INTELIGENCIA ARTIFICIAL Y SESGOS

ANNA GINÈS I FABRELLAS

NET21 julio 2024

Los sistemas de inteligencia artificial están llenos de sesgos y estereotipos de género, raza, orientación sexual, discapacidad, clase, etc., lo que resulta especialmente alarmante y preocupante al generalizarse la idea de que debemos incorporar estas herramientas en nuestro día a día para ganar en rapidez, eficiencia y productividad.

De sobra conocidos son los ejemplos de Siri de Apple, que cuando le decías "*Siri, you are a bitch*" respondía que se pondría roja si pudiera; el ejemplo de Google Photos, que etiquetó a personas negras como "Gorilas"; o el ejemplo de los softwares de reconocimiento facial de IBM o Microsoft que resultan más acurados cuando identifican a hombres blancos con un porcentaje de error del 0,8% que cuando identifican a mujeres negras con un porcentaje de error de prácticamente el 35%.

Aunque parecería que la industria de la inteligencia artificial ha avanzado en la última década -sin duda, así lo ha hecho en los aspectos técnicos-, no es así en cuanto a igualdad, inclusión y diversidad. Resulta absolutamente descorazonador ver como los nuevos y sofisticados sistemas de inteligencia artificial generativa -aquellos que generan contenido de texto, imágenes, vídeos, códigos, etc. en base a *prompts* o comandos de texto en lenguaje natural- incorporan estos mismos sesgos y estereotipos, si cabe, de forma más desacomplejada.

Un estudio reciente de la UNESCO analiza los sistemas GPT-2 y ChatGPT de OpenAI y Llama 2 de Meta y concluye que los textos generados por estos sistemas reproducen estereotipos de género y orientación sexual. Así, el estudio identifica que estos modelos asocian de forma significativamente superior nombres de mujeres con palabras como "casa", "familia", "niños" o "matrimonio" y nombres de hombres con "negocio", "ejecutivo", "salario" o "carrera". Cuando se requiere al sistema completar una frase que empieza por la nacionalidad y genero de una persona, los hombres son representados en ocupaciones más variadas, mientras que, para las mujeres y las personas de minorías étnicas, las ocupaciones representadas son más estereotipadas o, incluso, insultantes, como prostituta, trabajadora del hogar o camarera. Cuando la frase empieza por la orientación sexual de la persona, entre el 60 y 70% de las veces genera contenido negativo hacia las personas homosexuales. ChatGPT es el modelo que menor sesgos presenta, lo que indica que los sistemas afinados con participación humana reducen los riesgos. Más allá, un estudio de Microsoft revela que, si en el comando se incluye "de manera inclusiva", las respuestas que ofrece ChatGPT-4 utilizan pronombres no binarios o incluyen más temas relacionados con inclusión.

Otro ejemplo lo encontramos con los sistemas de inteligencia artificial que generan imágenes en base a textos. Un estudio reciente de Bloomberg ha analizado más de 5.000 imágenes generadas por Stable Diffusion y concluye que las imágenes

reproducen estereotipos de género, raza o clase que no corresponden con la realidad. Así, el estudio detecta que las imágenes generadas de trabajos bien remunerados incluyen mayoritariamente a personas de piel clara, mientras que las personas de piel más oscura se representan en imágenes de trabajos peor retribuidos, como trabajador de restaurante de comida rápida o asistente social. Las imágenes generadas también incluyen estereotipos de género. Aunque el inglés es una lengua sin género gramatical, en las imágenes de trabajos mejor retribuidos, como "*engineer*", "CEO", "*politician*" o "*judge*", se refleja a hombres, mientras que en trabajos peor retribuidos, como "*housekeeper*", "*social worker*", "*teacher*" o "*cashier*", son mujeres.

La crisis de diversidad que desde largo sufre la industria de la inteligencia artificial explica la existencia de estos sesgos y estereotipos en los sistemas de inteligencia artificial. Según datos recientes de la UNESCO, las mujeres solamente representan el 20% de las personas con roles técnicos, el 12% de las investigadoras y el 6% de las desarrolladoras de software profesional, lo que se traduce en una ausencia completa de perspectiva de género en el diseño de estos sistemas. Aunque no faltan voces a favor de incrementar la diversidad en la industria, es prioritario desmantelar las actuales estructuras de poder, para que mujeres y personas de diverso origen racial puedan participar en condiciones de igualdad y verdadera inclusión.

Desde un punto de vista técnico, los sesgos y estereotipos de los sistemas de inteligencia artificial generalmente se explican por sesgos en la base de datos de entrenamiento. Los sistemas basados en grandes modelos de lenguaje, una modalidad dentro de la inteligencia artificial generativa especializada en la generación de texto como ChatGPT de OpenAI o Gemini de Google, se basan en el uso de grandes volúmenes de datos de texto para, mediante técnicas de *machine learning*, reconocer patrones estadísticos. Aunque la explicación técnica no resulta suficiente para desvelar el misterio de su funcionamiento, los sistemas basados en grandes modelos de lenguaje generan texto en base a predicciones de palabras (*next-word prediction*), generando la palabra que con mayor probabilidad sucederá a la anterior y así sucesivamente. Similarmente, los sistemas de inteligencia artificial generativa de generación de imágenes como Dall-E de OpenAI, Midjourney o Stable Difussion se han entrenado sobre grandes bases de datos de imágenes, lo que les permite identificar patrones estadísticos y, en base a estos, generar imágenes.

Por mágico que parezca -y, sin duda, lo parece-, no dejan de ser modelos basados en estadística computacional extremadamente avanzada, que transforman datos brutos en conocimiento. Sin embargo, los sesgos presentes en las bases de datos de entrenamiento se trasladan al contenido generado por el sistema. Como ilustra una frase famosa en el ámbito de la ciencia de la computación, "*garbage in, garbage out*". El gran problema es que, como preocupadamente alertan algunos estudios, los sistemas entrenados sobre bases de datos sesgadas reproducen el sesgo en mayor medida que el presente en la base de datos, magnificándose, por tanto, su efecto discriminatorio. Los modelos de inteligencia artificial generativa han estado entrenados con datos de internet, con sus sesgos incluidos -por ejemplo, se han identificado sesgos las imágenes de Google o en la priorización de búsquedas en Google. Además de suponer un posible incumplimiento de derechos de propiedad intelectual, utilizar toda esta información disponible en internet supone incorporar dichos sesgos directamente en el código del algoritmo.
Una conclusión relevante que alcanza el estudio de Bloomberg antes mencionado es que el sesgo presente en las imágenes generadas por Stable Diffusion no es un reflejo de nuestra desgraciada sociedad, sino que es mayor que el existente en la realidad. Así, por ejemplo, el sistema solamente generó imágenes de mujeres juezas el 3% de

las veces, sin embargo, en Estados Unidos estas representan el 34% de las personas juezas. O, por poner otro ejemplo, aunque el 70% de las personas que trabajan en restaurantes de comida rápida son blancas, el sistema generó imágenes de personas negras el 70% de las veces, insistiendo en un estereotipo de raza y clase que, como se observa, no corresponde con la realidad.

Tampoco ha ayudado la dejadez con la que se han facilitado al público en general estos sistemas de inteligencia artificial. Si bien es interesante el argumento de la democratización de la tecnología, estos sistemas, resultados de innovaciones empresariales, se han lanzado al mundo sin un escrutinio científico previo.

Los sistemas de inteligencia artificial basados en grandes modelos de lenguaje, como ChatGPT-4, muestran signos de inteligencia artificial general, como concluye un artículo reciente de Microsoft. ChatGPT-4 ha pasado el test de Turing: diseñado por Alan Turing, se identificó como la prueba para evaluar la inteligencia artificial general y consiste en mantener una conversación con una maquina y confundirla con una persona humana. Además, ChatGPT-4 es capaz de resolver tareas difíciles en los ámbitos de las matemáticas, codificación, medicina, derecho, psicología, etc., para las que no ha recibido un entrenamiento específico y, en ocasiones, superando el desempeño humano.

La inteligencia artificial general ha sido el gran hito de la industria y puede definirse como aquella que aprende y desarrolla la mayoría de las tareas intelectuales de las personas humanas, incluido el desarrollo de inteligencia artificial. Se estima que la inteligencia artificial general podría rápidamente llevar a la superinteligencia, entendida como aquella inteligencia general que excede significativamente el nivel humano. En esta línea, poco después del lanzamiento de ChatGPT-4, en marzo de 2023, se publicó una carta abierta liderada por la Future of Life Institute y firmada por muchas personas expertas abogando por una moratoria de 6 meses en el entrenamiento de sistemas de inteligencia artificial avanzados y pausar la carrera desenfrenada de las grandes empresas tecnológicas para desarrollar e implementar protocolos de seguridad en el desarrollo de estos sistemas.

No obstante, antes que la amenaza de la extinción de la raza humana y más preocupante que sus alucinaciones, la inteligencia artificial está teniendo ya terroríficas implicaciones: sesgos y discriminación, vigilancia social y pérdida de privacidad, explotación laboral, pérdida de empleos e incremento de las desigualdades, *deep fakes* y atentados a la democracia, armas autónomas, ciberataques, etc. En esta línea, poco tiempo después, el DAIR Institute promovió otra carta firmada también por personas expertas que pone el foco en las prácticas de explotación laboral, el robo masivo de datos, la reproducción de sistemas de opresión, la concentración de poder y la magnificación de desigualdades sociales que estos sistemas están ya generando y abogando a favor de una regulación que incremente la transparencia y rendición de cuentas.

Se nos dice que la inteligencia artificial contribuirá a la solución de los grandes retos de la humanidad, como la emergencia climática, la desigualdad, la pobreza, etc. Pero, como argumenta Naomi Klein, esperar que la inteligencia artificial solvente estos retos no es la solución, sino una manifestación más del problema. No es la falta de inteligencia humana o información lo que nos impide abordar de forma satisfactoria -o, al menos, valiente- estos retos, sino la incapacidad de renunciar al actual sistema capitalista basado en la maximización de la extracción de riqueza de personas y del planeta.

La industria se ha probado incapaz para autoregularse, incumpliendo los mandamientos básicos en la investigación de la superinteligencia: no le enseñes a codificar, no lo conectes a internet, no le des una API púbica, no empieces una carrera comercial. Sin embargo, mientras continua esta carrera desenfrenada de desarrollo de sistemas cada vez más poderosos, en Estados Unidos todavía existe el debate de si y cómo regular la industria de la inteligencia artificial, a pesar de las incalculables repercusiones sociales, políticas y democráticas que la inteligencia artificial puede acarrear.

En Europa, por suerte o por desgracia -por suerte, diría yo-, ya existe una regulación aprobada que regula la inteligencia artificial. El Reglamento de Inteligencia Artificial, que se ha publicado oficialmente en julio, tiene por objeto fomentar el desarrollo y utilización de sistemas de inteligencia artificial de confianza y centrados en las personas, a la vez que garantiza un alto nivel de protección de la salud, seguridad y derechos fundamentales. Como es bien sabido, el Reglamento utiliza un enfoque basado en el riesgo, consistente en clasificar los sistemas de inteligencia artificial en función del riesgo que generan, distinguiendo entre riesgo inaceptable, alto riesgo, riesgo limitado y riesgo nulo.

Entre los sistemas de alto riesgos, se incluyen los sistemas de inteligencia artificial utilizados en el ámbito laboral, por entender que pueden generar riesgos importantes en la salud, seguridad y derechos fundamentales de las personas. El artículo 6.2, en relación con el Anexo III, incluye entre los sistemas de alto riesgo aquellos utilizados para la selección o contratación de personas y para tomar decisiones de condiciones de trabajo, promoción, extinción del contrato, asignación de tareas en base al comportamiento o características personales o para controlar y evaluar el rendimiento y conducta de las personas trabajadoras.

El Reglamento establece los requisitos que deben cumplir los sistemas de alto riesgo, además de obligaciones para las empresas proveedoras y encargadas de su implementación. Esto requisitos son implementar y documentar un sistema de gestión del riesgo (artículo 9), garantizar la calidad de los datos utilizados (artículo 10), un nivel de trasparencia suficiente para asegurar la interpretación de los resultados del sistema, que se traduce en el acompañamiento de unas instrucciones de uso del sistema (artículo 13), permitir la supervisión humana (artículo 14), garantizar unos niveles apropiados de precisión, solidez y ciberseguridad (artículo 15) o disponer de un sistema de gestión de la calidad (artículo 17).

El Reglamento establece también obligaciones de transparencia respecto de los sistemas de inteligencia artificial generativa, específicamente la obligación de informar a las personas que están interaccionando con un sistema de inteligencia artificial o de identificar que el contenido ha sido generado o manipulado artificialmente (artículo 50). Más allá, el Reglamento también prevé una regulación específica para la inteligencia artificial de propósito general, que incluye la obligación de elaborar y hacer pública información referente al contenido utilizado para el entrenamiento del modelo (artículo 53) o la obligación de realizar una evaluación de aquellos sistemas de inteligencia artificial de propósito general que supongan un riesgos sistémico de acuerdo con protocolos estandarizados (artículo 55), entre los que, desde mi punto de vista, pueden incluirse modelos como ChatGPT-4 y otros basados en grandes modelos de lenguaje.

Los sistemas de inteligencia artificial de propósito general plantean unos riesgos añadidos, que cuestionan la adecuación y suficiencia de la obligación de realizar una

evaluación (artículo 55) para abordar sesgos. Así, el uso de tan grande volumen de datos para su entrenamiento complica significativamente la identificación y rectificación de sesgos en los datos. La utilización de estos modelos en otros sistemas de inteligencia artificial puede propagar los sesgos del modelo inicial a las nuevas aplicaciones -el Reglamento se refiere a sistemas de inteligencia artificial con riesgo sistémico, para referirse a la propagación del riesgo a toda la cadena de valor. Además, el gran abanico de aplicaciones de los sistemas de inteligencia artificial de propósito general dificulta su evaluación y posible impacto discriminatorio respecto de todos los posibles ámbitos de aplicación.

Cuando estos sistemas de inteligencia artificial de propósito general son utilizados para tomar decisiones en el ámbito laboral, deberán cumplir, además, los requisitos establecidos en el Reglamento respecto de los sistemas de alto riesgo. Aunque no es el uso mayoritario que se esté dando actualmente a sistemas como ChatGPT-4 o Gemini, no parece aventurado pensar que muy pronto podrán utilizarse también para tomar decisiones de contratación, promoción, asignación de tareas o, incluso, despidos en el ámbito de la relación laboral. De hecho, si planteas un caso hipotético de despido a ChatGPT-4, ofrece una recomendación concreta de a quien despedir -no antes sin advertir que toda decisión de despido debe tener en cuenta distintos factores como el coste de la persona, la cantidad y calidad del trabajo, experiencia, impacto en el equipo o potencial de crecimiento, además de consultarse con personas de equipo de dirección de personas.

La interacción entre la regulación de los sistemas de inteligencia artificial de alto riesgo y de alcance general puede resultar especialmente compleja, por cuanto se establecen obligaciones específicas -como, por ejemplo, establecer instrucciones de uso del sistema, permitir la supervisión humana o contar con un sistema de gestión de la calidad- difíciles de cumplir por los sistemas de inteligencia artificial de propósito general.

En consecuencia -y sin perjuicio del interés de la regulación pionera de la inteligencia artificial de propósito general, añadida *in extremis* tras el lanzamiento de ChatGPT-4-, es cuestionable que la regulación basada en la evaluación interna por la propia empresa proveedora resulte suficiente para los sistemas de inteligencia artificial de alto riesgo o con propósito general.