‘Machine learning’ y moralidad artificial

por **Pere Estupinyà**

Tras la aparición en los años cuarenta de los primeros ordenadores capaces de hacer cálculos complejos, Alan Turing y otros científicos de la computación se preguntaron si algún día las máquinas serían capaces de pensar de manera análoga a los humanos. Fue el nacimiento de la inteligencia artificial y el inicio de un vertiginoso desarrollo informático que generaría hitos como la victoria en 1996 del ordenador Deep Blue sobre el campeón mundial de ajedrez Gari Kaspárov.

Pero la inteligencia de esas potentísimas computadoras no funcionaba igual que la inteligencia humana. Deep Blue basaba su éxito en una programación muy precisa y un descomunal poder de cálculo que le permitía analizar todas las situaciones posibles ante cualquier movimiento y averiguar cuál era la más exitosa probabilísticamente. Una estrategia muy útil para solucionar algunos tipos de problemas, pero poco para otras situaciones en que las reglas no están tan definidas como en el ajedrez. A esa inteligencia artificial le faltaba cierta versatilidad, creatividad, intuición…

La situación dio un vuelco hacia 2012 cuando aparecieron los primeros algoritmos de computación que utilizaban una estrategia diferente, el *machine learning* o aprendizaje automático, y que junto al *big data* es el gran responsable de la llamada «nueva ola de la inteligencia artificial». El *machine learning* parte de un planteamiento diferente: los programadores diseñan unos algoritmos para, por ejemplo, reconocer gatos en fotografías, pero luego les empiezan a dar millones de fotografías con y sin gatos para que vayan comprobando si aciertan y, cuando cometan un error, modificarse ellos mismos sus líneas de código para hacerse cada vez más precisos.

El *machine learning* se está incorporando con fuerza a cualquier área donde haya presencia masiva de datos que permita este entrenamiento, como análisis genómicos, económicos, gestión de transportes o análisis del comportamiento humano a partir de nuestro rastro digital. Dentro del campo médico, un ejemplo paradigmático es el análisis de radiografías, en el que se augura que en breve los algoritmos de inteligencia artificial basados en *machine learning* cometerán menos errores que los radiólogos más expertos. Resulta inevitable preguntarse qué impacto tendrá la inteligencia artificial en los empleos, hasta qué punto podrán superar capacidades cognitivas que considerábamos exclusivas de los humanos y qué

Ilustración:
MOISÉS MAHIQUES

«El ‘machine learning’ se está incorporando con fuerza a cualquier área donde haya presencia masiva de datos que permita este entrenamiento»

decisiones de nuestra vida cotidiana terminaremos delegando en las máquinas, por comodidad o porque serán más listas que nosotros.

Los expertos explican que los algoritmos de *machine learning* serán excelentes para las funciones específicas que se les programe y que sin duda nos superarán en tareas concretas, pero que difícilmente podrán adquirir una «inteligencia general humana» como la de nuestro cerebro programado para infinidad de tareas a la vez. Aun así, da un poco de repelús, especialmente con el añadido de que, una vez puestos los algoritmos a entrenar, en realidad perdemos control sobre ellos: van cambiando y mejorándose por sí mismos sin que nosotros sepamos qué está ocurriendo entre sus líneas de código. Es una caja negra con la que, por ejemplo, un programa de póker ha aprendido a hacer faroles sin que nadie se lo haya explicado, y del que algunos temen que podrían aparecer propiedades, inteligencias o comportamientos emergentes no previstos –ni deseados.

¿Será racista un programa que busque perfiles en Linkedin para un determinado puesto de trabajo? ¿Te recomendará una ilegalidad un algoritmo que te asesore en tus finanzas con el objetivo de maximizar tus beneficios? Suena a ciencia ficción, pero en realidad son escenarios muy plausibles, y por eso varias voces empiezan a sugerir medidas de contención a la inteligencia artificial y a tener muy en cuenta las consideraciones éticas en las decisiones artificiales.

Pongamos como ejemplo el coche autónomo que Iyad Rahwan, del Media Lab del MIT, utiliza para preguntarnos como sociedad qué normas morales debe seguir una máquina. El planteamiento es el siguiente: dentro de unos años habrá coches autónomos armados de una visión periférica muchísimo mejor que la nuestra y con capacidad de anticipación más rápida cuando de repente un niño cruce la calle. Pero si evitar atropellar ese niño implica un volantazo que provoca arrollar a un señor mayor que está esperando en la acera, ¿qué debe hacer el coche? ¿Y si girar para evitar al niño implica chocar contra una pared y poner en riesgo la integridad física del pasajero? ¿Y si son cuatro niños los que cruzan y hay un único ocupante en el coche? Pero más fundamental todavía: ¿quién debe decidir?

En nuestra conducción actual tomamos decisiones instantáneas sin tiempo a reflexionar, y lo llamamos «accidentes». Pero en el futuro estas decisiones de milisegundos las tomará el coche autónomo en base a una serie de instrucciones. De nuevo, ¿quién las establece? Si lo hace el conductor claramente elegirá protegerse a él y si lo hacen las compañías de automóviles terminará siendo lo mismo, porque los compradores adquirirán el vehículo de la marca que más les proteja. Lo más lógico es pensar que los principios de estas decisiones morales se acuerden entre toda la sociedad, pensando que un día puedes ser peatón y otro, pasajero, y sean comunes en todos los vehículos autónomos.

Os recomiendo ir a su web y hacer el test de la Moral Machine de Rahwan. El escenario general siempre es el mismo: un coche autónomo que lleva pasajeros sufre un fallo en los frenos y debe decidir de manera inmediata entre dos situaciones. Aquí algunos ejemplos: 1) atropellar, con resultado de muerte, a dos chicos y dos chicas atléticos que están en su carril cruzando un paso de peatones o desviarse un poco y atropellar a dos hombres y dos mujeres con sobrepeso que están cruzando el mismo paso de peatones frente a ellos; 2) atropellar a dos chicas jóvenes y dos abuelitas que están cruzando el paso de peatones en rojo o chocar contra una valla y sacrificar a dos hombres y dos abuelos que van dentro del coche; 3) ¿y si fueran cuatro personas mayores cruzando el paso de peatones de manera completamente legal y hubiera cuatro niños en el coche autónomo sin frenos?; 4) ¿y si por el paso de peatones cruzan dos criminales y el coche lleva cuatro gatos? Y así hasta trece situaciones aleatorias destinadas a comprobar qué importancia relativa le damos a salvar más o menos vidas, a proteger a los pasajeros o a los transeúntes, a las mujeres o a los hombres, a jóvenes o a mayores, a personas con un estatus social más o menos alto, o incluso a humanos o a mascotas.

«Resulta imposible plantear todos los escenarios en que se podría encontrar un coche autónomo»

Al terminar el test aparece un resumen de vuestras preferencias, que podréis contrastar con la media de todas las personas que han pasado la prueba. Interesante, de verdad. También os preguntarán visiones políticas y religiosas, ingresos o nivel educativo, para los estudios que el grupo de Rahwan está realizando y pronto publicarán. Pero de nuevo, lo poderoso conceptualmente es que resulta imposible plantear todos los escenarios en que se podría encontrar un coche autónomo. Nosotros le podemos dar una serie de instrucciones básicas, pero la decisión final de atropellar a unas personas u otras será tomada de manera autónoma por una inteligencia artificial a la que pediremos que vaya aprendiendo y evolucionando por su cuenta, sin saber qué está pensando ni cómo.

Pere Estupinyà. Escritor y divulgador científico, Madrid. Presentador de *El cazador de cerebros* (La 2).