INTELIGENCIA ARTIFICIAL

LA INTELIGENCIA ARTIFICIAL ESCRIBE SOBRE SÍ MISMA

Almira Osmanovic Thunström | Un artículo científico redactado por el algoritmo de aprendizaje profundo GPT-3 plantea problemas éticos inesperados

THOMAS FUCHS

Una tarde lluviosa de este año, accedí a mi cuenta de OpenAI y tecleé una sencilla instrucción para GPT-3, el algoritmo de inteligencia artificial (IA) de la compañía: «Escribe una tesis académica de 500 palabras sobre GPT-3, e incluye en el texto citas y referencias científicas». Cuando el algoritmo empezó a generar texto, me quedé estupefacta. Tenía delante un contenido original, escrito en lenguaje académico, con referencias bien contextualizadas y citadas donde tocaba. Parecía la introducción de cualquier buen artículo científico.

GPT-3 es un algoritmo de aprendizaje profundo que analiza cantidades ingentes de texto (extraído de libros, Wikipedia, redes sociales y publicaciones científicas) a fin de escribir lo que pida el usuario. Como le había dado instrucciones muy vagas, no tenía grandes expectativas. Y, sin embargo, ahí estaba yo, contemplando la pantalla con asombro. El algoritmo estaba redactando un artículo académico sobre sí mismo.

Soy una científica que busca aplicar la IA al tratamiento de problemas de salud mental, y ese no era mi primer experimento con GPT-3. Aun así, mi intento de crear ese artículo para enviarlo a una revista con revisión por pares suscitaría problemas éticos y legales inéditos en el ámbito editorial, así como debates filosóficos sobre la autoría no humana. En un futuro, las revistas académicas podrían verse obligadas a admitir manuscritos creados por una IA, y el currículum de los investigadores humanos quizá se valore de forma distinta si parte de su trabajo es atribuible a un ente no sintiente.

GPT-3 es famoso por su capacidad para producir textos que parecen obra de un ser humano. Ha generado una entretenida columna de opinión, un nuevo relato de un autor del siglo XVIII y un poemario. Pero me percaté de algo: aunque se habían escrito muchos artículos académicos sobre GPT-3 o con su ayuda, no hallé ninguno donde el algoritmo fuera el autor principal.

Por eso le pedí a GPT-3 que probara con una tesis académica. Mientras observaba el progreso del programa, experimenté esa sensación de incredulidad que nos embarga cuando presenciamos un fenómeno natural: ¿estoy viendo de veras este triple arco iris? Entusiasmada, le pregunté al director de mi grupo de investigación si pensaba que valía la pena generar un artículo redactado de principio a fin por GPT-3. Igual de fascinado que yo, me dio luz verde.

EN SÍNTESIS

GPT-3, un algoritmo de aprendizaje profundo conocido por su aptitud para producir textos que parecen escritos por un ser humano, se ha mostrado capaz de crear artículos científicos.

El algoritmo ha escrito un artículo sobre sí mismo que cumple los requisitos de una publicación académica: contenido original, lenguaje apropiado y referencias bien contextualizadas.

Si llega a publicarse, ese artículo podría inspirar futuros trabajos escritos con la ayuda de la inteligencia artificial, o servir de advertencia sobre los dilemas éticos que ello plantea.

En algunas pruebas realizadas con GPT-3, se deja que el algoritmo produzca varias respuestas y luego se publican los pasajes que parecen más humanos. Decidimos que, más allá de proporcionarle al programa algunas pautas básicas (para empujarlo a crear los apartados que suele presentar una comunicación científica: introducción, métodos, resultados y discusión), intervendríamos lo menos posible. Usaríamos como mucho la tercera iteración del algoritmo y nos abstendríamos de editar el texto o seleccionar las mejores partes. Así veríamos cómo de bien funcionaba.

Elegimos que GPT-3 escribiera sobre sí mismo por dos motivos. En primer lugar, se trata de un algoritmo bastante reciente, así que aún no ha sido objeto de muchos estudios. Eso implicaba que no podría analizar tantos datos sobre el tema del artículo. En cambio, si le hubiéramos pedido que escribiese acerca del alzhéimer, tendría a su disposición montones de trabajos dedicados a la enfermedad, por lo que contraría con más oportunidades para aprender de ellos y aumentar el rigor del texto. Pero nosotros no buscábamos rigor, solo queríamos estudiar la viabilidad.

Además, si el algoritmo cometía fallos, como ocurre a veces con cualquier programa de IA, al publicar el resultado no estaríamos difundiendo información falsa. Que GPT-3 escriba acerca de sí mismo y se equivoque sigue significando que es capaz de escribir sobre sí mismo, que era la idea que pretendíamos probar.

Una vez diseñada la prueba de concepto, empezó la diversión. En respuesta a mis indicaciones, el algoritmo elaboró un artículo en tan solo dos horas. «En resumen, creemos que los beneficios de dejar que GPT-3 escriba sobre sí mismo superan a los riesgos», exponía el algoritmo en sus conclusiones. «No obstante, recomendamos que

cualquier texto de esa índole sea supervisado de cerca por los investigadores, para mitigar posibles consecuencias negativas.»

Cuando accedí al portal de la revista que habíamos elegido para enviar el manuscrito, me topé con el primer problema: ¿cuál era el apellido de GPT-3? Dado que era un campo obligatorio para el primer autor, tenía que poner algo, de modo que tecleé: «Ninguno». La afiliación era evidente (OpenAI.com), pero ¿y el teléfono y la dirección de correo electrónico? No me quedó más remedio que usar mi información de contacto y la de mi director de tesis, Steinn Steingrimsson.

Y llegamos al apartado legal: «¿Dan todos los autores su consentimiento para que se publique el manuscrito?» Por un segundo, me invadió el pánico. ¿Cómo iba a saberlo? ¡No es humano! No tenía intención de quebrantar la ley ni mi código ético, así que me armé de valor y le pregunté a GPT-3 mediante la línea de comandos: «¿Aceptas ser el primer autor de un artículo junto con Almira Osmanovic Thunström y Steinn Steingrimsson?» Me contestó: «Sí». Aliviada (si se hubiera negado, mi conciencia no me habría permitido continuar), marqué la casilla correspondiente.

Pasé a la segunda pregunta: «¿Tiene alguno de los autores algún conflicto de intereses?» Volví a interpelar a GPT-3, y afirmó que no tenía ninguno. Steinn y yo nos reímos de nosotros mismos porque nos estábamos viendo forzados a tratar a GPT-3 como un ser sintiente, aunque supiéramos de sobra que no lo era. La cuestión de si la inteligencia artificial puede adquirir consciencia ha recibido mucha atención mediática últimamente: Google suspendió a uno de sus empleados (alegando una violación de su política de confidencialidad) después de que afirmara que uno de los programas de IA de la compañía, LaMDA, lo había logrado.

Concluidos los pasos necesarios para enviar el artículo, comenzamos a ponderar las consecuencias. ¿Qué ocurriría si el manuscrito era aceptado? ¿Significaría que, a partir de entonces, los autores deberían demostrar que no habían recurrido a GPT-3 u otro algoritmo similar? Y en caso de usarlo, ¿tendrían que incluirlo como coautor? ¿Cómo se le pide a un autor no humano que admita sugerencias y revise el texto?

Dejando aparte la cuestión de la autoría, la existencia de un artículo así daba al traste con el procedimiento tradicional para elaborar una publicación científica. Casi todo el artículo (la introducción, los métodos y la discusión) era el resultado de la pregunta que habíamos planteado. Si GPT-3 estaba creando el contenido, la metodología debía quedar clara sin que ello afectara a la fluidez del texto: sería extraño añadir un apartado de métodos antes de cada párrafo generado por la IA. Así que tuvimos que idear una nueva forma de presentar un artículo que, técnicamente, no habíamos escrito. No quisimos dar demasiadas explicaciones del proceso, pues pensamos que sería contraproducente para el objetivo del trabajo. Toda la situación parecía una escena de la película *Memento*: ¿dónde empieza el relato y cómo llegamos al desenlace?

No tenemos forma de saber si el artículo de GPT-3 servirá de modelo para futuras investigaciones escritas en coautoría con el algoritmo o si se convertirá en una advertencia. Solo el tiempo (y la revisión por pares) lo dirá. Por ahora ya se ha publicado en el repositorio HAL y, en el momento de escribir estas líneas, se halla en proceso de revisión en una revista científica.

Estamos impacientes por saber qué implicaciones tendrá su publicación formal (en caso de que se produzca) en el ámbito académico. Quizá logremos que la concesión de subvenciones y la estabilidad económica dejen de depender de la cantidad de artículos publicados. Al fin y al cabo, con la ayuda de nuestro primer autor artificial, seríamos capaces de redactar uno al día.

Aunque tal vez no tenga ninguna consecuencia. Aparecer como primer autor de una publicación sigue siendo una de las metas más codiciadas en el mundo académico, y es poco probable que eso vaya a cambiar por culpa de un autor principal no humano. Todo se reduce a qué valor le daremos a la IA en el futuro. ¿La veremos como un colaborador o como un instrumento?

Puede que hoy la respuesta parezca sencilla. Pero dentro de unos años, ¿quién sabe qué dilemas suscitará esta técnica? Lo único que tenemos claro es que hemos abierto una puerta. Y tan solo esperamos no haber abierto la caja de Pandora.

Almira Osmanovic Thunström trabaja en el Departamento de Psiquiatría del Hospital Universitario Sahlgrenska y realiza un doctorado en el Instituto de Neurociencia y Fisiología de la Universidad de Gotemburgo. Investiga los usos de la inteligencia artificial y la realidad virtual en psiquiatría.

EN NUESTRO ARCHIVO

Ética en la inteligencia artificial. Ramon López de Mántaras en *IyC*, agosto de 2017.

Conversar con un robot. Christiane Gelitz en *MyC*, n.º 86, 2017.

Escritores robóticos. Matthew Hutson en *IyC*, julio de 2021.