Editorial

Deus ex machina: intentando explicar la inteligencia artificial

Jiménez Llamas R[1], Jiménez Alés R[2]

[1]*Graduado en Matemáticas y Física. Máster en Física Teórica. Departamento de Estadística e Investigación Operativa. Universidad de Sevilla. Sevilla. España.*

[2]*Pediatra. UGC Puente Genil. AGS Córdoba-Sur. Servicio Andaluz de Salud. Córdoba. España.*

La inteligencia artificial (IA) es un campo de la ciencia en auge. Se nutre de las matemáticas y de la capacidad de computación de los ordenadores actuales para poder hacer predicciones a problemas muy variados y complejos. Es sorprendente, incluso para los que trabajan con IA, la capacidad que tienen muchos de los algoritmos creados. Pero ¿cómo puede ser sorprendente incluso para los creadores? ¿Acaso no saben lo que están haciendo? Pues, en cierta manera, no. Demos un ejemplo para el caso de redes neuronales: sin entrar en mucho detalle, imaginemos un análogo al funcionamiento de una red neuronal concreta que se encarga de reconocer si en una foto en blanco y negro hay un fenotipo X. A grandes rasgos, a cada píxel de la imagen se le puede asignar un número entero entre 0 y 1. El 0 si el píxel es de color negro y un 1 si es de color blanco. Los grises intermedios toman valores entre 0 y 1 dependiendo de si están más cerca del negro o del blanco. Nuestro análogo de una red neuronal toma cada uno de los números de cada píxel y les hace una serie de operaciones. Dependiendo de la red concreta, agrupa ciertos píxeles y realiza operaciones con sus valores, cada conjunto del área de la foto se computa de una manera o de otra. Toda la red funciona de manera que hay una especie de caja negra en la cual los números iniciales que representan la tonalidad de los pixeles entran en esta caja, y a la salida de la caja obtenemos un único número. Ese número se puede interpretar como la probabilidad de que en esa imagen haya un fenotipo X. Podemos decidir a partir de qué número consideraremos que tenemos una aceptable sensibilidad y especificidad para diagnosticar correctamente el fenotipo X. Los creadores de esta red podrían ajustar manualmente cada una de las operaciones y los grupos que se toman, en lo que sería una tarea inabarcable. Imaginemos una mesa gigantesca, con miles de ruletas con las cuales ajustamos cada una de esas operaciones. Girando una ruleta cambiamos qué números sumamos o qué operación hacemos a los números. En una red neuronal puede haber miles de millones de estas ruletas, que además aplican nuevas operaciones a los resultados emitidos por otras ruletas ordenadas en una sucesión de capas (Figura 1). ¿Cómo ajustar este complejo mecanismo si es casi imposible saber el efecto individual que cada ruleta tiene en el resultado final?

El algoritmo de *backpropagation* o de retropropagación[1] es un algoritmo que encuentra el valor al que se tiene que girar cada "ruleta" para que cuando aparezca un fenotipo X, se obtenga un valor final cercano a 1, y cuando no lo haya, sea cercano a 0. Para hacerlo, es necesario "entrenar" a la red neuronal. En el proceso, aparentemente "sencillo", se le da una imagen en la cual sabemos con "certeza" si hay o no una cara con ese fenotipo. Este algoritmo lo que hace en resumidas cuentas es ver hacia dónde tiene que girar cada ruleta para que el número final se acerque al 0 o al 1 en cada caso dependiendo de si la imagen que le hemos dado tiene o no el fenotipo. Alimentando

Cómo citar este artículo: Jiménez Llamas R, Jiménez Alés R. *Deus ex machina*: intentando explicar la inteligencia artificial. Form Act Pediatr Aten Prim. 2024;17(2):60-2.

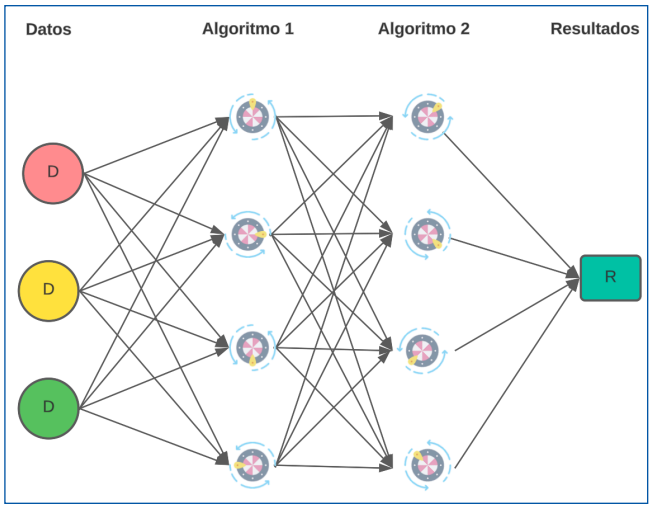

Figura 1. **Esquema de una red neuronal de 2 capas.**

el algoritmo con muchas imágenes (a ser posible, miles o millones, que es lo que hacemos cuando aseguramos a nuestro ordenador que no somos un robot marcando imágenes que contienen un semáforo o un paso de cebra), poco a poco las ruletas se van ajustando "solas". Al final del entrenamiento, se obtiene un algoritmo que es muy bueno en lo que se le ha entrenado, en este caso reconocer un fenotipo X, un semáforo o un paso de cebra. Pero no será bueno reconociendo ninguna otra cosa. Además, al tener millones de ruletas que ajustar, por mucho que se sepa el giro individual de cada ruleta final, eso no implica que se sepa por qué ese es el giro óptimo y no otro para reconocer un determinado fenotipo. De ahí que todo este proceso sea como una "caja negra". Es decir, obtenemos un algoritmo que funciona muy bien, pero no sabemos exactamente ni cómo ni por qué, ya que diciéndole al algoritmo si hace una predicción bien o mal, le hemos enseñado a hacer algo bien o mal, pero no hemos realizado ningún tipo de ajuste "manual". El aprendizaje humano no dista mucho de este tipo de entrenamiento. Los niños al principio funcionan por ensayo y error, llevando a cabo conductas más o menos simples, que progresivamente se van complicando, de modo que reciben refuerzos positivos o negativos, sin que realmente entiendan por qué algo está bien o mal. Las conductas se van ajustando con pequeñas variaciones y matices hasta que el niño recibe un refuerzo positivo óptimo cada vez que repite esa conducta. Al principio, el niño no sabe explicar por qué una conducta desencadena la aparición de un determinado efecto; simplemente sabe que esa conducta dará ese resultado. Conforme crece, el niño acumula más datos y los educadores se vuelven más exigentes para proporcionar el refuerzo. Algo similar ocurre con el entrenamiento de la IA, con la diferencia de que el ser humano puede llegar a interiorizar principios morales y éticos y a razonar por qué determinadas conductas son adecuadas o no, mientras una IA solo podría alcanzar a simular esos principios mediante la acumulación de datos que la retroalimenten. Si los datos con los que se entrena no son los adecuados, las respuestas de la IA no lo serán.

Esto es aplicable a decenas de campos. Un ejemplo de posible aplicación lo tenemos cuando una persona va a un banco a solicitar un préstamo; el banco usa los datos financieros (y no financieros) de una persona para decidir si conceder ese préstamo. Lo hacen mediante algoritmos de IA, porque estos algoritmos son mucho más rápidos que una persona, y además, a simple vista, parecen mucho más objetivos, puesto que simplemente usan "datos", no "sensaciones" ni "intuición", aunque la forma de procesar esos datos sea tan difícil de auditar como lo es explicar en qué consiste una "intuición" humana. Tras la implantación de estos algoritmos se observó que, para las mismas condiciones, se tendía a conceder menos préstamos a mujeres que a hombres y a personas de color respecto a personas blancas. Y es importante notar que esto no es porque la persona que creo el código o las personas del banco quisieran generar esa discriminación, intentando perder potenciales clientes, sino porque los datos de entrenamiento del banco estaban sesgados, como lo está la intuición cuando se ve alterada por "prejuicios". Los bancos usaron sus datos de años anteriores para entrenar a sus algoritmos, pero ¿cuántas mujeres y personas de color recibían un préstamo en los años 50 o 60? Comparado con hombres blancos, eran una minoría, y el algoritmo interpretaba el sesgo anterior, producto de los prejuicios humanos, como algo a reproducir. Es decir, estos algoritmos aprenden de los datos y, si los datos están sesgados de alguna manera, estos sesgos se reproducen e incluso se retroalimentan y aumentan, ya que los datos de nuevos préstamos no concedidos se usan para retroalimentar constantemente el algoritmo. Esto puede llevar a casos de discriminación o de poca precisión en la predicción por no ser los datos de entrenamiento suficientemente representativos de los datos a los que luego se va a aplicar el algoritmo.

En el campo de la Medicina es de especial importancia tener en cuenta la posible existencia de estos sesgos, puesto que de su fiabilidad puede depender la vida de alguien[2]. ¿De dónde sacan los creadores de estos algoritmos miles o millones de imágenes en las cuales se sepa que hay un fenotipo X o no lo hay? ¿Son suficientes? ¿Están representados todos los subtipos? ¿El diagnóstico humano fue correcto? ¿Se incluyeron las posibles variantes étnicas, las distintas edades y sexos?

A veces no queda otra alternativa que entrenar la IA con los datos de que disponemos. Es algo similar a lo que hacemos al

extrapolar los datos obtenidos en ensayos clínicos en adultos a poblaciones como mujeres gestantes o niños. Sin lugar a duda, es mejor eso que nada. Con el entrenamiento de una IA también es mejor entrenarla con datos de una población distinta a la que se aplicará que no entrenarla porque no se dispone datos suficientemente representativos. En estos casos, siempre hay que tener en cuenta las limitaciones y cuáles son los datos con los que ha sido entrenado el algoritmo. Si las compañías privadas de salud usan algoritmos de IA para estimar el riesgo de que una persona padezca una enfermedad y lo hacen a partir de los pacientes que ingresan, sin incluir datos de personas que no llegan a enfermar, es posible que sobreestimen el riesgo de enfermar y, por tanto, el coste de un seguro privado. Del mismo modo, si no se tiene cuidado, pueden aparecer sesgos discriminatorios en cuanto a sexo, raza y edad, en lugar tener en cuenta los factores exclusivamente médicos y de salud relevantes.

Ante este panorama, se ha considerado que las evaluaciones de algoritmos pueden contribuir a paliar dichos efectos, con la detección de aspectos problemáticos como la discriminación de determinadas poblaciones, la distorsión de la realidad o la explotación de información personal. Pero ¿cómo se pueden evaluar los algoritmos para detectar los potenciales problemas que contienen y/o que se derivan de su uso, así como contribuir a su mitigación?

Digital Future Society es una iniciativa transnacional sin ánimo de lucro que conecta a responsables políticos, organizaciones cívicas, expertos académicos y empresarios para explorar, experimentar y explicar cómo se pueden diseñar, usar y gobernar las tecnologías a fin de crear las condiciones adecuadas para una sociedad más inclusiva y equitativa. Su objetivo es ayudar a los responsables políticos a identificar, comprender y priorizar los desafíos y las oportunidades fundamentales, ahora y en los próximos diez años, en relación con temas clave que incluyen la innovación pública, la confianza digital y el crecimiento equitativo. Esta iniciativa ha emitido muy recientemente una serie de 6 recomendaciones para mejorar los procesos de evaluación de algoritmos[3], en un documento que invitamos a leer y sobre el que reflexionar.

BIBLIOGRAFÍA

1. Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G. Backpropagation and the brain. Nat Rev Neurosci. 2020;21(6):335-46.

2. Jiménez-Alés R. Inteligencia artificial. Desafíos y preocupaciones. Rev Pediatr Aten Primaria 2023;25:205-10.

3. Hacia un uso responsable de los algoritmos: métodos y herramientas para su auditoría y evaluación. En: Digital Future Society [en línea] [consultado el 03/05/2024]. https://digitalfuturesociety.com/es/report/towards_accountable_algorithms/