Humanos, demasiado humanos: los algoritmos y sus prejuicios

Cuanto más se extiende el uso de la inteligencia artificial, más dudas surgen acerca de la objetividad de las decisiones que toma. La aparición de sesgos en sus dictámenes puede tener consecuencias muy negativas para las personas, dado que, un funcionamiento incorrecto, a menudo crea situaciones de discriminación. Las máquinas inteligentes deben ser auditadas para asegurar que no perjudicarán a ningún ser humano o colectivo.

En el sistema judicial de Estados Unidos se utiliza un algoritmo informático para predecir la probabilidad de reincidencia de los convictos. La agencia de noticias independiente ProPublica denunció en 2016 que COMPAS -así se llama el sistema- mostraba un sesgo en sus dictámenes que le llevaba a predecir que los reos negros presentaban un mayor riesgo de volver a delinquir que los blancos.

Por su parte, PredPol es otro algoritmo creado para determinar dónde es posible que vayan a tener lugar los crímenes en una ciudad, para que la policía pueda estar preparada y responder con mayor prontitud. Investigadores de la organización Human Rights Data Analysis Group descubrieron que, en una prueba realizada con este software sobre delitos relacionados con las drogas en Oakland, California, un gran número de veces enviaba patrullas a barrios mayoritariamente habitados por minorías raciales, independientemente de las tasas de criminalidad existentes en los mismos.

Un experimento realizado en el MIT con tres algoritmos de inteligencia artificial para el reconocimiento facial –de IBM, Microsoft y de la empresa china Megvii -conseguía identificar a un varón blanco a través de una fotografía con una exactitud del 99%. El problema es que cuando el sujeto es una mujer de piel oscura, la precisión en el acierto baja al 35%. Si fuesen utilizados en situaciones reales, podrían identificar erróneamente a mujeres y a minorías raciales.

El famoso motor de búsqueda Google tampoco se libra de los sesgos. Un estudio de 2015 demostró que al buscar imágenes bajo término "CEO" (consejero delegado), tan solo un 11% de las fotografías que presentaba era de mujeres, a pesar de que en Estados Unidos casi la tercera parte de los cargos ejecutivos están ocupados por mujeres.

Armas de destrucción matemática

Los algoritmos cada vez están más presentes en nuestras vidas. Muchos procesos relacionados con las personas se automatizan, como pueden ser la preselección de candidatos para ocupar un puesto de trabajo o la concesión de créditos bancarios, por poner tan solo dos ejemplos. Un programa informático recibe toda la información sobre cada persona y, en función de los parámetros con los que ha sido programado, realiza la evaluación.

El problema es que, según para qué sean utilizados, estos algoritmos pueden tener la responsabilidad de la toma de decisiones importantes que afectan a la vida de la gente. Pueden determinar que consigamos o no un trabajo, que podamos estudiar o no en un colegio o universidad solicitado, e incluso –como hemos visto más arriba en el caso de COMPAS- que se nos conceda la libertad provisional.

La automatización de estos procesos persigue lograr una evaluación de cada tema mucho más objetiva, eliminando los prejuicios propios de los humanos, pero, paradójicamente, aparecen sesgos y fallos en los algoritmos, que les llevan a discriminar a determinadas personas y colectivos.

Es por ello, que la experta Cathy O´Neil los define como *armas de destrucción matemática* en su libro del mismo nombre. Se trata de programas que pueden hacer mucho daño a mucha gente. Y lo peor es que las víctimas de sus decisiones no saben bajo qué criterios se les ha evaluado, pues el funcionamiento de los algoritmos es demasiado complejo y solamente es conocido por los técnicos que los diseñan.

¿Pero, qué es un algoritmo?

Hoy en día todos hablamos de algoritmos, pero no pocos ignoran exactamente qué significa esa palabra. Un algoritmo es un proceso o, como explica Wikipedia, *"un conjunto prescrito de instrucciones o reglas bien definidas, ordenadas y finitas que permiten llevar a cabo una actividad mediante pasos sucesivos que no generen dudas a quien deba hacer dicha actividad".*

Cathy O´Neil explica en su libro de forma muy didáctica lo que es un algoritmo, con un ejemplo doméstico: las cenas que cocina para su familia. Su "modelo" se nutre de información -las preferencias culinarias de sus hijos, los alimentos disponibles y su propia energía como madre y cocinera cada noche-, y tiene un resultado, qué plato decide cocinar y cómo. El grado de éxito lo mide en función de los comentarios de los comensales y

de si se han comido todo el contenido del plato.

La medida de disfrute de su familia le permite corregir y actualizar el proceso de cara a la próxima cena, creando lo que se conoce como un modelo dinámico. Si metiese toda esa información en un ordenador e introdujese ciertas reglas (por ejemplo, *la prohibición relativa a la comida basura se relaja en las cenas de cumpleaños*), dispondría de un "modelo automatizado", que podría utilizar cualquiera sin estar ella presente.

Un modelo es una simplificación de la realidad, en el que incluimos solamente las cosas que consideramos importantes, y que nos arroja conclusiones que nos ayudan a tomar decisiones. Un algoritmo utiliza grandes volúmenes de datos para elaborar un patrón de éxito. La clave de un modelo es cómo se define el éxito del mismo y, aquí precisamente, radica todo el peligro. Aquel que defina cuáles son los resultados esperados que se considerarán positivos, tendrá en sus manos todo el poder.

El maestro y el libro de texto

Los algoritmos no desarrollan sesgos discriminatorios por sí mismo, sino que reproducen los prejuicios de sus educadores. Como en el caso de un niño, su aprendizaje depende en gran medida del maestro que tenga y del libro de texto que utilice.

En esta metáfora, concebida por el investigador del MIT Rahul Bhargava, el *libro de texto* para la inteligencia artificial serían los datos con los que es entrenada para aprender a tomar decisiones. Se trata de grandes cantidades de datos que constituyen ejemplos de situaciones que se han resuelto a nuestra conformidad o que responden correctamente a la pregunta que queremos que el algoritmo aprenda a contestar.

Por ejemplo, si estamos entrenando a un sistema para que establezca la probabilidad que presentan los solicitantes de un crédito de devolver la deuda contraída a tiempo, nutriremos a la inteligencia artificial con información sobre casos de créditos cancelados correctamente, para que pueda extraer de ellos un patrón que describa al prestatario más proclive a cumplir sus obligaciones y, de esta manera, poder clasificar a los solicitantes en función de su riesgo de insolvencia.

El segundo elemento que interviene en el aprendizaje de la máquina es el maestro, es decir, la persona que hace las preguntas y que determina qué conjunto de datos debe considerar el algoritmo para elaborar una respuesta. En el ejemplo anterior, se puede indicar al sistema que tenga en cuenta datos del solicitante como la cantidad solicitada, el tiempo establecido para su devolución o su nivel de ingresos, pero también se pueden incluir otros, como su situación familiar, género o raza.

De esta forma, los sesgos discriminatorios que presentan los algoritmos son reflejos de nuestros propios prejuicios, dado que dependen de los datos con los que alimentamos al sistema y de las preguntas que le hacemos.

Imaginemos, siguiendo con el mismo ejemplo, que alimentamos el algoritmo creado para evaluar la solvencia de los solicitantes de crédito con historiales crediticios mayormente de personas de raza blanca. Seguidamente, le indicamos, entre los parámetros que debe utilizar para la toma de decisiones, la etnia del solicitante.

El patrón que genera el sistema sobre cómo es una persona solvente podría considerar que los prestatarios negros no lo son, dado que no encuentra entre la información que ha recibido suficientes ejemplos de ciudadanos de piel oscura que cancelan sus deudas y, lo que es peor, su maestro le ha indicado que la raza es un factor importante de cara a establecer un juicio. En consecuencia, dictamina un elevado riesgo de prestar fondos a solicitantes que no son blancos.

Rechazo social

Es un tema complejo y difícil de entender para los que no son técnicos, pero cada vez las sociedades son más conscientes de que la inteligencia artificial empieza a formar parte de las vidas de todos nosotros. Comienza a jugar un papel importante en muchos aspectos que nos afectan directamente. Y eso despierta preocupación.

Una encuesta de Pew Research Center, realizada este año, arroja el dato de que la mayoría de los ciudadanos estadounidenses considera inaceptable el uso de algoritmos para tomar decisiones que tengan consecuencias reales para los humanos.

En concreto, el 56% se muestra contrario a que se utilicen para evaluar el riesgo de reincidencia de convictos en libertad provisional y el 57% no cree que sea justo usarlos para analizar automáticamente las solicitudes y currículos de los candidatos a ocupar un puesto de trabajo. Sube el porcentaje, hasta el 67%, de los que consideran que no se debe automatizar el análisis de los vídeos de entrevistas de trabajo. Finalmente, casi el 70% de los encuestados ve mal que un algoritmo clasifique la solvencia de los solicitantes de crédito en función de variables relacionadas con el consumo.

En las preguntas abiertas de la encuesta afloran las causas que justifican este elevado rechazo a aplicar la inteligencia artificial en temas que afectan a las personas. Básicamente, son las siguientes:

- Viola la privacidad individual, pues en muchos casos hace uso de datos personales para la toma de decisiones.
- No es justa. Muchos de los encuestados consideran que son procesos que pueden derivar en la discriminación de las personas.
- Elimina el elemento humano de las decisiones importantes.
- Los humanos somos complejos y estos sistemas son incapaces de identificar los matices.

Auditorías de algoritmos

Un concepto emergente, vista la amenaza que plantean los sesgos, es el de la necesidad de auditar los sistemas de inteligencia artificial y el big data, para identificar y prevenir los posibles errores que conllevan. Hablamos de un proceso que examina cada paso del proceso llevado a cabo por la ciencia de datos.

Resulta interesante compartir aquí la metodología al respecto, que plantea la consultora ORCAA, y que se resume en cuatro etapas: los datos, la definición, la construcción y la monitorización.

Desde la perspectiva de los datos, es necesario plantear las siguientes preguntas:

- ¿Qué datos has recopilado? ¿Son relevantes, tienes suficientes y son los correctos?
- ¿Qué grado de objetividad tienen estos datos? ¿Presentan sesgos? ¿Hay algunos que sean menos precisos? ¿Cómo lo evalúas?
- ¿Tu información deja fuera de forma sistemática datos de carácter importante? ¿Están representando por defecto o por exceso determinados eventos, comportamientos o personas?
- ¿Estás limpiando los datos, teniendo en cuenta los datos perdidos, los periféricos o los que son irracionales?

En relación con la definición de la información que quieres obtener, hay que cuestionarse:

- ¿Cómo defines el éxito de tu algoritmo? ¿Existen otras definiciones de éxito relacionadas, que piensas que pueden aparecer si fuerzas esa definición?
- ¿Qué atributos eliges en la búsqueda para potencialmente asociarla con éxito o fracaso? ¿Hasta qué punto los atributos constituyen filtros y cómo podrían hacer fallar el proceso?

Y a la hora de construir el modelo en cuestión, hay que considerar:

- ¿Qué tipo de algoritmo debo utilizar?
- ¿Cómo calibro el modelo?
- ¿Cómo determino cuándo el modelo ha sido optimizado?

Finalmente, hay que planificar la monitorización del algoritmo:

- ¿En qué medida está trabajando el modelo en la producción de resultados?
- ¿Necesita ser actualizado regularmente?
- ¿Cómo se distribuyen los errores?
- ¿Está el modelo produciendo consecuencias no intencionadas?

Reacciones contra los sesgos

Afortunadamente, muchas de las instituciones más punteras en el campo de la inteligencia artificial están cobrando consciencia de los fallos que acarrean los algoritmos que desarrollan, y, en consecuencia, tratan de poner medios para poder corregirlos.

IBM anunció en septiembre el lanzamiento de una herramienta que analiza en tiempo real cómo y por qué los algoritmos toman decisiones. Se denomina Fairness 360 y es capaz de rastrear signos de que el sistema produce sesgos e incluso recomendar ajustes para corregirlos.

Ese mismo mes, Google presentaba What-If Tool, una aplicación que explora el funcionamiento a lo largo del tiempo de los sistemas de inteligencia artificial. Y también empresas como Microsoft y Facebook están trabajando en kits para detectar posibles sesgos en el funcionamiento de la inteligencia artificial.

¿Quién asume la responsabilidad?

Después de esta breve exposición sobre los riesgos que implica el uso de algoritmos para la toma de decisiones que afectan a las personas, creo que una cosa queda clara: el problema no es si un software concreto funciona bien o presenta fallos, lo preocupante es por qué depositamos esa responsabilidad sobre una máquina.

Todd Breyfogle de Aspen Institute, entrevistado dentro del ciclo Tech & Society[1], afirmaba que nos arriesgamos, en nombre de la eficiencia, a abdicar nuestras responsabilidades como seres humanos en los sistemas inteligentes.

Ciertamente, hay funciones y tareas que siempre deberían tener a una persona detrás, aunque estén fuertemente apoyadas por la computación. En palabras de Breyfogle:

> “Al final, mi preocupación no es si la inteligencia artificial es una amenaza para la humanidad, sino si los humanos acabaremos degradando nuestra humanidad al delegar nuestras obligaciones de cuidado moral en sistemas lógicos que parecen exhibir tan solo la apariencia del conjunto de experiencias que nos hacen humanos. Nosotros, los humanos, somos nuestra mayor amenaza.”

[Photo by Markus Spiske temporausch.com from Pexels]

Bhargava, R. “The Algorithms Aren’t Biased, We Are” en *MIT Media Lab*. Disponible en: https://medium.com/mit-media-lab/the-algorithms-arent-biased-we-are-a691f5f6f6f2

Kleinman, Z. “IBM launches tool aimed at detecting AI bias” en *BBC News* (2018). Disponible en: https://www.bbc.com/news/technology-45561955

NewScientist (2018) “Discriminating algorithms: 5 times AI showed prejudice”. Disponible en: https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/

O´Neil, C. (2017) “Armas de destrucción matemática. Cómo el big data aumenta la desigualdad y amenaza la democracia”. Madrid. Capitán Swing

ORCAA (2017) “What is a data audit?”. Disponible en: http://www.oneilrisk.com/articles/2017/1/24/what-is-a-data-audit

Smith. A (2018) “Public Attitudes Toward Computer Algorithms”. Pew Research Center. Disponible en: http://www.pewinternet.org/2018/11/16/public-attitudes-toward-computer-algorithms/