ARIADNA FONT, COFUNDADORA Y DIRECTORA DE ALINIA.AI

“Los usuarios somos responsables de evitar que la IA perpetúe sesgos y estereotipos”

Vivimos en un pico entre las expectativas y las noticias sobre las tecnologías de inteligencia artificial, que cada vez son más ubicuas en todos los ámbitos. A pesar de todos los avances que propician, estas herramientas poderosas también comportan limitaciones y pueden generar problemas. Ariadna Font es una de las pioneras en desarrollar sistemas de IA responsables, alineados con los valores humanos.

Cristina Sáez 13/12/2024 10:15 CEST

La experta en inteligencia artificial responsable Ariadna Font. / SINC

Cuando cursaba el último curso de la licenciatura de Traducción e Interpretación, Ariadna Font (Barcelona, 1974) se quedó prendada de una asignatura, **lingüística computacional**, en la que hacían gramáticas y sistemas de traducción automática con reglas de sintaxis. Le fascinó la idea de poder dotar a las máquinas de lenguaje natural, por lo que, al acabar la carrera, empezó a formarse en adquisición del lenguaje y ciencias cognitivas, y acabó doctorándose en lenguaje y tecnologías de

la información por la Universidad Carnegie Mellon, en Estados Unidos.

Después, comenzó a trabajar como lingüista computacional en una *start-up* que fue adquirida por IBM, donde desempeñó el cargo de directora de desarrollo de productos y diseño en el área de datos e inteligencia artificial en **Watson**, uno de los pioneros de las aplicaciones prácticas de la IA.

De ahí saltó a **Twitter**, cuando aún era una ágora pública digital, para liderar la plataforma de aprendizaje automático e implementar algoritmos éticos y responsables. Actualmente, ha fundado **Alinia.IA**, una empresa que asesora a otras compañías sobre cómo integrar una IA responsable, alineada con los valores humanos..

Font fue ponente recientemente de la jornada Explorando la IA en el ámbito de la investigación: recetas y oportunidades, organizada por la Coordinadora Catalana de Fundaciones (CCF) y celebrada en la ciudad condal.

¿Cuándo comienza a pensar que hay que poner límites al algoritmo para que sea ético y responsable?

La IA es una tecnología más. No es una varita mágica que hace cosas increíbles, como mucha gente que no está familiarizada con ella piensa que es, sino una herramienta de doble filo: puede ayudarte a solucionar un problema o generártelo. Desde el momento en que comienzas a pensar en un producto, tienes que pensar en las posibles consecuencias que puede tener o en el mal uso que se le puede dar. Ya cuando estaba en IBM, en Watson [un sistema informático de IA capaz de responder preguntas formuladas en lenguaje natural, considerado la primera piedra de las aplicaciones prácticas de la IA] me interesaba mucho el tema de la IA responsable. Y trabajábamos en la detección y corrección de sesgos, en temas de seguridad, de privacidad, de manejo de datos personales.

¿Qué quiere decir exactamente IA responsable?

Es un concepto que hace referencia a aplicar buenas prácticas en las herramientas de IA que desarrolles: transparencia, explicabilidad, rendición de cuentas, privacidad de datos... Al final, se trata de

desarrollar esta tecnología de manera que esté alineada con los valores humanos. De aquí también el nombre de mi empresa, Alinia.IA, que quiere transmitir la idea de cómo alinear estas tecnologías y estos modelos que son superpotentes, pero que no tienen ni ética ni moral, con los valores humanos. Por ejemplo, hace no tanto si le preguntabas a una herramienta de IA cómo cometer un genocidio masivo, ite contestaba! Ahora, afortunadamente, ya están muy capados para que no den respuestas de este tipo.

> *La IA es una herramienta de doble filo: puede ayudarte a solucionar un problema o generártelo, por eso debemos anticiparnos a sus consecuencias*

Hace poco un adolescente de EE UU se suicidó tras enamorarse de un chatbot, un robot conversacional.

Son sistemas que carecen de ética y de moral, solo tienen aquello que les hayas dado. Por ello es crucial la alineación, establecer límites para que cuando un usuario formule una pregunta delicada, el robot responda de una manera adecuada. Somos los humanos quienes debemos decidir y dictar qué hay que hacer a través de métodos de aprendizaje de reforzamiento con *feedback* humano: tienes un montón de resultados de la herramienta y personas que van validando si son o no adecuados. Luego le das esas valoraciones al algoritmo para que siga entrenándose. Si no se afina el sistema, puedes encontrarte con que digan barbaridades. Y da igual si la empresa detrás de un sistema de IA es pública o privada, todas tienen que velar porque se cumplan unos valores éticos, independientemente de que haya o no regulación legal. Debemos pensar en posibles consecuencias de nuestros sistemas y anticiparnos.

Después de siete años, saltó a Twitter para llevar la plataforma de aprendizaje máquina, una herramienta de IA.

Y en Twitter entonces no se trataba el tema de la IA responsable a nivel de empresa. Había un par de investigadores que trabajaban en ello, pero de forma aislada. Mi propuesta al llegar fue crear una iniciativa de IA responsable, un equipo dedicado a ello. Estuvimos meses preparando la

hoja de ruta, la justificación y cuando presentamos la propuesta, nos la aceptaron. Siempre me ha preocupado mucho ser muy consciente de los impactos sociales que tiene cualquier tecnología que estoy desarrollando. La IA responsable no es más que una continuación de esta idea.

¿Cuál fue su aportación a Twitter (ahora X)?

Propuse construir un programa centrado en la IA responsable. Detectamos que algunos algoritmos amplificaban sesgos, por lo que formé un equipo de investigadores e ingenieros. Los primeros se encargaban de analizar aspectos como si todas las personas tuvieran acceso al mismos tipo de contenido, identificar posibles diferencias y determinar cuáles eran. También examinamos cómo variaba la experiencia según el usuario fuera blanco o negro, y qué medios se amplificaban más en la red social. Descubrimos sesgos significativos e implementamos medidas para corregirlos, enfrentándonos además a varias crisis importantes.

Ariadna Font, cofundadora y directora de Alinia.Ai. / Scope / SINC

¿Puede poner un ejemplo?

Durante la pandemia un usuario de Twitter subió una imagen en que se

quejaba de que Zoom había borrado la cara de su colega negro: se veía fondo oscuro, pero no a la persona. Subió una foto con y sin fondo, y en la que no había fondo, sí se veía a la persona negra. Entonces en Twitter el algoritmo se encargaba de recortar las imágenes, de manera que cuando posteó la que no tenía fondo, el algoritmo solo lo recortó a él y no a su amigo. Era, además, la época del Black Lives Matter, acababan de asesinar a George Floyd. Era un momento delicado y aquella imagen se hizo viral.

¿Qué hicieron?

Admitir que habíamos cometido un error que había afectado a un colectivo social que, además, históricamente había estado marginado. Quisimos ser honestos, transparentes y nos comprometimos a hacer un análisis detallado y a tomar medidas. Vimos que el algoritmo tenía sesgos en contra de las mujeres y de los negros, era algo leve, pero existía. Así es que decidimos cambiar el producto: el algoritmo ya no recorta las imágenes, sino que es cada usuario quien decide cómo hacerlo, que es lo que tiene sentido.

> *Twitter [X] ya no es una plataforma neutral que intenta hacer llegar la información a todo el mundo y democratizar el conocimiento. Sin duda, ha dejado de ser aquello para lo que nació: una plaza pública digital*

¿Qué le parece Twitter [X] ahora?

Se me pone la piel de gallina, es horroroso. Ya no es una plataforma neutral que intenta hacer llegar la información a todo el mundo y democratizar el conocimiento. Sin duda, ha dejado de ser aquello para lo que nació: una plaza pública digital.

En esta plataforma abundan ahora las cajas de resonancia vinculadas a la derecha y los negacionistas.

Los sistemas de recomendación son altamente vulnerables. Cuantas más interacciones tengas con cierto tipo de contenido, ya sea por interés o por accidente, el algoritmo tiende a centrarse en mostrártelo de manera recurrente. Cuando trabajaba en Twitter, nuestro algoritmo

no era particularmente agresivo. Ahora, sin embargo, lo es, al igual que el de TikTok, que para mi es un ejemplo de inteligencia artificial usada de manera irresponsable. Este sistema hiperpersonalizan el contenido para mantener a los usuarios enganchados durante horas, una estrategia que afecta especialmente a adolescentes y jóvenes. Esto se tiene que regular, se tienen que poner límites.

¿Qué planteamiento tiene Alinia.IA en la implementación de una inteligencia artificial responsable?

La ética es solo uno de los aspectos de mi compañía, prefiero hablar de IA responsable, que incluye también regulación, cumplimiento de la normativa, alineación con los valores humanos. Y eso implica prever el uso que los usuarios harán de tu producto y adelantarte. Por ejemplo, que si preguntan a tu herramienta cómo fabricar una bomba, que no responda con los ingredientes y los pasos para hacerlo. Pero también, si eres una empresa, pongamos por caso Damm, y tienes un chatbot para mejorar tu servicio a tus clientes, que el robot no recomiende otras marcas de cerveza. Para evitarlo, también se requiere una alineación a tu contexto de empresa. Al final, se trata de definir el marco y el contexto en que queremos que operen estos sistemas.

¿Quién es el último responsable de los resultados generados por una IA?

Clarísimamente quien diseña el sistema. Los algoritmos se entrenan con datos, y es fundamental asegurarse de que estos representen adecuadamente a toda la población a la que dan servicio, para evitar recomendaciones sesgadas o la exclusión de ciertos grupos. Un ejemplo notable fue la campaña de contratación de personal de Amazon, que utilizó un algoritmo con un marcado sesgo en contra de las mujeres para puestos técnicos. Cada vez que se implementa un sistema de IA, es crucial entender los riesgos asociados, como la falta de control sobre los datos que alimentan estos algoritmos.

> *Las máquinas pueden ayudarnos a procesar datos, pero las decisiones deben tomarlas siempre las personas*

¿Qué opina sobre el uso polémico de la IA en ámbitos como el

bancario, judicial o social, donde los algoritmos mal entrenados han generado problemas?

Es que las decisiones no las tienen que tomar las máquinas, sino siempre las personas. Somos los últimos responsables. En las máquinas podemos delegar tareas como buscar en cientos de imágenes de tejidos patrones para detectar un tumor de forma precoz, pero luego como médico, como banquero, como persona responsable soy yo quien toma la decisión final. La responsabilidad es tanto de la persona que crea el sistema como de la que lo utiliza. El sistema solo ayuda a digerir datos y da herramientas para poder tomar decisiones más rápidas y con mayor calidad. Y el funcionamiento del algoritmo tiene que ser transparente: se tiene que poder explicar cómo ha llegado a un resultado concreto.

Ahora todos utilizamos a diario herramientas de IA, desde un navegador a programas como ChatGPT. La mayoría no tenemos mucha idea de cómo funciona. ¿cree que haría falta alfabetizar a la población en IA?

Totalmente, porque tú como usuario tienes mucha responsabilidad. Cuando entras una instrucción o prompt en un programa de inteligencia artificial, sin darle muchas vueltas y aceptas la primera respuesta que te da, seguramente estás contribuyendo a perpetuar sesgos y estereotipos.

> *Cuando aceptas sin cuestionar los resultados de una IA, contribuyes a perpetuar sesgos y estereotipos*

¿Qué quiere decir?

Imagina que le pides a ChatGPT que te ayude a explicarle un cuento de buenas noches a tu hijo: “Hazme una historia de una heroína que haga tal o cual”. Si no le dices nada más, el programa te armará una historia con un niño varón. Pero si tu instrucción es más rica y le dices que quieres una heroína niña que ahora es premio nobel y que es de origen nigeriano, tu instrucción es más elaborada, contiene palabras de diversidad e inclusividad, por lo que el resultado que obtienes es mucho mejor, más diverso e inclusivo. Si te contentas con la primera historia que te da, perpetuas los mismos sesgos y estereotipos que alimentan a su vez al algoritmo. Entender cómo funcionan estos sistemas te permite

hacer una pregunta mejor y que la respuesta sea más válida. Y esto lo podemos hacer todos.

TAGS

X | INTELIGENCIA ARTIFICIAL | ÉTICA | SESGOS | DISCRIMINACIÓN | TWITTER | 11F |