AUDITOR ÉTICO, UNA NUEVA FIGURA EN LA ERA DE LA IA

Inteligencia artificial y la búsqueda de una sociedad más justa

DAVID CASACUBERTA

Usamos inteligencia artificial para tomar decisiones que afectan a nuestras vidas diarias. Es necesario asegurarnos no solo de que esas decisiones se basen en datos fiables sino de que sean justas. En este artículo exploramos de qué manera podemos usar inteligencia artificial para asegurar la justicia en las decisiones tomadas por bancos, servicios sociales básicos, universidades o escuelas.

Ethical auditor, a new figure in the era of AI

ARTIFICIAL INTELLIGENCE AND THE SEARCH FOR A MORE JUST SOCIETY

We use artificial intelligence to make decisions that affect our daily lives. It is necessary to ensure not only that these decisions are based on reliable data but that they are fair. In this article we explore how we can use artificial intelligence to ensure justice in decisions made by banks, basic social services, universities or schools.

Keywords: *algorithm, ethics, artificial intelligence, machine learning, algorithmic injustice.*

Palabras clave: algoritmo, ética, inteligencia artificial, *machine learning*, injusticia algorítmica.

ILUSTRACIÓN: **PALOMA LIRANZO**

Los medios de comunicación, cuando hablan de inteligencia artificial, suelen saltar del optimismo al pesimismo postapocalíptico. Esas supuestas "máquinas que aprenden cualquier cosa sin intervención humana" y que "están a punto de alcanzar la autoconsciencia" nos traerán el martes por la mañana increíbles logros; el jueves por la tarde se desvelará en una tertulia televisada cómo hay tantos usos positivos como negativos del *machine learning* (aprendizaje robótico) y el suplemento dominical de ciencia nos revelará cómo la "IA fuerte" volverá la humanidad obsoleta.

Y llegamos a la distopía cuando hablamos del rol que la inteligencia artificial podría asumir o está asumiendo en el ámbito político y social. En ese contexto, los algoritmos son solamente fuente de injusticia, manipulación y vigilancia, y nada ni nadie puede detener la ola de fascismo y paternalismo digitales.

Quiero aquí ofrecer un término medio y argumentar que, aunque los peligros de digitalizar la comunicación política, la sanidad pública o las decisiones judiciales son reales y plausibles, las oportunidades también lo son y los algoritmos de aprendizaje automático pueden convertirse en herramientas en pro de la justicia social y política.

Se habla mucho últimamente de la necesidad de una ética para la inteligencia artificial. Desgraciadamente, la mayor parte de la discusión se centra en problemas lejanos e irreales. Sería ciertamente irresponsable desarrollar una inteligencia artificial "fuerte", una similar o incluso superior a la humana sin añadir una serie de controles éticos muy claros. Pero es aún más irresponsable llenar libros y pantallas sobre esa futurible singularidad, cuando hoy en día ya hay problemas éticos asociados al uso de algoritmos de aprendizaje automático para tomar decisiones que parecen generar mucho menos interés.

Y no me refiero aquí a la cuestión tan debatida de a quién debería matar un coche autónomo cuando sea necesario escoger entre dos posibles trayectorias de colisión: ¿evitar un choque frontal que mataría al conductor, aunque ello implique atropellar a cinco personas que esperan al autobús? ¿Llevarse por delante a la anciana o al bebé? La literatura es ahora mismo interminable y una rápida búsqueda de "ética" y "coche autónomo" nos dará en un momento cientos de referencias.

El dilema del tranvía aplicado a los coches autónomos es el ejemplo perfecto de la solución en busca de un problema. En una peculiar versión del lecho de Procusto[1], algunos ingenieros y economistas quieren convertir los problemas éticos en procesos de asignación de probabilidades a partir del *feedback*[2] de usuarios sobre sus preferencias. De ahí su interés en convertir cualquier problema ético en una versión del dilema del tranvía. Para bien y para mal, la ética es mucho más complicada.

De la misma forma que prácticamente nadie en el año 2005, cuando se hablaba de la web 2.0, se imaginaba que las redes sociales se convertirían en instrumentos de manipulación política, pienso que la mayoría de los problemas éticos asociados a la inteligencia artificial no son predecibles y se irán definiendo según vayamos dando uso a esas nuevas tecnologías. Considero así más útil discutir problemas que ya han surgido y buscarles solución y esperar a que los nuevos problemas empiecen a cobrar sentido antes de desarrollar metodologías para tratar con ellos.

La mayoría de los problemas éticos asociados a la inteligencia artificial no son predecibles y se irán definiendo con el uso

Injusticia algorítmica

Uno de los problemas más claros y acuciantes ahora mismo es el de la injusticia algorítmica.

Actualmente se usan algoritmos de aprendizaje automático para tomar todo tipo de decisiones en el ámbito social: usamos algoritmos para decidir si una persona será capaz o no de devolver un crédito al banco; quién es el mejor candidato para ocupar un cargo en una empresa; qué estudiantes sacarán más provecho de una carrera universitaria con una gran demanda; o si una persona detenida debería ir a prisión preventiva o salir bajo fianza, entre otros muchos temas. Para desarrollar esos algoritmos recopilamos datos de decisiones anteriores, indicamos

1 Es una expresión proverbial que se refiere a quienes pretenden acomodar siempre la realidad a sus intereses o su visión de las cosas. Más información: https://es.wikipedia.org/wiki/El_lecho_de_Procusto

2 La palabra inglesa *feedback* equivale en español a reacciones, comentarios, opiniones, impresiones, sensaciones, e incluso a retorno, respuestas o sugerencias.

Es mucho más fácil descubrir una decisión sesgada en un algoritmo que en una persona

cuándo las decisiones fueron correctas y confiamos en que ese algoritmo será capaz de capturar las regularidades y correlaciones relevantes que permitan tomar decisiones como mínimo tan adecuadas como las personas que lo hicieron anteriormente, si no mejores.

Este uso de la inteligencia artificial, si no se es cuidadoso, puede llevarnos a tomar decisiones injustas.

La base de datos perfecta no existe. Siempre hay errores de transcripción, correlaciones espurias o sesgos en la toma decisiones. Un ejemplo bien significativo es COMPAS o *Correctional Offender Management Profiling for Alternative Sanctions*, un programa en uso en Estados Unidos para ayudar a los jueces a automatizar la decisión de si una persona detenida ha de ir a prisión preventiva o puede salir bajo fianza. La revista de investigación periodística *ProPublica* analizó COMPAS en detalle y declaró que el programa tendía a asignar una probabilidad mucho mayor de poder salir bajo fianza a las personas de raza blanca que a las de raza negra, asignando así, por ejemplo, una probabilidad de un riesgo bajo de reincidencia a un hombre de raza blanca que ya había cometido varios robos a mano armada y, en cambio, asignarle la máxima probabilidad de reincidencia a un joven negro que solo tenía en su haber delitos menores. Las razones de este posible sesgo son variadas. Pero si la base de datos captura realmente los criterios que el sistema judicial americano usa para decidir sobre la libertad bajo fianza, es forzoso admitir que se trata de un sistema racista.

Imaginemos ahora que estos sistemas automáticos se implantan de forma generalizada en juzgados y que, al estilo de AlphaGo[3], las decisiones tomadas por los algoritmos se usan para seguir entrenando sistemas de inteligencia artificial de aprendizaje automático. El resultado sería como mínimo el mantenimiento del sesgo y, muy probablemente, su extensión y ampliación. Si el algoritmo ha descubierto un criterio sencillo, aunque sesgado, de tomar una decisión, y ese criterio se ajusta bien a los resultados esperados, que lo explote para encontrar soluciones sencillas y efectivas es prácticamente inevitable.

Observemos que no hay soluciones fáciles, como borrar de la ficha la información sobre la raza de la persona. Si los jueces que tomaron las decisiones previas que el algoritmo usa para aprender utilizaron el dato de origen étnico como criterio —ya fuera consciente o inconscientemente—, el programa puede reconstruir la categoría "etnicidad" a partir de un *proxy* como el barrio en el que vive esa persona más su nombre y apellido.

Paradójicamente, situaciones como la generada por COMPAS pueden ser un poderoso instrumento para hacer que las decisiones en temas sociales clave sean más justas. Después de todo, es mucho más fácil descubrir una decisión sesgada en un algoritmo que en una persona. Y, lo más importante, es harto más fácil cambiar un algoritmo para evitar injusticias que la mentalidad de una jueza y un juez. En nuestro horizonte cercano tenemos así una nueva profesión, el auditor ético, que examina sistemas de toma de decisiones para establecer si los resultados son realmente justos.

Pero ello no nos debería llevar a utopías al estilo Kurzweil[4] donde cedemos nuestras decisiones más importantes a la supuesta sabiduría imparcial de la máquina. Como ya explicitó el filósofo von Wright en su célebre distinción entre explicación y comprensión, el mundo de lo social se sostiene desde la argumentación de razones y no meramente la exposición de causas. Para establecer si una decisión es justa o injusta, recurrir a regularidades estadísticas es insuficiente. Necesitamos explicitar las razones de la decisión. Es vital asegurarnos de que hay humanos al final de la cadena de decisiones y que pueden detallar criterios racionales de por qué se ha tomado esa decisión y no otra.

Debemos ver así a los algoritmos de aprendizaje automático como herramientas útiles para establecer regularidades y descubrir sesgos. Y entender que la decisión de suplantar mentes humanas por máquinas no es tecnológica, sino política.

Frente al hecho de que ya hay algoritmos de aprendizaje automático capaces de detectar tumores a partir de una radiografía con una fiabilidad mayor que un humano, hay dos posibles vías: despedir a los radiólogos y sustituirlos por máquinas consiguiendo así un dudoso recorte de gastos en la Seguridad Social o equipar a nuestros radiólogos con esos algoritmos para que así puedan dedicar su tiempo a investigar y mejorar técnicas de detección contra el cáncer, liberándolos de la tarea rutinaria de revisar radiografías en busca de tumores. Se trata de una decisión política, basada en una comprensión correcta de las ventajas y defectos de la inteligencia artificial que ponga la calidad de vida y la justicia por encima de otras consideraciones.

3 AlphaGo es un programa informático de inteligencia artificial desarrollado por Google DeepMind para jugar al juego de mesa Go. En octubre de 2015 se convirtió en la primera máquina de Go en ganar a un jugador profesional de Go sin emplear piedras de handicap en un tablero de 19x19. Más información: https://es.wikipedia.org/wiki/AlphaGo

4 Raymond Kurzweil es un inventor estadounidense, además de músico, empresario, escritor y científico especializado en Ciencias de la Computación e Inteligencia artificial. Desde 2012 es director de Ingeniería en Google. Más información en: https://www.kurzweilai.net

Bibliografía

Latorre, J.I. (2019): *Una ética para las máquinas*. Barcelona, Ariel.

O'Neil, C. (2019): *Armas de Destrucción Matemática*. Madrid, Capitán Swing.

Wallach, W. y Allen, C. (2010): *Moral Machines: Teaching Robots Right from Wrong*. Oxford y Nueva York, Oxford University Press.