Ética de la inteligencia artificial

Colección Teorema
Serie mayor

Mark Coeckelbergh

Ética de la inteligencia artificial

Traducción de Lucas Álvarez Canga

CÁTEDRA
TEOREMA

Título original de la obra:
AI Ethics

1.ª edición, 2021

Juan Ignacio Luca de Tena, 15. 28027 Madrid
Depósito legal: M. 27.435-2020
ISBN: 978-84-376-4212-3
Printed in Spain

Índice

Para Arno

Agradecimientos

Este libro no solo se basa en mi propio trabajo en esta materia, sino que también refleja el conocimiento y la experiencia de todo el campo de la ética en la inteligencia artificial. Sería imposible hacer una lista con toda la gente con la que he discutido y aprendido a lo largo de los años, pero en las relevantes y cada vez más numerosas comunidades que conozco se incluyen investigadores en inteligencia artificial [IA] como Joanna Bryson y Luc Steels, compañeros filósofos de la tecnología como Shannon Vallor y Luciano Floridi, académicos que trabajan en innovación responsable en Países Bajos y Reino Unido como Bernd Stahl en la Universidad De Montfort, gente que conocí en Viena como Robert Trappl, Sarah Spiekermann y Wolfgang (Bill) Price, y mis compañeros de los órganos consultivos orientados a la política pertenecientes al Grupo de Expertos de Alto Nivel en IA (Comisión Europea) y al Consejo austriaco de robótica e inteligencia artificial, por ejemplo Raja Chatila, Virginia Dignum, Jeroen van den Hoven, Sabine Köszegi y Matthias Scheutz, por nombrar unos pocos. También me gustaría dar afectuosamente las gracias a Zachary Storms por ayudarme con la corrección y la preparación del libro, y a Lena Starkl e Isabel Walter por ayudarme con la búsqueda de bibliografía.

Capítulo 1

Espejito, Espejito

La exageración de la IA y sus miedos: espejito, espejito, ¿quién es el más listo del reino?

Cuando se anunciaron los resultados, los ojos de Lee Sedol se llenaron de lágrimas. AlphaGo, un programa de inteligencia artificial (IA) desarrollado por DeepMind de Google consiguió la victoria por 4 a 1 en el juego de Go. Es marzo de 2016. Dos décadas antes, el gran maestro de ajedrez Garry Kasparov perdió contra la máquina Deep Blue, y ahora un programa de ordenador ganó contra el dieciocho veces campeón del mundo Lee Sedol en un complejo juego al que se creía que solo los humanos podían jugar, utilizando su intuición y su pensamiento estratégico. El ordenador ganó siguiendo no solo las reglas dadas por los programadores, sino empleando también un sistema de aprendizaje automático basado en millones de partidas anteriores de Go y jugando contra sí mismo. En este caso, los programadores prepararon las bases de datos y crearon los algoritmos, pero no podían saber cuáles serían los movimientos que elaboraría el programa. La IA aprende por sí misma. Después de varias jugadas inusuales y sorprendentes, Lee tuvo que rendirse (Borowiec 2016).

Se trata de un logro increíble para la IA, pero también suscita inquietudes. Hay una admiración por la belleza de las jugadas, pero también tristeza, incluso miedo. Existe la esperanza de que IAs más inteligentes puedan incluso ayudarnos a revolucionar los servicios sanitarios

o a encontrar soluciones para todo tipo de problemas sociales, pero también la preocupación de que las máquinas tomen el control. ¿Podrán las máquinas superarnos en inteligencia y controlarnos? ¿Es la IA una mera herramienta, o poco a poco y de forma segura se está convirtiendo en nuestro amo y señor? Estos miedos nos recuerdan las palabras del ordenador HAL en la película de ciencia ficción de Stanley Kubrick *2001: Una odisea en el Espacio,* quien en respuesta a una orden humana, «abre las puertas del hangar», responde: «lo siento Dave, me temo que no puedo hacer eso». Y si no es miedo, entonces puede que sea un sentimiento de tristeza o decepción. Darwin y Freud destronaron nuestras creencias de que éramos excepcionales, nuestros sentimientos de superioridad y nuestras fantasías de control; hoy día, la inteligencia artificial parece asestar otro golpe a la autoimagen de la humanidad. Si una máquina puede hacer esto, ¿qué queda para nosotros? ¿Qué somos? ¿Somos simplemente máquinas? ¿Somos máquinas *inferiores,* con demasiados defectos? ¿Qué será de nosotros? ¿Nos convertiremos en los esclavos de las máquinas? O, peor, ¿en una mera fuente de energía, como en la película *Matrix?*

El verdadero y generalizado impacto de la IA

Pero los avances de la inteligencia artificial no se limitan a los juegos o al mundo de la ciencia ficción. La IA se da ya en la actualidad, y está generalizada, a menudo integrada de forma invisible en nuestras herramientas cotidianas como parte de complejos sistemas tecnológicos (Boddington 2017). Dado el crecimiento exponencial de la potencia de los ordenadores, la disponibilidad de grandes conjuntos de datos debida a las redes sociales y al uso masivo de miles de millones de smartphones y redes móviles de gran velocidad, la IA, y especialmente el aprendizaje automático, ha logrado avances significativos. Este hecho ha permitido a los algoritmos hacerse cargo de muchas de nuestras actividades, incluyendo la planificación, el habla, el reconocimiento facial y la toma de decisiones. Las aplicaciones de la IA se dan en muchos ámbitos, incluyendo transporte, *marketing,* servicios sanitarios, finanzas y aseguradoras, la seguridad y el ámbito militar, ciencia, educación, trabajo de oficina y asistencia personal (por ejemplo, Google Duplex)[1], entreteni-

[1] Véase <https://www.youtube.com/watch?v=D5VN56jQMWM>.

La IA se da ya en la actualidad, y está generalizada, a menudo integrada de forma invisible en nuestras herramientas cotidianas.

miento, artes (recuperación de información musical y composición), la agricultura y, por supuesto, la fabricación.

La IA es creada y utilizada por empresas de tecnología de la información [TI] y de internet. Por ejemplo, Google siempre ha usado la IA para su motor de búsqueda. Facebook usa IA para la publicidad dirigida y el etiquetado de fotos. Microsoft y Apple usan IA para potenciar sus asistentes digitales. Pero la aplicación de la IA abarca mucho más que el sector TI definido en sentido estricto. Por ejemplo, hay muchos planes concretos y experimentos con coches autónomos. Esta tecnología también está basada en la IA.

Los drones usan IA, así como las armas autónomas que pueden matar sin intervención humana. Y la IA ya se ha empleado para tomar decisiones en juzgados. En los Estados Unidos, por ejemplo, el sistema COMPAS se ha utilizado para predecir quién es más probable que vuelva a delinquir. La IA entra también en campos que normalmente consideramos que son más personales o íntimos. Por ejemplo, las máquinas ahora pueden leer nuestras caras: no solo para identificarnos, sino también para interpretar nuestras emociones y recuperar todo tipo de información.

La necesidad de discutir los problemas éticos y sociales

La IA puede tener muchos beneficios. Se puede usar para mejorar los servicios públicos y comerciales. Por ejemplo, el reconocimiento de imágenes es una buena noticia para la medicina: puede ayudar en el diagnóstico de enfermedades como el cáncer o el alzhéimer. Pero las aplicaciones cotidianas de la inteligencia artificial muestran también que las nuevas tecnologías plantean problemas éticos. Daré algunos ejemplos de conflictos éticos relacionados con la IA.

¿Deberían tener los coches autónomos restricciones éticas y, si así fuera, qué tipo de restricciones, y cómo deberían determinarse? Por ejemplo, si un coche autónomo se encuentra en una situación en la que debe escoger entre atropellar a un niño y chocar contra un muro para salvar la vida del niño, pero potencialmente matar a su pasajero, ¿qué debería escoger? Y ¿deberían estar siquiera permitidas las armas letales autónomas? ¿Cuántas decisiones y cuánto de estas decisiones queremos delegar a la IA? Y ¿quién es responsable cuando algo sale mal? En cierto caso, los jueces depositaron más confianza en el algoritmo COMPAS

que en los acuerdos a los que llegaron la defensa y el fiscal[2]. ¿Se confía demasiado en la IA? El algoritmo COMPAS es también enormemente controvertido, puesto que la investigación ha demostrado que los falsos positivos del algoritmo (personas que predijo que iban a volver a delinquir pero que no lo hicieron) se daban desproporcionadamente entre gente de piel negra (Fry 2018). La IA puede, así, reforzar prejuicios y discriminaciones injustas. Problemas similares pueden surgir con algoritmos que recomiendan decisiones sobre solicitudes de préstamos y de empleo. O considérese la llamada policía predictiva: los algoritmos se usan para predecir dónde es probable que ocurran los delitos (por ejemplo, en qué área de una ciudad) y quién puede cometerlos, pero el resultado puede ser que un grupo específico socioeconómico o racial esté señalado desproporcionadamente por la vigilancia policial. La policía predictiva ya se ha usado en los Estados Unidos y, según lo que muestra un informe reciente de AlgorithmWatch (2019), también en Europa[3]. Y la IA basada en la tecnología de reconocimiento facial se utiliza a menudo para la vigilancia y puede violar la privacidad de las personas. También puede predecir de manera más o menos aproximada las preferencias sexuales. No se necesita información del número de teléfono ni datos biométricos. La máquina hace su trabajo a distancia. Con cámaras en la calle y en otros espacios públicos, podemos ser identificados y «leídos», incluso por lo que respecta a nuestro estado de ánimo. Mediante el análisis de nuestros datos, se puede predecir nuestra salud mental y corporal (sin que lo sepamos). Los empresarios pueden usar la tecnología para monitorizar nuestro rendimiento. Y los algoritmos que están activos en las redes sociales pueden propagar discursos de odio o información falsa. Por ejemplo, los *bots* políticos pueden aparentar ser personas reales y publicar contenido político. Un caso conocido es el del *chatbot* de Microsoft en 2016, llamado Tay y diseñado para tener conversaciones lúdicas en Twitter, pero que, cuando se volvió más inteligente, empezó a *tuitear* contenidos racistas. Algunos algoritmos de IA

[2] Véase el caso de Paul Zilly contado por Fry (2018, 71-72). Más detalles en Julia Angwin, Jeff Larson, Surya Mattu y Lauren Kirchner, «Machine Bias», *ProPublica,* 23 de mayo de 2016, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

[3] Por ejemplo, en 2016 una parte de la policía local de Bélgica comenzó a usar *software* de policía predictiva para predecir atracos y robos de vehículos (AlgorithmWatch 2019, 44).

pueden incluso crear vídeos de discursos falsos, como el vídeo que se compuso para inducir falsamente al público a pensar que parecía un discurso dado por Barack Obama[4].

Las intenciones normalmente son buenas. Pero estos problemas éticos son normalmente consecuencias no deseadas de la tecnología: la mayoría de estos efectos, como los prejuicios o los discursos de odio, no eran algo que pretendieran hacer los que desarrollan o son usuarios de la tecnología. Además, una cuestión crítica que debe siempre preguntarse es: ¿mejora para quién? ¿El gobierno o los ciudadanos? ¿La policía o aquellos que la policía tiene en su punto de mira? ¿El vendedor o el cliente? ¿Los jueces o los acusados? Las cuestiones relacionadas con el poder entran en acción, por ejemplo, cuando la tecnología se forma solo para algunas pocas mega corporaciones (Nemitz 2018). ¿Quién determina el futuro de la IA?

Esta cuestión subraya el significado social y político de la IA. La ética de la IA se ocupa del cambio tecnológico y su impacto en las vidas de los individuos, pero también de las transformaciones que se producen en la sociedad y en la economía. Las cuestiones que tienen que ver con el prejuicio y la discriminación ya indican que la IA resulta relevante socialmente. Pero también está cambiando la economía y, por tanto, quizás la estructura de nuestras sociedades. De acuerdo con Brynjolfsson y McAfee (2014), hemos entrado en una segunda edad de la máquina en la que las máquinas no son únicamente un complemento de los humanos, como en la Revolución Industrial, sino que también los sustituyen. Ya que profesiones y trabajos de todo tipo se verán afectados por la IA, se ha predicho que nuestra sociedad cambiará drásticamente a medida que pasen al mundo real ciertas tecnologías antaño descritas en la ciencia ficción (McAfee y Brynjolfsson 2017). ¿Cuál es el futuro del trabajo? ¿Qué tipo de vidas tendremos cuando la IA asuma puestos de trabajo? ¿Y quién somos «nosotros»? ¿Quién saldrá ganando con esta transformación, y quién perdiendo?

[4] BuzzFeedVideo, «You Won't Believe What Obama Says in This Video!», <https://www.youtube.com/watch?v=cQ54GDm1eL0&fbclid=IwAR1oD0AlopEZa00XHo3WNcey_qNnNqTsvHN_aZsNb0d2t9cmsDbm9oCfX8A>.

La ética de la IA
se ocupa del cambio
tecnológico y su
impacto en las vidas
de los individuos,
pero también de las
transformaciones
en la sociedad
y en la economía.

Ciertos avances espectaculares han provocado un gran problema de exageración en torno a la IA. Y esta ya se está utilizando en un amplio grupo de áreas de conocimiento y de prácticas humanas. Lo primero ha hecho surgir especulaciones disparatadas sobre el futuro tecnológico e interesantes discusiones filosóficas sobre lo que significa ser humano. Lo segundo ha creado una sensación de urgencia en parte de los estudiosos de la ética y los políticos para asegurar que esta tecnología nos beneficia en lugar de propiciar desafíos insuperables para determinados individuos y sociedades. Estas últimas preocupaciones son más prácticas e inmediatas.

Este libro, escrito por un filósofo académico que tiene también experiencia en asesoramiento para el establecimiento de algunas políticas, lidia con ambos aspectos: trata de la ética en tanto que relacionada con todas estas cuestiones. Busca dar al lector una buena visión general de los problemas éticos que surgen en conexión con la IA entendida de forma amplia, desde las influyentes narrativas sobre el futuro de la IA y las cuestiones filosóficas sobre la naturaleza y el futuro de lo humano, hasta las preocupaciones éticas sobre la responsabilidad, el prejuicio y el modo de lidiar con cuestiones prácticas del mundo real que surgen de la tecnología mediante la aplicación de políticas (preferiblemente antes de que sea demasiado tarde).

¿Qué pasará cuando sea «demasiado tarde»? Algunos escenarios son distópicos y utópicos al mismo tiempo. Comenzaré con algunos sueños y pesadillas sobre el futuro tecnológico, relatos de gran repercusión que, al menos *a primera vista,* parecen relevantes para evaluar los potenciales beneficios y peligros de la inteligencia artificial.