Reseña de Liao, M. (ed.), 2020, *Ethics of Artificial Intelligence*, New York: Oxford University Press.

ISBN: 978-0190905040

Germán Massaguer Gómez: "Reseña de Liao, M. (ed.), 2020, Ethics of Artificial Intelligence, New York: Oxford University Press"
ILEMATA, Revista Internacional de Éticas Aplicadas, nº 35, 79-83

Los caminos de la filosofía y la Inteligencia Artificial (IA) parecen estar irremediablemente ligados desde hace décadas. Este vínculo se ha incrementado a raíz del avance de esta tecnología en años recientes. Los dilemas éticos parecen multiplicarse a la par que los retos informáticos e ingenieriles, sobre todo en lo referente al pronóstico de un cambio radical en las sociedades actuales, un cambio que incumbe a nuestra forma de pensar, de entendernos como seres humanos y de relacionarnos con el mundo que nos rodea. La complejidad de los problemas que atañen a la IA requiere un riguroso y exhaustivo análisis filosófico que remite a cuestiones éticas, epistemológicas, políticas e incluso metafísicas que han habitado el pensamiento humano desde hace milenios. Este análisis debe ser llevado a cabo *ex ante*, ya que los riesgos, por poco probables que puedan parecer, son preocupantemente elevados. Esta es precisamente la posición defendida por los autores y las autoras cuyos artículos son recogidos en el libro *Ethics of Artificial Intelligence* editado por Matthew Liao. Peter Railton, Frances M. Kamm, Jean-François Bonnefon, Azim Shariff, Iyad Rahwan, Andrea Loreggia, Nicholas Mattei, Francesca Rossi, K. Brent Venable, Stephen Wolfram, Aaron James, Peter Asaro, Cathy O'Neil, Hanna Gunn, Kate Devlin, Nick Bostrom, Allan Dafoe, Carrick Flynn, Stuart Russel, Jessica Taylor, Elizer Yudkowsky, Patrick LaVictoire, Andrew Critch, Wendell Wallach, Shannon Vallor, Steve Petersen, Susan Schneider, Eric Schwitzgebel, Mara Garza y el propio Matthew Liao exponen a lo largo de este volumen sus puntos de vista respecto a problemas éticos sobre IA. Los autores y las autoras que participan en este libro aluden a cuestiones específicas y hacen propuestas significativas sobre cómo investigarlos, tratarlos e incluso resolverlos. Tras una magistral introducción de Liao en la que se resumen

prácticamente todas las cuestiones relativas a la IA, el libro se divide en cuatro grandes bloques, que abarcan consideraciones tanto presentes como futuras.

La primera cuestión que se plantea trata el diseño de máquinas morales autónomas —es decir, una IA que decida su propio curso de acción en un determinado momento sin intervención explícita humana, teniendo el resultado de dicha acción repercusiones morales. Más allá de la dificultad científica que suponga poder implementar un código moral a una máquina, se reincide sobre el problema de tener que determinar una teoría moral correcta antes de dicha implementación. Esta es una cuestión no resuelta y previsiblemente irresoluble. No obstante, las IAs con cierto grado de autonomía van a incorporarse a la sociedad y, por tanto, hay que pensar en modos de "educarlas éticamente" para evitar situaciones perjudiciales. Existen diferentes métodos para dicha educación. Railton propone un "bottom-up approach": si introducimos en una IA ciertas características similares a las de los humanos, como la curiosidad, la confianza, la cooperación o la deliberación, se llevarán a cabo los procedimientos propios del aprendizaje natural humano en la máquina. En otras palabras, el aprendizaje ético de las máquinas debe ser análogo al de los humanos y partir desde la experiencia y la observación (p.64). Cierto grado de incertidumbre es un seguro frente a la posibilidad de que una IA persiga su objetivo (dado) sea como sea (p.66). Los métodos opuestos a este son los "top-down", que sugieren que se deben codificar normas morales inquebrantables. En este libro encontramos una propuesta de este tipo en el capítulo de Loreggia et al, quienes defienden que las prioridades y preferencias de las IAs deben ser establecidas de antemano (p.132).

¿Cómo se manifiesta este aprendizaje ético en situaciones reales que podamos entender mejor? En la literatura sobre máquinas morales autónomas se presenta recurrentemente el 'Problema del tranvía' (*Trolley Problem*) y su adaptabilidad y similitud con el conocido caso del coche autónomo. Este dilema recorre todo el primer bloque, pero es importante reconocer que el problema del tranvía no es análogo al de las máquinas morales, como previene Kamm, ya que no implica una decisión consciente en el momento crítico (pp.89-91). Conviene subrayar que, independientemente de la resolución de los complejos problemas teóricos, los avances de la IA deben mejorar la vida de los usuarios y la sociedad en general, sea a través de un algoritmo que ayude a los jueces, que determine quién recibe un órgano o a quién debe "sacrificar" el coche autónomo. Por tanto, la sociedad debe aceptar las decisiones que se tomen, porque puede decidir prescindir de dichas tecnologías en caso de considerarlas injustas e inmorales. Como proponen en su artículo Bonnefon et al, la opinión popular es imprescindible a la hora de diseñar y educar a las máquinas (p. 123).

Como ya hemos señalado, el coche autónomo es uno de los casos más sonados en el debate ético sobre la IA y su integración a nuestra cotidianeidad. Pero existen otras cuestiones que amenazan con cambiar por completo el mundo tal y como lo conocemos. En el segundo bloque de este libro se tratan tres temas específicos: el desempleo, las armas autónomas y los robots sexuales. En el primer caso, es conocida la preocupación, existente desde la Revolución Industrial, sobre el remplazo de puestos de trabajo humanos por máquinas. A medida que la inteligencia de las máquinas vaya incrementando, más tareas cognitivas podrán desempeñar (James, p.183). Si bien no hay certeza de que esto pueda conllevar una situación de desempleo crítica a escala mundial, el riesgo es suficientemente elevado y probable como para tomar medidas ex ante. James propone un ingreso mínimo vital pre-

ventivo, ya que el remplazo laboral solo tiene sentido si el objetivo es mejorar las vidas de los humanos y no incrementar la pobreza a nivel global (p. 186, 189). El segundo capítulo de este bloque, escrito por Asaro, repasa las problemáticas que plantean las armas autónomas, entendidas como un sistema que elige los blancos y emplea la fuerza sin un control humano significativo. Los dilemas éticos que surgen van desde posibles ataques a civiles a la incertidumbre e imprevisibilidad que generan, así como la posibilidad de ser "hackeados" (p.215-223). Sin embargo, para Asaro existe un problema clave que permanecería incluso si estos problemas fuesen resueltos: si una máquina decide matar a un humano, sean cuales sean las circunstancias, esto estaría atentando contra su dignidad humana, dado que una máquina no puede comprender el valor de la vida (p.229). Por tanto, la integración de armas autónomas en el escenario bélico amenaza con atentar contra el derecho fundamental a la dignidad. Estos problemas son ya característicos del presente, así como el de los robots sexuales, tratado en este libro por Devlin, y otros como la vigilancia, las "cajas negras" o los algoritmos sesgados. Es necesario, como proponen O'Neil y Gunn en su capítulo, que a la hora de estudiar las implicaciones éticas de las inteligencias artificiales que tengan repercusiones sobre vidas humanas, se tengan en cuenta las consideraciones e intereses de todas las personas involucradas, especialmente de aquellas cuyas vidas van a verse efectivamente alteradas por estas tecnologías (p.242).

Las cuestiones planteadas en los primeros dos bloques tratan sobre implicaciones morales que tienen un impacto inmediato sobre nuestra sociedad. Sin embargo, existe una gran preocupación sobre lo que pueda ocurrir en el futuro por muy lejano y ficticio que pueda parecer. El tercer bloque trata sobre el fenómeno conocido como súperinteligencia, explosión de inteligencia o singularidad, a saber, el momento en que una IA supere significativamente a la inteligencia humana en todos los dominios cognitivos relevantes.

El ejemplo ya clásico de Bostrom que es nombrado para ilustrar esta posibilidad alude a una súperinteligencia cuyo objetivo sea crear clips de papel y acabe desarrollando una tecnología que le permita convertir al mundo entero en clips de papel, incluyendo a los seres humanos. Uno de los conceptos clave es la alineación de valores ('value alignment') que implica tener la seguridad de que la súperinteligencia conozca los valores éticos de la humanidad para que, independientemente del grado de inteligencia que pueda alcanzar, siempre trabaje y desarrolle sus tareas en concordancia con estos. Los cinco capítulos de este bloque son tanto una advertencia sobre los posibles peligros que entraña esta tecnología como una lista de propuestas y soluciones que deben ser tomadas con anterioridad, puesto que el riesgo es altamente elevado: nada más ni nada menos que la extinción del ser humano. Bostrom et al. se centran en los retos políticos que emanan de estas tecnologías y en cómo los gobiernos deben actuar de antemano frente a todas las posibles consecuencias en diferentes ámbitos según vayan avanzando las tecnologías (pp. 312-314). Russell, por su parte, incide sobre la necesidad de incorporar cierta incertidumbre en las máquinas respecto a los objetivos humanos. Esta visión se asemeja a la expuesta por Railton, quien considera a la incertidumbre y el aprendizaje por experiencia como fundamentos claves para el desarrollo ético de las máquinas. Si los objetivos no están impuestos, y la máquina aprende a partir de la observación de la conducta de los humanos, se puede evitar, por un lado, que esta tome decisiones indeseables para alcanzar dicho objetivo y, por otro lado, que busque formas para no ser apagada (p.334). Se trata, por tanto, de otro "bottom-up approach". Este puede ser útil para una IA limitada, como propone precisamente Railton,

ISSN 1989-7022

ISSN 1989-7022

Dilemata, año 13 (2021), nº 35, 79-83

pero parece complejo de aplicar a la hora de hablar de súperinteligencia. Esto mismo defienden Wallach y Vallor en su artículo. Frente a la alineación de valores pura, proponen su combinación con la encarnación de la virtud ('virtue embodiment') (p.383). Para garantizar la seguridad, debe conseguirse que exista una percepción del mundo físico similar a la de los humanos; una consideración análoga a la que tenemos de nuestros cuerpos y el mundo que nos rodea (p.386).

Otro de los problemas a largo plazo que atañen a las discusiones filosóficas sobre IA es el relativo a la conciencia. Más allá de la pregunta ¿puede existir una conciencia artificial?, las preguntas se encaminan alrededor de qué pasaría si se llegase a dar una conciencia en una entidad artificial. Esto no es un problema aislado, sino que afecta a todos los temas discutidos en este volumen. Esta discusión es relegada muchas veces a un segundo plano por altamente improbable, pero sin duda requiere de nuestra atención. En los artículos relativos a este tema, los del cuarto y último bloque, se incide sobre todo en lo que significaría para las entidades artificiales tener conciencia. Los autores de estos tres capítulos concuerdan en que, en caso de darse cierto grado de conciencia en las máquinas, estas deberían recibir cierta consideración moral, y, por tanto, no pueden ser tratadas como una lavadora o un teléfono móvil (Schneider, p.440; Schwitzgebel y Garza, p.460; Liao, p. 481). Es importante, por tanto, tener siempre presente la posibilidad, por remota que parezca, de que se esté tratando con entidades potencialmente conscientes y, por tanto, merecedoras de cierto estatus moral, puesto que existe un riesgo de estar infligiendo un daño real (Schneider, p.454). Por tanto, hay que incluir a la entidad artificial dentro de la matriz ética propuesta por O'Neil y Gunn. Este monumental volumen acaba con el capítulo escrito por el propio Liao, advirtiendo, en su última página, sobre inteligencias artificiales que sean auténticos agentes morales con conciencia, sentimientos, capacidad de querer, de sufrir, de razonar; en definitiva, capacidades cognitivas análogas a las de los seres humanos —aquello que nos hace, propiamente, ser humanos. Liao llega a usar la palabra "alive" (vivas), lo cual implica, sin duda, un enorme paso conceptual (p.497). Se está abriendo la puerta a la posibilidad de que haya vida más allá de la biología. Si esto llegará a ser una realidad o no permanece una incógnita, pero es una discusión que está en el centro mismo del estudio ético de la IA. Los debates sobre armas y coches autónomos, sobre robots sexuales u otros robots de compañía, sobre aprendizaje moral, sobre súperinteligencia, etcétera, se verían alterados por completo si las inteligencias artificiales pudiesen desarrollar al menos cierto grado y cierto tipo de conciencia.

Este libro no es un manual introductorio al estudio de la ética de la IA. No se trata de un repaso de los temas y problemáticas más candentes, es decir, no da vueltas sobre el estado de la cuestión. Ciertos temas no reciben un capítulo específico (sobre todo destacado el problema de la vigilancia ligada a la IA). Cada autor y cada autora que aparecen en este volumen trata sobre problema específico sobre el que se hacen propuestas, se emiten dudas, y se demuestra un conocimiento remarcable sobre el tema. Muchos de los capítulos parecen desligados de los demás; islas aparte —por ejemplo, el capítulo de Wolfram sobre el desarrollo de un lenguaje codificado y su aplicación al derecho (contratos). Repito, este libro no es un manual. Es más bien una antología de artículos científicos. Cada uno refleja una investigación exhaustiva y propuestas ciertamente interesantes respecto a los temas que tratan. Por ejemplo, el artículo de Devlin "The Artificial Lover", no es un repaso a las discusiones que generan los robots sexuales, sino una propuesta para reconvertirlos en

aparatos no antropomórficos. Esto, por supuesto, no quiere decir que no aparezcan señalados los problemas básicos que implican los robots sexuales para usuarios y sociedad. Sin esta base la argumentación carecería de solidez. Simplemente considero relevante subrayar que este libro no es descriptivo sobre los problemas, sino que trata cuestiones muy específicas. No recomendaría este libro como introducción al tema, pero sin duda lo haría como profundización.

Germán Massaguer Gómez
Universidad Carlos III de Madrid
100314957@alumnos.uc3m.es

ISSN 1989-7022