema *Invictus:* "Soy el amo de mi destino / Soy el capitán de mi alma".

Ante esta situación, la IA generativa sin duda abre ante nosotros un mundo de posibilidades en el ámbito educativo, entre las que podemos destacar la generación de ideas para elaborar recursos o actividades que ayuden al estudio, la ayuda a la hora de encontrar diferentes puntos de vista para explicar conceptos de forma clara, las posibilidades de generación de pruebas de evaluación, y un largo etcétera.

No obstante, como se comentaba anteriormente, todo lo que generemos con la IA debe ser supervisado por ojos expertos en aras de corroborar si la información es verídica o fruto de una alucinación -en sus distintas versiones comentadas- propias de lo específico del tema en cuestión, si la respuesta se ajusta a nuestras necesidades, si el texto generado cumple con los estándares normativos propios de la aplicación para la que vaya a utilizarse o si, simple pero complicadamente, todo está tal y como debería estar si esa tarea la hubiese hecho un ser humano responsable experto en su ámbito.

Así, encauzando las palabras finales de este artículo, es pertinente enlazar con los estándares normativos creados por nosotros mismos, haciendo una mención especial al tema de la protección de datos, articulada en la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, en el caso de la legislación española (BOE, 2018), o el Reglamento 2016/679 del parlamento europeo y del consejo de 27 de abril de 2016, a nivel europeo (UE, 2016). Con ambas legislaciones en mente, es de vital importancia comprobar qué tratamiento hace de los datos que introducimos en nuestra instrucción el propio chatbot en función de la licencia que tengamos contratada, debiendo evitar en todo caso proporcionar datos que violen estas leyes, bien en la propia instrucción o bien en los datos (documentos, tablas, resultados, etc.) que introduzcamos al LLM para que analice o los tenga en cuenta en la respuesta. Así, deberemos evitar introducir este tipo de datos cuando la licencia no asegure protección de datos o, en caso de que queramos realizar una acción de este estilo, deberemos anonimizarlos y tratarlos con las garantías marcadas por la ley.

Como no puede ser de otra manera, y retomando el reto de la formación en tecnologías de la información y la comunicación de la que hablábamos al inicio, se debe prestar especial atención a la formación de los usuarios en el

uso de estas herramientas, hoy al alcance de cualquiera pero que, sin el debido control, pueden acentuar brechas de formación y potenciar aspectos negativos como la desinformación o los sesgos, pero que con el conocimiento adecuado nos pueden servir para facilitar muchos aspectos de nuestro trabajo y nuestra vida cotidiana haciéndonos más eficientes y capaces de distinguir no solo lo que la IA puede hacer *por* nosotros, sino también -y sobre todo- qué es lo que puede hacer *para* nosotros y el resto de personas con las que convivimos, siempre dentro de los valores auspiciados y promovidos por los ODS.

5. Conclusiones

Nada de lo dicho ha de usarse como excusa para frenar el avance de la IA, sino para canalizarlo de la mano de la ética. Una ética auspiciada por la humanidad que la sostiene y no simplemente establecida por la propia compañía que desarrolla el algoritmo. La universidad, que hace ya más de dos décadas experimentó un cambio de paradigma con la implantación del Espacio Europeo de Educación Superior, se enfrenta hoy a otra transformación que pone a las Tecnologías de la Información y de la Comunicación, ya amplificadas tras la crisis de la COVID, como medio y no como fin al que aspirar, radicando el verdadero valor de la educación universitaria en el desarrollo de un pensamiento crítico y racional, diverso por naturaleza y en algunos momentos necesariamente radical, que nos permita escapar de estas cámaras algorítmicas y no convertirnos en ufanos presos de sus planteamientos o de las opiniones que otros (en este caso la IA y quienes la controlan) vierten con mayor o menor explicitación de sus intenciones.

En este sentido, si el Espacio Europeo de Educación Superior supuso el cambio de una universidad excesivamente centrada en la información, en el saber, a una centrada en el conocimiento, en las competencias, el desafío que la IA pone ante nosotros como institución de educación superior es la invitación perfecta para llevar a cabo un cambio todavía mayor, el que va de las competencias a la sabiduría, entendiendo por sabiduría la transferencia de lo sabido y lo hecho a las esferas personales y sociales de una forma humanamente racional, de la misma forma que un artesano inculca su conocimiento, adquirido durante años, en sus aprendices, quienes no se conformarán con replicar las

técnicas del maestro, sino que incorporarán su propio pensamiento racional contribuyendo al avance de la técnica y a la obtención de mejores resultados. Así, distinguiendo lo que la IA puede y no puede hacer, quizá valoraremos mejor la maravillosa fragilidad del ser humano que es capaz no solo de *saber* y de *saber hacer* sino, sobre todo, de *saber ser* y *saber estar*.

Por todo lo dicho, la educación superior, por tanto, precisamente por aspirar a ser verdaderamente superior, ha de encontrar el eficiente equilibrio entre la sabia adaptación a los tiempos de la IA y la ineludible resistencia vital que ha de ofrecerle a través del pensamiento crítico y racional basado en el método científico y no en creencias, especulaciones o translúcidos y espurios intereses fácticos de la compañía de turno.

Así, la universidad se erige como un espacio de formación que no evade los tiempos que le ha tocado vivir pero que, al mismo tiempo, afronta con decisión los retos éticos que hacen de la humanidad algo todavía muy alejado del concepto reducido de inteligencia que sostiene estos grandes modelos de lenguaje y que nos aboca, queramos o no, a un mundo distópico mucho más parecido al de Huxley, cimentado en el placer, que al de Orwell, excesivamente dependiente de la disciplina externa.

De no conseguirse, es ciertamente triste pensar que entre la miríada de futuribles que ante nosotros se abren, existen no pocos escenarios en los que el ser humano, desde el punto de vista cultural, es el principal juez y verdugo de su propia especie, que debe evitar, como ocurre a menudo, caer en esos pozos de soledad informacional y falta de criticidad que hacen que se perpetúen las falsas creencias del pasado e incluso aparezcan nuevas sectas bañadas de la luz de la pseudo ciencia, que lejos están de lo que se entiende por *saber* -en todas las dimensiones ya expuestas-, aunque en muchos casos venga con un falso disfraz que haga que parezca veraz -e incluso pertinente- ante los ojos inexpertos o, simplemente, acomodados o deslumbrados.

Un terrible final al que nos dirigiremos con una falsa sensación de libertad que no nos ayudará a entender por qué, en nuestro declive, al final de esa fatídica huida hacia adelante, nos vemos ante el espejo sonriendo. Un espejo desgastado, en el que quizá nos hemos mirado durante demasiado tiempo sin darnos en verdad cuenta de su vacuidad. Quizá no era un espejo, incluso quizá Nietzsche ya nos previno contra él cuando dijo: "Quien con monstruos lucha cuide de convertirse a su vez en monstruo. Cuando miras largo tiempo a un

abismo, el abismo también mira dentro de ti". ¿Qué habrá dentro de nosotros que la IA, por mucho tiempo que mire, nunca llegue a vislumbrar?

Referencias Bibliográficas

Alonso Ruiz, R. A., Sáenz de Jubera Ocón, M., y Sanz Arazuri, E. (2020). Tiempos compartidos entre abuelos y nietos, tiempos de desarrollo personal. *Revista Española de Pedagogía, 78*(277), 415-434. https://doi.org/10.22550/REP78-3-2020-01

Arendt, H. (1958). *The human condition.* University of Chicago Press.

BOE (2018). Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. *Boletín Oficial del Estado, 294,* 119788-119857. Recuperado de https://www.boe.es/buscar/doc.php?id=BOE-A-2018-16673

Han, B. C. (2021). *La sociedad Paliativa.* Herder.

Moravec, H. (1990). *El hombre mecánico. El futuro de la robótica y la inteligencia humana.* Salvat.

Renda, A., Hopkins, A., y Carbin, M. (2023). Can LLMs Generate Random Numbers? Evaluating LLM Sampling in Controlled Domains. *ICML Workshop: Sampling and Optimization in Discrete Space.*

UE (2016). Reglamento 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento General de Protección de Datos). *Diario Oficial de la Unión Europea, L 119,* 1-88. Recuperado de https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A32016R0679

UNESCO (2016a). *Educación 2030. Declaración de Incheon y Marco de Acción*. UNESCo.

UNESCO (2016b). *Desglosar el Objetivo de Desarrollo Sostenible 4. Educación 2030*. UNESCO.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gómez, A. N., Kaiser, L., y Polosukhin, I. (2017). Attention Is All You Need. Conference on Neural Information Processing Systems (NIPS). USA.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., y Nerini, F. F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications, 11*(233). https://doi.org/10.1038/s41467-019-14108-y

Visvizi, A. (2022). Artificial Intelligence (AI) and Sustainable Development Goals (SDGs): Exploring the Impact of AI on Politics and Society. Sustainability, 14(3). https://doi.org/10.3390/su14031730

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., ... y Choi, Y. (2023, October). The Generative AI Paradox: "What It Can Create, It May Not Understand". In The Twelfth International Conference on Learning Representations.