[03] ¿LOS ALGORITMOS DE INTELIGENCIA ARTIFICIAL, SON SEGUROS Y FIABLES?

La respuesta, según Virginia Eubanks y otros muchos autores, es categórica: no lo son.[1]

Podríamos argumentar con razón que pocas cosas, en este mundo, son realmente seguras y fiables. Nosotros mismos somos vulnerables e imprevisibles, y todo lo que vemos, creamos y construimos está sujeto a multitud de imponderables que pueden acabar descontrolando su comportamiento.

Pero, en relación con los algoritmos de IA, constatamos dos hechos: estos sistemas son mucho menos fiables de lo que nos explican, y tanto sus desarrolladores como sus panegiristas y ensalzadores suelen tender a esconder fallos y errores, dando una visión sesgada de su grado de fiabilidad. Porque el nivel de fallos y errores de los algoritmos y sistemas de inteligencia artificial es muy superior al de los algoritmos clásicos,[2] deterministas y no basados en Big Data, aquellos que nos ayudan en nuestros desplazamientos (como el sistema GPS), los que calculan la cuenta de nuestra compra en el supermercado, o los que nos permiten interaccionar durante las video llamadas, por citar solo algunos ejemplos.

No existe ningún sistema de IA capaz de contextualizar y de hacer el tipo de inferencias básicas que incluso un niño realiza sin esfuerzo.[3] Y de hecho, los sistemas de reconocimiento de imágenes basados en IA se han demostrado altamente inestables. Ciertos cambios que son imperceptibles para los ojos humanos pueden hacer que un sistema de IA deje de reconocer una imagen como la de un león, pasándola a clasificar como la de una biblioteca, por ejemplo.[4] Y es por ello que un buen número de activistas se marca la cara con algunas líneas de pintura.[5] Con ello, consiguen que muchos sistemas de reconocimiento facial con IA se descontrolen y no puedan reconocerles.

La fiabilidad de los algoritmos de IA depende de la aplicación y de la complejidad de la salida deseada (no es lo mismo diseñar un sistema para clasificar imágenes de productos envasados en dos categorías correcto/defectuoso que construir un sistema que deba inferir caras de personas a partir de información sobre las mismas, por ejemplo), del diseño estructural de la red neuronal, de si el aprendizaje continúa o no durante el uso de la red, del tamaño de la propia red... y de los datos utilizados para el entrenamiento. En este caso, la fiabilidad de los resultados es función del número de datos y los posibles sesgos inherentes a los mismos, sesgos que terminan siempre perjudicando a las minorías raciales y de género y a las personas más vulnerables.[6] Los datos que se suministran a los sistemas de IA para su aprendizaje están habitualmente sesgados, heredando los prejuicios de aquellas personas que han intervenido en los procesos y en el negocio de los datos. En consecuencia, los sistemas de IA acaban reproduciendo estos sesgos y reduciendo su grado de fiabilidad. Pero además, debido a su estructura masivamente heurística y a un

proceso de aprendizaje que es necesariamente subóptimo, adolecen de una fiabilidad que es intrínsicamente limitada, algo que es inherente a su estructura y que implica una probabilidad de error no despreciable.

En ciertas aplicaciones críticas como las de diagnóstico en medicina, y ante errores del orden del 12%,[7] los expertos entienden que la intervención de los expertos en la toma de decisiones es imprescindible y que hay que incorporar evaluaciones clínicas con pacientes durante los necesarios procesos de validación previa.[8] Porque, gracias a ONGs como Big Brother,[9] sabemos de la baja fiabilidad de los sistemas de reconocimiento facial que se utilizaron durante los carnavales de Candem: sólo un 5% de las identificaciones de criminales hechas a través del sistema de IA fueron correctas, con un error promedio del 95%.[10]

Todos los nuevos sistemas que utilizamos deben someterse a pruebas y procesos de certificación para determinar su fiabilidad y seguridad. Deberíamos pedir lo mismo con los sistemas de IA y sobretodo con los basados en datos y deberíamos disponer de información cuantitativa sobre su grado de fiabilidad. Pero, además de disponer de metodologías de verificación y validación adecuadas, debemos exigir la creación de agencias independientes de certificación que validen todas las nuevas aplicaciones de IA antes de que se utilicen de forma generalizada.[11]

Notas:

1. Virginia Eubanks (2018), "Automating Inequality", St. Martin's Press: https://virginia-eubanks.com/
2. La definición de algoritmo que proporciona Richard Dawkins en su libro "Escalando el monte improbable" es especialmente interesante: es una manera muy adecuada de resumir el conocimiento que tenemos sobre cualquier conjunto de reglas". Los algoritmos de los sistemas GPS o de las cajas de cobro en los supermercados siguen paso a paso conjuntos de reglas precisas y no ambiguas que garantizan resultados correctos.
3. Ramón López de Mántaras (2020), "El traje nuevo de la inteligencia artificial", Investigación y ciencia, Julio de 2020: 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.
4. Ver: Nguyen, Anh (2015), "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images", IEEE CVPR 2015: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/ - Ver también; Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard (2016), "DeepFool: a simple and accurate method to fool deep neural networks", Proceedings of the IEEE CVPR, https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Moosavi-Dezfooli_DeepFool_A_Simple_CVPR_2016_paper.pdf
5. Ver por ejemplo: https://i-d.vice.com/en_uk/article/jge5jg/dazzle-club-surveillance-activists-makeup-marches-london-interview
6. Virginia Eubanks: Una respuesta al DHS del condado de Allegheny, sobre la herramienta de evaluación familiar de Allegheny (AFST): "Creo que el sistema es injusto y discriminatorio. Además, la declaración del condado del 31 de enero sugiere que para el 55% de las familias, la mayoría, la recepción de servicios públicos de hecho aumenta su puntuación AFST, dejándolos desproporcionadamente vulnerables en las investigaciones sobre bienestar infantil": https://virginia-eubanks.com/2018/02/16/a-response-to-allegheny-county-dhs/
7. Xinyuan Zhang, Shiqi Wang, Jie Liu & Cui Tao (2018), "Towards improving diagnosis of skin diseases by combining deep neural network and human knowledge", BMC Medical Informatics and Decision Making: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0631-9?optIn=true
8. Emma Beede (2020), "Healthcare AI systems that put people at the center": https://www.blog.google/technology/health/healthcare-ai-systems-put-people-center/
9. Ver: https://bigbrotherwatch.org.uk/campaigns/stop-facial-recognition/#facial-recognition-uk
10. Roser Martínez Quirante y Joaquín Rodríguez (2020), "El costat fosc de la intel·ligència artificial - El cas dels sistemes d'armament letal autònom o els Killer Robots", Revista Idees: https://revistaidees.cat/el-costat-fosc-de-la-intelligencia-artificial/
11. Luca Steels y Ramón Lólez de Mántaras (2018), "The Barcelona declaration for the proper development and usage of artificial intelligence in Europe": https://content.iospress.com/articles/ai-communications/aic180607