Mediaciones Sociales
ISSN-e: 1989-0494

https://dx.doi.org/10.5209/meso.78322

EDICIONES COMPLUTENSE

Etica para máquinas

Latorre, J. I. (2019). *Ética para máquinas*. Barcelona: Ariel.

La Cuarta Revolución Industrial nos está transportando a un nuevo mundo y una nueva sociedad, en la que las máquinas son igual o más inteligentes que los propios seres humanos. Hemos pasado de los molinos para triturar trigo al robot Da Vinci para realizar operaciones quirúrgicas con la guía de un médico. Esta nueva revolución marcará un antes y un después en nuestra historia, puesto que las máquinas nos gobernarán. ¿O podemos evitarlo dándoles una ética humana?

En *Ética para máquinas*, José Ignacio Latorre reflexiona sobre cómo educaremos y qué ética daremos a unas nuevas máquinas que nos superarán en todos los sentidos, tomarán el control y, nos gobernarán.

José Ignacio Latorre es catedrático de física teórica en la universidad de Barcelona y director del Centro de Ciencias Benasque Pedro Pascual y el autor del libro "Ética para máquinas" publicado en 2019 por la editorial Ariel. Este es su cuarto libro publicado después de "*Cuántica: tu futuro en juego*" (2017), "*Fonaments de física*" (1998) y "*Notes sobre relativitat especial*" (1982).

La primera parte del libro, titulada "Máquinas sin alma", lleva al lector por un recorrido histórico que comienza con el primer mono que usó un hueso como arma contra un enemigo, hasta llegar a la inteligencia artificial que estamos desarrollando actualmente. Explica de forma amena y con ejemplos sencillos cada gran avance histórico: cómo las máquinas han ido ocupando las tareas físicas que realizaba el hombre y la forma en la que actualmente llevan a cabo las tareas mentales. Recorreremos cientos de años y comprenderemos como funcionan las máquinas "sin alma" y aquellas que si la tienen (las que disponen de inteligencia artificial). Desde máquinas veloces como el avión, pasando por máquinas extremadamente precisas como un reloj hasta llegar a máquinas asesinas, como las bombas de nitrógeno o atómicas. En esta parte del libro aprenderemos también sobre el origen del cálculo y los números; este viaje nos llevará desde la invención del ábaco en Mesopotamia, la máquina diferencial de Ada Lovelace de mitad del siglo XIX y el primer ordenador programable o "Zi" de Konrad Zuse en 1936. Recibiremos una sencilla clase de programación y conoceremos un hecho histórico del que poco se habla: el día que la máquina ganó al ser humano. Gary Kasparov, el mejor jugador de ajedrez de la historia contra Deep Blue, un superordenador entrenado por grandes maestros del ajedrez.

La llegada de las máquinas ha provocado el declive tanto físico como mental del ser humano, ya no tenemos que realizar grandes esfuerzos físicos y, por tanto, nos hemos debilitado físicamente al no necesitar tener una gran capacidad física. Mentalmente también hemos dejado de realizar esfuerzos. De pequeños se nos enseña a realizar divisiones y multiplicaciones y al terminar el instituto, somos capaces de realizar logaritmos neperianos, derivadas e integrales. Pero ahora, años después de terminar nuestra formación, utilizamos la calculadora del móvil para dividir entre 2 la cuenta del restaurante. Estamos cediendo nuestro espacio intelectual a las máquinas.

En la segunda parte del libro titulada *Máquinas que parecen inteligentes* el autor reflexiona sobre el proceso de aprendizaje de los humanos y lo poco que sabemos de él. El aprendizaje del ser humano se compara con el de las máquinas, puesto que si no sabemos como aprende un humano adulto o incluso un bebé, que suele ser por ensayo y error, ¿cómo vamos a enseñar a las máquinas a pensar? Debemos empezar por la creación de sistemas expertos, que no es otra cosa que un humano experto en un problema concreto y que es aumentado con toda la potencia de cálculo de un ordenador. Y así fue como IBM creó a Deep Blue: unieron a los grandes maestros del mundo de ajedrez con la potencia de un ordenador. Desde entonces se han creado máquinas como Stockfish que gana sin problema a cualquier ajedrecista y máquinas que compiten contra sí mismas y aprenden mucho más que jugando contra humanos.

Pero ahora llegamos al quid de la cuestión de este libro, ¿hasta qué punto podemos aplicar la filosofía de sistemas expertos a problemas de relación entre máquinas y humanos? Un gran problema que encontraron los investigadores fue que las máquinas no entendían a los humanos, no entendían su idioma. Los sistemas expertos sufrieron un estrepitoso fracaso: eran incapaces de resolver problemas difíciles de modelar. Tras este fracaso apareció un nuevo reto: entender como aprendemos los humanos, para así enseñar a las máquinas a pensar. En 1943, McCulloch y Walter H. Pitts idearon una neurona artificial que funcionaba de manera similar a una neurona humana pero más sencilla. La información entra en la neurona artificial, se procesa, se genera una nueva información y esta pasa a otra neurona. Este gran invento generó muchísimas preguntas como las que establece Latorre "¿Qué significa el procesamiento que hace la neurona? ¿entendemos lo que está

haciendo? ¿cómo enseñamos a una neurona artificial la forma en la que ha de comportarse? ¿cómo entrenamos a una neurona artificial? ¿cómo la educamos?" (p. 106). Más de dos décadas más tarde Frank Rosenblatt propuso crear una variante de neurona artificial llamada *perceptrón.* Esta variante se veía como el inicio a la inteligencia artificial, pero resultó ser un estrepitoso fracaso, puesto que solo podía resolver problemas lineales. El problema de este nuevo invento fue que un perceptrón no puede trabajar por si solo, necesita una red: una red neuronal artificial, una primera capa de neuronas inicia el procesamiento, la segunda toma el resultado y la elabora más y así capa a capa hasta llegar a la última que nos da el resultado final. El autor señala que ya se ha dado con el corazón del problema: un bebé que aprende está fijando sus sinapsis, entonces una red neuronal artificial que aprenda debe ser capaz de fijar sus sinapsis artificiales, sus conexiones. Es decir, aprender a aprender.

Una vez hemos enseñado a las máquinas a aprender, empezará a decidir: desde toda la flota de coches de un país controlada por IA, el sistema judicial o los futuros médicos que salvarán nuestras vidas. Se espera que las máquinas con IA tomen el control y las decisiones que los humanos somos demasiado imparciales para tomar.

Llegamos a un punto de inflexión: los humanoides. Estos nuevos seres generarán lo que se llama la teoría del valle inquietante: miedo a los robots excesivamente realistas. Nuestra sociedad no está preparada para convivir con humanoides que pueden comunicarse y actuar como un humano. E iremos más allá con la aparición de los cíborgs: un híbrido de humano y máquina. Esta nueva tecnología nos llevará a donde muchos llevan años intentando llegar: la amortalidad humana. Este término recuerda a la teoría del superhombre de Nietzche: un nuevo ser que no responde ante nadie más que a sí mismo, que no se iguala a nadie. Un nuevo hombre que domina a quienes le rodean y decide según sus valores. (Fernández-Blanco Inclán 2020). El transhumanismo está permitiendo a los humanos crear "la velocidad de escape de la longevidad", fenómeno que recoge Cordero y Wood en La muerte de la muerte, fenómeno que permite que por cada año que vivimos, ganemos otro más. (Cordeiro y Wood, 2018)

Nos adentramos en la tercera parte del libro, *Ética para máquinas*, que comienza con una reflexión sobre la caída de la ética protestante y la llegada del Estado de derecho, puesto que nuestra sociedad estará regulada por inteligencia artificial que no se dejará sobornar por las grandes corporaciones y trabajará para hacer nuestra sociedad más justa. Dejaremos atrás la eterna lucha de izquierda/derecha derechas y comenzará la lucha humanos/máquinas: tendremos que luchar por no perder nuestro puesto de trabajo frente a una máquina.

Entrando de lleno en la ética para máquinas, el autor nos propone un imaginario: "una persona entra con un arma de fuego a robar un banco. Se muestra alterada, se mueve sin control y grita de forma muy agresiva. Varios policías le plantan cara y el atracador termina disparando y matando a uno de ellos. Un robot de seguridad que controla un fusil tiene en el punto de mira al atracador. ¿Qué hace?" (p. 193) El imperativo categórico kantiano de no matar le parece irrefutable, puesto que no tienes que hacer lo que no quieres que te hagan, en este caso, matar:

El autor reflexiona la filosofía de Kant y concluye que debemos considerar la posibilidad de programar a las máquinas usando como ética el imperativo categórico de Kant:

> Acto o proposición que se lleva a cabo por el hecho de ser considerada necesaria, sin que existan más motivos para ser llevada a cabo que dicha consideración. Serían las construcciones que se realizan en forma de "debo", sin estar condicionados por ninguna otra consideración, y serían universales y de aplicación en cualquier momento o situación (Castillero Mimenza 2021).

Pero si no hace nada, más gente puede morir. El robot se encuentra en un dilema ético que no puede resolver porque los humanos tampoco sabemos resolverlo. Una solución a esta situación puede ser la corriente filosófica llamada utilitarismo de Bentham, quien defendió que las acciones deben regirse por el principio del bien común, defendiendo que el mayor número de personas sean beneficiadas (p. 194). Esta filosofía soluciona rápidamente el dilema del ladrón: si el robot dispara, solo muere el ladrón. Si no lo hace, mueren más personas.

Entramos en otro de los dilemas de la inteligencia artificial: la legislación. ¿Deben las máquinas con inteligencia artificial responder ante la ley? ¿o debe ser su creador? Pongamos un ejemplo para explicar esta situación: un coche controlado por inteligencia artificial se estrella contra un coche aparcado en la calle.

¿Debe ser el coche el que pague las consecuencias, como si se tratase de una persona física? ¿o debe pagar el conductor? ¿o el fabricante de la IA en su lugar? Estas cuestiones han abierto un gran debate y muchos piensan que la máquina no debe ser responsable de este tipo de errores, puesto que si se hace la empresa responsable no tendría ningún tipo de penalización y no tendría que responsabilizarse de que sus máquinas fallen y puedan matar a alguien.

Nos adentramos en la cuarta y penúltima parte del libro llamada "Máquinas que nos superarán" en la que el autor realiza una reflexión crítica sobre el futuro de los seres humanos y la singularidad. Llegaremos a un punto en el que las máquinas dispondrán de una inteligencia artificial tan avanzada que se podrá mejorar así misma, al igual que ha pasado con máquinas que juegan al ajedrez y juegan contra sí mismas para ser mejores. La singularidad nos llevará a un momento en el que los humanos no tengan nada que hacer contra

las máquinas y no podremos modificar su comportamiento o su ética. Solo nos quedará esperar que la persona que las programó introdujese una buena ética en ellas. Por ejemplo, podemos programar una inteligencia artificial para que proteja el medio ambiente y esta puede decidir dos cosas: acabar con los humanos, ya que somos la especie más destructiva o cambiar el sistema en el que vivimos a uno completamente autosostenible y limpio.

Para Steven Pinker autor del libro *The Better Angels of Our Nature: Why Violence Has Declined*, los humanos seguimos un proceso de pacificación: las probabilidades de morir asesinado decrecen significativamente a lo largo de la historia. La violencia es parte de nosotros porque nos permite sobrevivir, pero actualmente somos una sociedad volcada en la inteligencia y eso significa que poco a poco la violencia desaparecerá de nosotros porque no nos va a ayudar a sobrevivir, pero ser inteligentes sí.

Llegamos finalmente a la última parte del libro: "Imitar el alma". Está claro que la primera carga de ética y alma a la inteligencia artificial tiene que hacerla un humano, pero luego esta irá cambiando para ajustarse mejor a su objetivo, al igual que hemos hecho nosotros en un proceso de pacificación. Hemos aprendido que ningún tipo de violencia nos reporta nada positivo, ni las guerras, ni la violencia machista ni la doméstica, ni el maltrato animal entre muchas. Este capítulo reflexiona sobre los posibles sentimientos de la inteligencia artificial, como si será capaz de amar o de sentir emociones humanas y a que punto le llevará esta humanización.

La obra de Latorre es un ensayo para todos los públicos que invita a reflexionar sobre el futuro de la humanidad y nuestra forma de relacionarnos con las máquinas. El autor pone el punto de mira en la ética que debemos dar a las máquinas, de ahí el título del libro, y este es un tema del que no se habla a menudo. Nos perdemos en la tecnología y en todo lo que se puede conseguir que nos olvidamos de que todo se puede dar la vuelta y terminar como en las películas apoteósicas.

Referencias bibliográficas:

Fernández-Blanco Inclán, J. (2020, 17 febrero). La fuerza natural del superhombre según Nietzsche. Recuperado de https://www.filco.es/fuerza-natural-superhombre-nietzsche/

Cordeiro, J. L., & Wood, D. (2018). La muerte de la muerte. *La posibilidad científica de la inmortalidad física y su defensa moral. Barcelona: Deusto*.

Castillero Mimenza, O. (2021, 12 enero). El imperativo categórico de Immanuel Kant: ¿qué es? Recuperado de https://psicologiaymente.com/psicologia/imperativo-categorico-kant

Andrea Estebaranz Escribano
Universidad Complutense de Madrid