Hacia una inteligencia artificial MÁS HUMANA

EL año pasado se firmó el Llamamiento de Roma por la Ética de la Inteligencia Artificial, promovido por la Pontificia Academia para la Vida y firmado por una larga lista de instituciones públicas y privadas. En el breve texto se subraya que para que el avance tecnológico se alinee con el verdadero progreso de la raza humana y el respeto por el planeta, debe cumplir con tres requisitos: la no discriminación, el foco en el bien de la humanidad y la integración de una perspectiva basada en la sostenibilidad. Los firmantes, entre los que se encuentran numerosas instituciones públicas y privadas, se comprometen a incluir las consideraciones éticas necesarias desde el diseño inicial de los algoritmos.

LLAMAMIENTO A UNA IA MÁS HUMANA

Este Llamamiento de Roma por la Ética de la Inteligencia Artificial se une a una multiplicidad de guías y reglamentos que están intentando orientar los desarrollos de la IA y el uso de los datos, que ha dado grandes pasos desde la introducción del Reglamento General de Protección de Datos. Entre ellas podemos distinguir las «Guías para una IA digna de confianza», propuestas por un grupo experto en la Comisión Europea. Entre ellas se distinguen siete principios: protección de la agencia humana y supervisión por parte de seres humanos, robustez técnica y seguridad, protección de la privacidad, transparencia, no discriminación, foco en el bienestar social y ambiental y gobernanza responsable de la IA.

Este código se construye sobre otros, como las recomendaciones de la UNESCO, que buscan también el desarrollo de una IA más humana basándose en principios muy similares.

2022 PASARÁ A LA HISTORIA DE LA TECNOLOGÍA

Cuando hablamos de inteligencia artificial (IA) más humana, podemos referirnos a dos conceptos bien diferentes. Primero, a una IA que tenga capacidades más similares a las del ser humano. En esto, 2022 ha sido un año clave. Ya veníamos de un punto en el que la traducción automática pasaba de ser una mera curiosidad con poca utilidad práctica dados sus fallos a una herramienta efectiva en entornos personales y profesionales, este año, las capacidades de procesamiento del lenguaje de la IA han aumentado aún más.

La última de las noticias ha sido GPT3, un modelo basado en procesamiento del lenguaje natural y aprendizaje profundo que ha desatado la locura. Es capaz de clasificar textos de acuerdo con su tema y sentimiento (si es positivo o negativo), generar escritos nuevos siguiendo un tema dado o condensar la información disponible en resúmenes que bien pueden pasar por el borrador inicial de un informe.

Para que el avance tecnológico se alinee con el verdadero progreso de la raza humana y el respeto por el planeta, debe cumplir con tres requisitos: la no discriminación, el foco en el bien de la humanidad y la integración de una perspectiva basada en la sostenibilidad.

Unos meses antes, en julio de este año, nos sorprendían las noticias ligadas a otro de estos modelos, LMDA, un chatbot capaz de aparentemente sostener conversaciones con un interlocutor humano. Uno de los ingenieros responsable del modelo fue despedido tras difundir conversaciones en las que el modelo –aparentemente– aseguraba que tenía sentimientos y temía ser apagado.

DALL-E y otras IAs aplicadas a la generación de imágenes han sido también indudables protagonistas: producen figuras que responden a una descripción concreta, incluso estilística (por ejemplo, «una persona leyendo la revista *Mensajero* en estilo de Picasso»).

Estas IAs, cuyas obras ya han resultado ganadoras de concursos de arte, han generado incluso debates sobre la propiedad intelectual que no habían aparecido hasta el momento. A la vez, 2022 ha sido un año en el que han dominado las conversaciones sobre el metaverso, blockchain y herramientas afines, que tienen como resultado ampliar aún más el ámbito de actuación de la tecnología sobre nuestras vidas. Este incremento e intensificación de nuestra relación con la tecnología debería llevarnos a examinar cuidadosamente sus posibles consecuencias negativas. ¿Cómo podemos construir una IA más humana?

PRINCIPIOS ÉTICOS QUE NECESITAMOS

Probablemente la primera vez que los peligros del mal uso del Big Data pasaron a la consciencia pública fue con el escándalo de Cambridge Analytica. La historia comenzó en 2014, cuando un científico de datos creó un test de personalidad y lo hizo disponible en Facebook. La aplicación recopiló datos no solo de los usuarios que respondieron el cuestionario, sino también de sus contactos, lo que permitió recopilar datos de millones de usuarios de Facebook. Estos datos se vendieron a la a consultora Cambridge Analytica, que los utilizó para crear publicidad política personalizada para las elecciones presidenciales de Estados Unidos de 2016. Así, la explotación de datos personales pudo ser decisiva en el éxito de Trump entre otros. Del escándalo impactó no sólo la profundidad de los datos recopilados y la vio-

→

lación de privacidad a gran escala, sino que pudieran ser utilizados para manipular decisiones tan relevantes como unas elecciones.

Si el caso de Cambridge Analytica nos mostraba el potencial impacto de los algoritmos a nivel social, el escándalo de COMPAS mostró su capacidad de impactar en vidas individuales. Este algoritmo fue desarrollado en Estados Unidos con el objetivo de predecir si los presos estadounidenses tenían un riesgo alto o bajo de reincidir, información que se utilizaba directamente para decidir otorgarles la libertad condicional. Después de ser utilizado durante años y tras una investigación periodística, se encontró que el algoritmo era, de alguna manera, «racista»: a los presos afroamericanos, independientemente de su historial, se les asignaban peores predicciones de reincidencia que los blancos. El problema de COMPAS se debía a que, en los datos empleados para el entrenamiento del algoritmo (que comprendían miles de historiales de presos), se encontraban sobrerrepresentados los afroamericanos reincidentes. Esto llevó al algoritmo a aprender que los afroamericanos reinciden en mayor medida, y lo incorporó así a sus predicciones. El sesgo en contra de las personas de raza negra no lo introduce el programador, ni es un error en sentido estricto; lo introduce el mismo algoritmo funcionando de manera correcta según los datos que se le han proporcionado. De ahí su nombre: sesgo algorítmico.

La mayoría de los algoritmos de aprendizaje automático, especialmente el aprendizaje profundo (una versión cuantitativamente más potente), son lo que denominamos cajas negras. Proporcionan respuestas, pero sin justificarlas. En los algoritmos de caja negra, es solo posible detectar el sesgo algorítmico a posteriori y realizando análisis específicos para ello. Por ejemplo, en el caso de COMPAS enviaríamos perfiles similares de presos de diferente raza a la caja negra y compararíamos las respuestas. Pero, si no nos hemos dado cuenta de que la raza era un problema al construir nuestra base de datos, ¿cómo se nos podría ocurrir realizar estas pruebas? No es aceptable la delegación de decisiones con un impacto relevante en un algoritmo oscuro sin posibilidad de supervisión. Afortunadamente, existen alternativas a los algoritmos de caja negra que, en una gran variedad de aplicaciones, pueden presentar un rendimiento similar empleando criterios transparentes. Estos modelos configuran lo que se ha bautizado como IA Explicable o Interpretable, y conforman un paso crucial en la transición hacia una IA centrada en el ser humano.

Las principales preocupaciones éticas tienen que ver con la privacidad, la transparencia y el sesgo. Algunas aplicaciones parecen especialmente vulnerables a estos problemas.

Incluso más preocupante que el sesgo algorítmico es que se empleen filtros discriminatorios de manera consciente en aras de la eficiencia económica. Algo similar sucedió en Holanda con el escándalo de las ayudas al cuidado de niños, que llevó a la dimisión en bloque de su gobierno. Estas ayudas se retiraron a más de 26 000 familias en base a unos indicadores de riesgo desarrollados por la administración que incluían, por ejemplo, la raza o el hecho de tener doble nacionalidad. Al retirarse las ayudas de manera retroactiva, resultaban fácilmente en deudas de decenas de miles de euros, cantidades capaces de llevar a muchas familias a la pobreza. En más de mil ocasiones, las deudas llevaron incluso a que los padres perdieran la custodia de sus hijos. Estos hechos nos hicieron conscientes del enorme potencial negativo del uso de datos y cómo las administraciones –y no solo las empresas– pueden ser culpables de un mal uso de los datos y la tecnología.

Dada la velocidad de los desarrollos, no dejarán de aparecer problemas nuevos. Como señalan los ejemplos presentados, las principales preocupaciones éticas tienen que ver con la privacidad, la transparencia y el sesgo. Algunas aplicaciones parecen especialmente vulnerables a estos problemas y están creciendo espectacularmente,

como el filtrado automático de CVs o la estratificación de riesgo de fraude para la optimización del gasto público. No sería extraño que nos encontrásemos con más problemas en estas o aplicaciones similares.

Del mismo modo, la generación de contenidos cada vez más parecidos a creaciones humanas nos llevará a nuevos problemas. De alguno de ellos no se habla demasiado, como el uso de chatbots con el objetivo no solo de realizar tareas de soporte, sino de remplazar relaciones humanas. El ejemplo principal en este ámbito es Xiaolce, una IA que promete tener lo más parecido a una relación con una compañera perfecta, que aprende los gustos del usuario y apoya sus opiniones en las conversaciones o le envía mensajes de buenos días y buenas noches. Ya tiene decenas de miles de usuarios, y miles de ellos aseguran que prefieren a Xiaolce antes que una persona de carne y hueso. Es crucial que distingamos las interacciones humanas de las que no lo son, y eduquemos a los usuarios para actuar en consecuencia.

Es el momento de conocer la tecnología, anticipar sus problemas y establecer guías que permitan aprovechar los potenciales y evitar los peligros.

¿IA QUE SE PARECE A UN HUMANO O QUE AYUDE AL SER HUMANO?

Como hemos visto, los avances más sorprendentes de la IA la llevan a realizar tareas complejas con resultados lo más parecidos posible a los que podría dar un ser humano. Sin embargo, el objetivo de la IA –y de la tecnología en general, como herramienta ampliadora de las capacidades humanas– no es la de *imitar* al ser humano, sino la de *apoyarle*. La meta es mejorar la condición humana y la de la casa común.

Clarificar qué desarrollos y aplicaciones tienen un impacto positivo y evitar los peligros es cada vez más importante. La IA, combinada con las enormes bases de datos a las que necesita tener acceso para entrenarse, resulta en un enorme poder. Puede bombardearnos con publicidad dirigida o ayudarnos a escoger de manera más eficiente, multiplicar las posibilidades de creación artística o generar Deepfakes. Este es el momento de conocer la tecnología, anticipar sus problemas y establecer guías que permitan aprovechar los potenciales y evitar los peligros. El Llamamiento de Roma es un paso adelante en este camino.

SARA LUMBRERAS
Instituto de Investigación Tecnológica
Universidad Pontificia Comillas