Daimon. Revista Internacional de Filosofía, (en prensa) reseña aceptada para ser publicada en un próximo número de la revista.
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico)

LARA, F., DECKERS, J. (eds.). (2023). *Ethics of Artificial Intelligence*. Cham: Springer (The International Library of Ethics, Law and Technology, vol. 41).
https://link.springer.com/book/10.1007/978-3-031-48135-2

La inteligencia artificial (IA), aunque en sentido estricto no es una novedad del presente siglo, tiene relativamente poco tiempo siendo parte de la cotidianeidad de muchas personas y estando en el debate público. Hoy en día, casi todas las personas utilizan asistentes virtuales como Alexa o Siri; algunas otras tienen acceso a sistemas de realidad virtual; algunos hospitales y bancos utilizan máquinas inteligentes para diagnosticar y otorgar préstamos, respectivamente; los robots cuidadores ya son una realidad; y todos sabemos que se han utilizado armas autónomas en el conflicto de Ucrania. Todo esto es posible gracias a la IA.

La ética de la inteligencia artificial, es decir, la disciplina académica que investiga las implicaciones y controversias éticas del uso, regulación y *naturaleza* misma de la IA, es también relativamente nueva y, sin embargo, muy prolífera. Pero a pesar de la importante cantidad de obras académicas en el área en los últimos años, sus conceptos y alcances no están bien delimitados todavía, lo cual la hace un área de especial interés para los profesionales de la filosofía.

El libro *Ethics of Artificial Intelligence*, coeditado por Francisco Lara (Universidad de Granada) y Jan Deckers (Universidad de Newcastle), es una obra que pretende proporcionar una visión general de los problemas éticos más importantes en torno a la IA, en un lenguaje lo suficientemente accesible para el público general, pero con el rigor necesario para contribuir al debate y la investigación entre los especialistas. Los editores han acertado tanto en la selección de los autores a quienes se comisionó cada capítulo, como en el orden que le dieron a los mismos. El libro está estructurado en tres partes, la primera dedicada a qué tan éticamente pueden funcionar las máquinas y qué estatus moral tienen, la segunda a las controversias éticas específicas de algunas aplicaciones de la IA y la tercera a la regulación de la IA.

El lector, dependiendo de sus intereses, puede leer el libro completo si quiere tener una visión panorámica, alguna de sus tres partes si quiere enfocarse en esa temática general o algún(os) capítulo(s) si requiere de una lectura más detallada y profunda sobre esa problemática particular que a su vez le lleve a otras lecturas y le permita avanzar su investigación. Debido a que cada capítulo explica los conceptos relevantes de la problemática que aborda, no hay, como en otros libros sí, que leer el capítulo previo para entender el siguiente.

Por razones de espacio y de la naturaleza misma del libro (i.e. la diversidad de temas específicos) esta reseña no puede abordar detalladamente cada capítulo. Hacerlo no sería justo ni para el libro como un todo ni para cada capítulo particular. Por esa razón, en lo siguiente me dedicaré a

exponer las características y problemas generales de cada una de sus partes, mencionando brevemente los aspectos que me parecen más importantes de cada capítulo.

Las preguntas fundamentales de la primera parte son: ¿puede la IA discriminar?, ¿sus decisiones son transparentes? y ¿tiene estatus moral?

En el Cap. 1 Jorge Casillas no solo responde afirmativamente a la primera pregunta, sino también que la discriminación ejercida por las máquinas es más grave debido a la mayor cantidad de personas a las que puede afectar. Si bien es cierto que la automatización permite mayor rapidez y menor esfuerzo, un problema en el diseño -intencional o no- de una máquina utilizada, por ejemplo, para otorgar préstamos, significaría una discriminación en masa, por decirlo de algún modo. A mí parecer, el autor descarta muy pronto la posible objeción de que no es la máquina la que discrimina sino los humanos que la diseñan. Parece quedar abierta la pregunta de si la intencionalidad es necesaria para adjudicar culpabilidad, en este caso.

En el Cap. 2 Alberto Fernández aborda lo que se conoce como opacidad de la IA. La opacidad no es otra cosa que la falta de transparencia en los resultados o sugerencias de una máquina que utiliza *Machine Learning*. Lo que sucede aquí es que el usuario desconoce las *razones* o el medio por el cual la máquina llegó a esa conclusión, aunque esta sea muy acertada. Comúnmente, a los humanos no nos gusta eso. Las razones de fondo para una conclusión sirven para justificarla. Al desconocer cómo es que la máquina concluyó tal cosa, parece que dicha conclusión carece de justificación. El autor aborda esta problemática y presenta algunos prospectos para mejorar las explicaciones de estas máquinas.

Finalmente, en el Cap. 3 Joan Llorca Albareda, Paloma García y Francisco Lara abordan el tema del estatus moral de las entidades de IA. Ahora que ya no es tan controversial preguntarse por el estatus moral de otras entidades biológicas como los animales o las plantas, nos enfrentamos al problema del estatus moral de entidades no biológicas como la IA. Los autores analizan si, de acuerdo con sus características, las entidades de IA encajan en algunas concepciones de agencia moral, responsabilidad moral, objeto de consideración moral y de derechos, así como en las dos propuestas del giro relacional de Coeckelbergh y Gunkel. Este tema es de suma importancia ya que, de dar con una respuesta sólida argumentalmente y basada en evidencia, supondría un giro en la forma de ver moralmente a las máquinas y, sobre todo, a la adjudicación de responsabilidad sobre las acciones realizadas por ellas. A este respecto, la temática abordada aquí podría dar un poco más de luz a la pregunta que queda abierta en el primer capítulo.

La segunda parte del libro se enfoca en controversias específicas de algunas aplicaciones de la IA.

Un tema recurrente en esta parte es el de la autonomía. Por ejemplo, en el Cap. 5 Juan Ignacio del Valle, Joan Llorca Albareda y Jon Rueda se preguntan si el uso cada vez más habitual de los asistentes virtuales puede generar dependencia cognitiva en los usuarios y, en última instancia, una obsolescencia humana. Sin embargo, también abordan el tema de si los asistentes virtuales pueden ayudar a mejorarnos como humanos, más específicamente en el ámbito moral. Varias propuestas se han hecho en torno a

esto, pero los autores parecen preferir la del asistente virtual socrático de Lara y Deckers. Los usuarios de este sistema mantendrían una conversación con una especie de Sócrates, el cual, lejos de decirles qué hacer, les llevaría mediante el diálogo a tomar sus propias decisiones. De este modo, el usuario mantendría un papel principal y no secundario, como en otras propuestas.

En el Cap. 6, dedicado a la realidad virtual, Blanca Rodríguez López hace un breve recuento histórico de esta, se pregunta si es real o no, aborda el tema de los avatares y expone algunos beneficios, perjuicios y rarezas de la realidad virtual. En relación con la autonomía y la privacidad, los sistemas de realidad virtual pueden hacer más fácil la manipulación de los usuarios debido a la posible confusión entre realidades y a la información biométrica delicada obtenida por los dispositivos utilizados en ella.

En el Cap. 7 Rafael Cejudo se centra en las deepfakes: contenidos audiovisuales creados por la IA. Dos de los principales problemas relacionados con las deepfakes son su uso para engañar y manipular al público y lo relacionado con la autoría de los contenidos creados por esta. El autor propone una revisión detallada de la normativa del copyright para que incluya lo realizado por la IA.

En torno a los robots cuidadores, María Victoria Martínez López, Gonzalo Díaz Cobacho, Aníbal M. Astobiza y Blanca Rodríguez López hacen una taxonomía de estos distinguiendo sus diferentes usos. En relación con la autonomía, los usuarios cognitivamente disminuidos presentan un mayor problema, pues entre menos capacidad cognitiva tienen, menor es su autonomía y las decisiones sobre su persona son delegadas a otros (en este caso, a los robots cuidadores). Los autores abordan esto muy someramente diciendo que el problema es el mismo en el caso de cuidadores humanos. Sin embargo, las implicaciones sobre la adjudicación de responsabilidad serían diferentes dada la incertidumbre en torno al estatus moral y legal de los robots cuidadores.

Por último, en el Cap. 9 Juan Ignacio del Valle y Miguel Moreno exploran el tema de las armas autónomas. Primero, los autores abordan el concepto de autonomía en sistemas armamentísticos y proveen una clasificación técnica de los diferentes tipos de armas autónomas; después, exponen las bases legales del uso de dichas armas; finalmente, abordan brevemente seis problemas en torno a ellas, uno de los cuales es el de la responsabilidad moral y legal. En el ámbito militar, hay una cadena de mando lo cual hace que muchas manos estén involucradas en el uso de un arma. Comúnmente, la responsabilidad del uso armamentístico cae en los gobiernos que los usan (y estos tienen que reparar el daño a los estados afectados). Sin embargo, se discute cada vez más si individuos particulares deben ser considerados como responsables del uso armamentístico y los daños provocados.

La tercera parte del libro versa sobre temas relacionados con la gobernanza y la regulación de la IA.

En el Cap. 10, Pedro Francés Gómez argumenta que la autorregulación de empresas privadas que producen IA, basadas en códigos éticos empresariales, no es suficiente, y que se requiere su adecuación a un sistema de gobernanza pública que, sin embargo, no está acabado. A este respecto, el autor expone a detalle lo que considera ser, al momento, el mejor modelo de gobernanza en torno a la IA: las *Directrices éticas para una IA fiable*, así como algunas de las propuestas de la Ley de IA, ambas de la Unión Europea.

Francés Gómez concluye remarcando algunas cuestiones que quedan abiertas para investigación y legislación futura, como el problema de la definición de la IA, su responsabilidad y su ambigüedad. Siempre será difícil tener una definición rígida de IA debido al rápido y constante progreso en su desarrollo; en torno a la responsabilidad, la estructura propuesta por la Ley de IA es cuestionable debido a que, similar al caso de las armas autónomas, hay muchas personas involucradas en su diseño, producción, distribución, etc. Tener un solo esquema de responsabilidad para tan diversos productos con IA y tantas personas involucradas, es problemático; la ambigüedad se refiere a que no es claro qué tipo de riesgos se corren con el uso de la IA. Nos es más fácil identificar riesgos materiales o tangibles, por lo cual los riesgos intangibles podrían pasar desapercibidos y sin regulación adecuada.

En el Cap. 11, Cristian Moyano Fernández y Jon Rueda reflexionan sobre la IA y la sustentabilidad, desde una perspectiva ético-ambiental. Su trabajo se centra en exponer y analizar el balance entre ventajas y desventajas morales del uso de sistemas de IA en relación con el ambiente. Un ejemplo es el siguiente: es posible utilizar IA para almacenar y procesar una enorme cantidad de datos para predecir y prevenir catástrofes naturales o identificar riesgos de contaminación y posibles soluciones, pero al mismo tiempo el vasto almacenaje y procesamiento significa el uso de grandes cantidades de energía.

Cerrando el libro, Antonio Diéguez y Pablo García Barranquero nos exponen el que es probablemente el tema con el que más se asocia -catastróficamente- a la IA desde el ámbito público: máquinas superinteligentes, nuestra interacción con ellas y el posible dominio sobre nosotros. Primero, los autores exponen varias definiciones de la Super Inteligencia Artificial General (hasta el momento solo tenemos Inteligencia Artificial Particular), así como el concepto de la Singularidad (el control de las máquinas), para después abordar el tema de la transferencia mental, cuyo objetivo último es el de lograr la inmortalidad. Los autores son escépticos en ambos casos. No creen plausible que el desarrollo técnico de la IA nos lleve a la creación de máquinas superinteligentes y por ende, a que estas tomen el control del mundo (al menos no voluntaria y conscientemente), ni que la transferencia mental -en caso de llegar a ser técnicamente posible- signifique la preservación de nuestra identidad personal en un mundo digital, es decir, que independientemente de que se pudiera transferir una copia idéntica de nuestra mente a una máquina, nosotros no estaríamos, como tal, presentes en esa máquina. A pesar del escepticismo en ambos casos, los autores advierten la importancia de controlar el desarrollo de la IA, así como de crear una gobernanza sólida en torno a ella y, por último, de no dejar que las decisiones más importantes sean tomadas por sistemas de decisión autónomos. Esto para asegurar que la IA contribuya al beneficio humano.

Para concluir, me parece que todos los autores incluidos en este libro, independientemente de la temática particular que abordan, hacen esta última advertencia. Se puede decir que este libro es, en su mayoría, precautorio, pues si bien no propone en ningún momento el cese de la investigación en inteligencia artificial, sí advierte que esta tiene que llevarse a cabo con cuidado, buscando siempre potenciar las cosas positivas de la IA y reducir las negativas. A este respecto, vale la pena

mencionar que gran parte de la literatura en torno a la Singularidad se basa en el advenimiento o la creación de Super Inteligencia Artificial General. Sin embargo, siguiendo la advertencia latente en todo el libro y explícita en el último capítulo, podría ser que, sin darnos cuenta, las máquinas tomen el control sin necesidad de llegar a ser superinteligentes, mediante una hiperdependencia a ellas.

No me gustaría terminar sin decir que este libro es una muestra de la firme actividad académica española en torno a la ética de la inteligencia artificial y del compromiso de algunas instituciones con la investigación sobre este tema. Esperemos más.

Robert Anthony Gamboa Dennis
(Universidad de Granada)