La inteligencia artificial escribe sobre sí misma

Una tarde lluviosa de este mismo 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 en los lugares adecuados. Parecía la introducción de cualquier publicación científica de calidad.

GPT-3 es un algoritmo de aprendizaje profundo que analiza cantidades ingentes de texto (extraído de libros, Wikipedia, conversaciones en redes sociales y publicaciones científicas) a fin de escribir lo que solicite el usuario. Dado que le había facilitado instrucciones muy vagas, no tenía muchas 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 explora cómo aplicar la IA en el tratamiento de problemas de salud mental, y este no era mi primer experimento con GPT-3. Aun así, mi intento de redactar 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 firmados 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 escritos por un ser humano. Ha generado una entretenida columna de opinión, un poemario y nuevas obras de un autor del siglo XVIII. No obstante, me percaté de algo: aunque se habían publicado muchos artículos académicos sobre GPT-3, y también con la ayuda de este, no pude localizar ninguno que tuviera al algoritmo como autor principal.

Esa es la razón por la que le pedí a GPT-3 que probara suerte con una tesis académica. Mientras observaba el progreso del

programa, experimenté esa sensación de incredulidad que le embarga a uno cuando presencia un fenómeno natural: ¿estoy viendo de veras este triple arco iris? Entusiasmada, me puse en contacto con el director de mi grupo de investigación y le pregunté 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.

Algunas aplicaciones basadas en GPT-3 permiten que el algoritmo produzca varias respuestas, y solo se publican los mejores pasajes, aquellos 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 los mejores fragmentos. Así comprobaríamos cómo de bien funcionaba.

Elegimos que GPT-3 escribiera acerca de sí mismo por dos sencillas razones. En primer lugar, se trata de un algoritmo bastante reciente, por lo 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 páginas y páginas dedicadas a la enfermedad, por lo que dispondría de más oportunidades para aprender de los estudios existentes y aumentar el rigor del texto. Pero nosotros no buscábamos rigor, solo queríamos estudiar la viabilidad.

Por otro lado, si el algoritmo cometía fallos, como ocurre en ocasiones 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 significa igualmente que es capaz de escribir sobre sí mismo, que era la idea que pretendíamos probar.

Una vez que diseñamos la prueba de concepto, empezó la verdadera 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 los textos de esa índole sean supervisados de cerca por los investigadores, para mitigar cualquier posible consecuencia negativa.»

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 ese campo era 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 proporcionar mi información de contacto y la de mi director de tesis, Steinn Steingrimsson.

Y entonces 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 infringir la ley ni mi código ético, así que me armé de valor y le pregunté directamente a GPT-3 a través de 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 seguir adelante), 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 me aseguró que no tenía ninguno. Steinn y yo nos reímos de nosotros mismos porque, llegados a ese punto, nos estábamos viendo obligados a tratar a GPT-3 como un ser sintiente, aunque sabíamos de sobra que no lo era. La cuestión de si la inteligencia artificial puede llegar a ser consciente ha recibido en los últimos tiempos mucha atención mediática: 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.

Una vez concluidos los pasos necesarios para enviar el artículo, empezamos a reflexionar sobre las consecuencias de nuestra acción. ¿Qué ocurriría si el manuscrito era aceptado? ¿Significaría que, a partir de ese momento, los autores deberían demostrar que NO habían recurrido a GPT-3 ni a 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?

Aparte de 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 documentación debía ser visible sin que ello afectara a la fluidez del texto; quedaría extraño añadir la sección de métodos antes de cada párrafo generado por la IA. Así

que tuvimos que inventar una nueva forma de presentar un artículo que, técnicamente, no habíamos escrito. No quisimos dar demasiadas explicaciones del proceso, ya que 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 modo en que decidimos presentar este artículo servirá de modelo para futuras investigaciones escritas en coautoría con GPT-3 o si se convertirá en una advertencia. Solo el tiempo (y la revisión por pares) lo dirá. El artículo de GPT-3 ya se ha publicado en el repositorio HAL y, en el momento de escribir estas líneas, se encuentra en proceso de revisión en una revista científica.

Aguardamos con impaciencia para conocer las implicaciones de 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 repercusión. Aparecer como primer autor de un manuscrito publicado sigue siendo una de las metas más codiciadas en el mundo académico, y es poco probable que eso cambie por culpa de un autor principal no humano. Todo se reduce a una pregunta: ¿Qué valor le daremos a la IA en el futuro? ¿La veremos como un colaborador o como un instrumento?

Tal vez parezca algo sencillo de responder ahora, pero dentro de unos años, ¿quién sabe qué dilemas suscitará esta tecnología? Lo único que sabemos es que hemos abierto una puerta. Y tan solo esperamos no haber abierto la caja de Pandora.

Con información revista Información y Ciencia