[05] ¿EL COMPORTAMIENTO Y LA MANERA DE ACTUAR DE LOS SISTEMAS DE INTELIGENCIA ARTIFICIAL, ES FÁCILMENTE EXPLICABLE?

Cuando los sistemas de Inteligencia Artificial basados en aprendizaje profundo (IA) aciertan y dan el resultado esperado, lamentablemente no podemos saber por qué han funcionado bien. Pero tampoco sabemos por qué fallan cuando se equivocan. De hecho, es algo que no saben los usuarios pero que tampoco pueden saber los diseñadores de estos sistemas de IA.[1] Es el denominado "problema de la caja negra", que hace que sea prácticamente imposible explicar las decisiones que toman estos sistemas. Ésta es una de las diferencias esenciales con otros artilugios que usamos a diario. Si una lámpara no se enciende, sabemos que puede ser un problema de la bombilla, del cable o del interruptor. Y si un coche deja de funcionar, el mecánico sabrá encontrar fácilmente la causa, sabiendo explicar el motivo y pudiendo proceder a su reparación. Poder explicar el por qué de un problema es el primer paso hacia su resolución.

Pero cuando los sistemas de IA tienen un comportamiento erróneo e inesperado, no hay forma de entender qué ha pasado. Por ello, decimos que los sistemas actuales de Inteligencia Artificial no son explicables. En los casos en que no funcionan, nadie puede explicar por qué han fallado. Es uno de sus graves inconvenientes. Son cajas negras.

Este comportamiento no explicable es consecuencia de la extraordinaria complejidad de las redes neuronales que conforman estos sistemas, de su carácter heurístico,[2] y de la ingente cantidad de parámetros que los gobiernan. Pero también es consecuencia de la multiplicidad de sesgos que con toda probabilidad contenían los datos de aprendizaje, y de su carácter inestable. Como resultado de estos sesgos y de la enormidad de implicaciones que acaban teniendo en el proceso de ajuste de los parámetros de la red, los sistemas de IA es probable que tomen "decisiones" incorrectas[3] a la vez que extrañas y por ello inexplicables. Por otra parte, su bien demostrada inestabilidad acaba agravando este carácter no explicable, porque cambios muy sutiles en la información que el sistema debe clasificar pueden llevar, inexplicablemente, a decisiones extraordinariamente distantes. Hay muchos ejemplos que han mostrado esta inestabilidad, por ejemplo, en los sistemas de clasificación y detección a partir de imágenes. Y de hecho, la modificación de unos pocos píxeles en la imagen de entrada, un cambio que es totalmente imperceptible para el ojo humano, es suficiente para inestabilizar el sistema, haciendo que salte de una clasificación a otra.[4] Un típico ejemplo muestra una fotografía de ImageNet con un autobús escolar; cuando se distorsiona de manera inapreciable, el sistema de IA sorprendentemente la clasificó como un avestruz. Son comportamientos inesperados y por ello inexplicables. En este contexto, parte de los actuales investigadores en

el campo de la IA están estudiando maneras que permitan que el propio sistema, además de dar su respuesta al problema planteado, nos dé información que explique el por qué de dicha respuesta. Todo ello tendrá, no obstante, nuevas limitaciones.[5]

Esta falta de explicabilidad está relacionada com la opacidad de las "cajas negras" y con no poder detectar cuales han sido los fallos internos que han llevado a determinados resultados erróneos. Todo ello impide repararlos e imposibilita que en el futuro podamos evitar errores similares.

Notas:

1. Ramón López de Mántaras (2020), "El traje nuevo de la inteligencia artificial", Investigación y ciencia: https://www.investigacionyciencia.es/revistas/investigacion-y-ciencia/una-nueva-era-para-el-alzhimer-803/el-traje-nuevo-de-la-inteligencia-artificial-18746 - Ramon López de Mántaras es fundador y exdirector del Instituto de Investigación en Inteligencia Artificial del CSIC, en Barcelona. Como explica López de Mántaras, "las personas tampoco podemos explicar siempre nuestras decisiones. Sin embargo, hay una diferencia fundamental: los humanos tendemos a confiar unos en otros porque creemos que los mecanismos de pensamiento de los demás son similares a los nuestros. Es lo que los psicólogos llaman tener una «teoría de la mente» sobre los demás. No obstante, ninguno de nosotros tiene una teoría de la mente sobre ninguna máquina, ni desde luego ninguna máquina la tiene sobre nosotros. Por ello, resulta perfectamente razonable exigir más explicaciones a una máquina que a una persona".
2. Véanse las respuestas a las preguntas 2 y 4.
3. Pere Brunet (2020), "El negocio de las armas que van contra la ética y las personas", El Salto Diario, enero de 2020: https://www.elsaltodiario.com/industria-armamentistica/negocio-armas-contra-etica-personas
4. Jiawei Su, Danilo Vasconcellos Vargas, Kouichi Sakurai (2019), "One Pixel Attack for Fooling Deep Neural Networks", IEEE Transactions on Evolutionary Computation, Vol. 23: https://arxiv.org/pdf/1710.08864.pdf
5. Dado que la fiabilidad de los sistemas de IA es siempre limitada, sus "explicaciones" tendrán una cierta probabilidad de ser erróneas. Véase la respuesta a la pregunta 3.