INTELIGENCIA ARTIFICIAL Y MORALIDAD

La inteligencia artificial se usa en multitud de aspectos de la vida porque la máquina, que almacena millones de datos, puede tomar en cuestión de segundos una impresionante cantidad de decisiones en ámbitos muy diversos. Automóviles sin conductor, drones, gestión de urgencias en enfermos… son ámbitos de la vida en que la Inteligencia Artificial puede tanto puede hacer maravillas como producir el mal. ¿Es posible programar las máquinas de acuerdo a criterios morales?

En 2016 el Instituto Tecnológico de Massachusetts puso en línea el Test de la Máquina Moral que tenía por objeto sondear nuestras intuiciones morales cuando un coche autónomo, del tipo Google Car, se encuentra ante una tesitura dramática. Supongamos que solo hay dos opciones: que de estampe el vehículo contra un muro, matando al conductor, o que mate a tres niños que salen de la escuela. ¿Qué haríamos? ¿Y si los niños han atravesado con el semáforo en rojo? ¿Y si hay un bebé en el coche? Se trata de una variante del dilema del tranvía que propuso en 1967 Philippa Foot, en que un tranvía a gran velocidad se lanza cuesta abajo y amenaza con matar a cinco personas en la vía. ¿Se debería accionar el mecanismo de guardagujas y desviarlo para que solo mate a una persona? En 1985 Judith Jarvis Johnson propuso una variante de esta cuestión. Imaginemos que sea posible parar el tranvía cuesta abajo lanzando desde un puente a un hombre obeso, que moriría, pero salvaría de la muerte a otros cinco ¿Lo haríais?

EL CASO DEL VEHÍCULO SIN CONDUCTOR

Un mundo donde todos nos desplazásemos en vehículos sin conductor sería deseable porque los coches autónomos reducirían los accidentes en más del 90% de los casos. Además, en un coche sin conductor no sería necesario pagar un seguro, porque no lo conduce nadie, y eso representa un buen ahorro. Pero aun así un padre de familia estándar sería reticente a comprarlo si se plantease el dilema moral que hemos visto. En principio, si uno es utilitarista, es decir, si cree que lo mejor moralmente es producir el máximo bien (o el mínimo mal) para el máximo número), entonces debiera preferir (o lamentar menos) la muerte de una persona que la de cinco. Pero activar la palanca haría morir a alguien inocente que, en circunstancias normales no habría muerto. Tirar por un puente a un señor mayor gordo, incluso podría ser considerado un asesinato – dejando aparte el hecho de que no tiene culpa de ser gordo y poder, por lo tanto, impactar con mayor fuerza sobre el tranvía. Supuesto que uno no es Dios, ¿puede disponer de las vidas de los otros? Pero, aunque uno no sea Dios, ¿por qué no minimizar el dolor de cinco, aunque sea a costa del dolor de uno?

¿Podemos construir un automóvil programándole la orden moral de “no matarás en ningún caso”? Parece que no resultaría práctico, porque entonces desarrollaría muy poca velocidad. Incluso programar un coche así podría ser contraproducente, porque sabiendo que no hay ningún peligro potencial a la vista, la gente que va en el coche podría emborracharse o hacer cualquier otra cosa más peligrosa para ellos o para los demás. Incluso puede argumentarse que si la gente no muriese de accidentes de circulación disminuiría mucho la cantidad de donantes potenciales de órganos necesarios para operaciones de trasplante (de corazón, de hígado, de pulmón, etc.), que salvan muchas vidas.

Podríamos complicar más el caso. Supuesto que los compradores de ese tipo de coches serían gente de una cierta edad y de un gran poder adquisitivo; ¿sería factible programar un coche para que solo matase pobres o para que solo matase gente joven – que podríamos

aprovechas para trasplantes? Evidentemente, eso no sucederá porque si ese criterio se divulgase aparecerían en la prensa informaciones como “los coches de la marca X son asesinos de niños” y la gente no los compraría (excepto quienes odian a los niños, eventualmente). Los constructores de tales tipos de vehículos no se atreverían a construirlos por miedo a un proceso judicial. Hoy por hoy cuando se produce una muerte en carretera no se imputa a la marca del vehículo (Seat, Honda, Renault), sino al conductor (Pepito, Luisita...). Pero en el caso de un coche autónomo, la responsabilidad de un error sería del fabricante, lo que hace improbable que exista un coche así, porque no parece que los fabricantes quieran arriesgarse a procesos penales, excepto en el caso de que sea el Estado quien cubra los gastos.

EL CASO DEL DRON ASESINO

El dron fue básicamente concebido para matar. Hoy por hoy muchos de ellos son dirigidos por control remoto por militares, que físicamente pueden encontrarse a miles de quilómetros del teatro de operaciones. Pero están en estudio drones que podrían decidir por sí mismo sus objetivos. ¿Habría que programarlos con instrucciones que prohibiesen atacar niños o población civil? En 2015, el filósofo Peter Asaro cofundador del ICRAC (*Internation Comittee for Robot Arms Control*), una organización internacional que agrupa a investigadores que pretenden llagar a criterios éticos el diseño de las armas de guerra, publicó un manifiesto: “Prohibid los robots asesinos”, que firmaron 1.500 personalidades, entre las cuales Elon Musk, Stephen Hawking o el filósofo Daniel Dennett.

De hecho, sería muy fácil programar un dron con un dispositivo de reconocimiento visual para hacerle capaz de distinguir si había niños o personas mayores en el objetivo y evitar abrir fuego. El derecho de guerra prohíbe matar a no combatientes; pero ¿podría distinguir un dron si alguien que le lanza una piedra es o no un combatiente, aunque sea inofensivo?

Sospechamos además que el dron asesino podría provocar más guerras que el desplazamiento de soldados humanos. Resulta mucho más sencillo declarar una guerra y justificarla ante la opinión pública si nuestros líderes saben que no va a morir nadie entre los nuestros, porque “solo” bombardearemos al enemigo mediante programas de ordenador con instrumentos mecánicos y sin intervención física, – mientras que, en cambio, difícilmente podría soportarse manifestaciones de madres viendo llegar los féretros de sus hijos caídos en combate.

Otro tema delicado en el ámbito de la guerra justa es el de la proporcionalidad. Estamos de acuerdo en que no debe destruirse una ciudad entera para matar a cinco personas. Pero, ¿qué pasa si un asesino se esconde en un hospital infantil? ¿El dron debería estar programado para evitar bombardear ese hospital, o no? ¿Y en una escuela? Hay que tener en cuenta, además, la cuestión de la intencionalidad. Una guerra nunca es justa cuando los combatientes encuentran un placer sádico en la muerte y en la destrucción. Nihilistas y sádicos son individuos considerados generalmente como amorales o inmorales. El militar que actúa con justicia detesta la muerte. La guerra solo es justa cuando quienes participan en ella sienten que preferirían no haber de hacerla y la hacen pese a sus deseos. Pero esa tensión moral al tomar decisiones no existe en la máquina, que calcula y no se emociona.

WITTGENSTEIN Y LA CERTEZA

El problema ético que plantea la Inteligencia Artificial es el de la contextualidad de las decisiones. Excepto si usted es un kantiano un poco antiguo, una decisión moral normalmente solo tiene sentido cuando se sitúa en un contexto. En general los humanos no

nos contentamos con seguir una regla. Las reglas solo existen y se vuelven significativas cuando se entrelazan con otras reglas y cuando conocemos las instrucciones de uso que las acompañan y, a veces, las limitan o las matizan Una regla moral funciona en un paisaje social, político, cultural, etc. Cuando ignoramos el contexto, resulta poco significativo decir, por ejemplo, que "la vida es sagrada" o que "una decisión moral debe ser imparcial". *Summa lex, suma iniuria*, propusieron los clásicos. Wittgenstein en sus apuntes *Sobre la certeza*, que se publicaron póstumamente, aporta cosas muy a tener en cuenta cuando especulamos sobre el problema de la Inteligencia Artificial. Nos recuerda, por ejemplo que la relación entre certeza y verdad es complicada porque la certeza implica una regla práctica y las reglas prácticas, a veces incluso pueden ser contradictorias, de manera que lo único que nos permite discriminar entre ellas es su uso. Para el filósofo austríaco «La certeza es, por así decirlo, un modo de voz en que se declara como son las cosas, pero que no concluye». Para él "creer" y "saber" eran modos indistinguibles: «Pensar que estados diferentes han de corresponder a las palabras "creer" y "saber" es como pensar que personas distintas han de corresponder a las palabras "Yo" y "Ludwig", porque los conceptos son diferentes». Lo expresó en un par de aforismos, de una manera incluso brutal: «Lo que sé, lo que creo» (177) y «A fin de cuentas, el saber se funda sobre el reconocimiento» (378).

La Inteligencia Artificial es, en buena manera, reconocimiento (de formas, modelos e incluso de rostros o de conductas) para establecer reglas de uso. Pero – ¡y ese es el problema! – los seres humanos difícilmente nos contentamos con seguir las reglas de manera pasiva. Nuestros juicios morales suponen una cantidad ingente de presuposiciones, de experiencias previas, de conocimientos explícitos o implícitos, etc., que hoy por hoy, por lo menos en el estadio actual de nuestra tecnología, no es posible programar y de los que, muchas veces, ni tan siquiera somos conscientes. Mientras lo que sabemos y lo que creemos resulte indistinguible, como pensaba Wittgenstein, el juicio moral seguirá siendo necesario. Y en el ámbito de las tecnologías de la Inteligencia Artificial fácilmente se puede falsificar información, como sabemos desde el ya lejano julio de 2017 cuando investigadores de la Universidad de Washington lograron "hacer pasar" como auténtico a todo el mundo un discurso que el presidente norteamericano Obama no había pronunciado jamás, reivindicando su legado. A Obama eso mismo le volvió a suceder en 2018 cuando el actor Jordan Peele le suplantó en un discurso haciéndole decir cosas como "El presidente Trump es un completo imbécil" mediante el uso de After Effects y la aplicación FakeApp.

El uso de la Inteligencia Artificial como herramienta de control político al servicio del totalitarismo puede tener consecuencias siniestras y no es algo que pueda ser descartado como ciencia ficción. China ha intentado erigirse en la dictadura perfecta controlando a millones de individuos a todas horas y todos los días mediante la combinación de la tecnología de reconocimiento facial y un de carnet por puntos en el que se refleja la mansedumbre de los súbditos. Los puntos que se pierden en el carnet de ciudadanía cuando alguien comete alguna infracción a las reglas (decididas por los dirigentes del partido) significan que el infractor cae directamente en la categoría de paria social y no puede, por ejemplo, acceder al carnet de conducir, conseguir un ascenso en su trabajo o estudiar en alguna universidad. En 2003 el partido único inició un "Proyecto Escudo Dorado", que empezó a funcionar en 2006 y, que de hecho era una formidable máquina de censura. Se clonó Twitter, Google, YouTube, que en China se convierten en Qzone, Weibo, Baidu, Youku… controladas por el gobierno. Desde entonces el control de la vida privada mediante el uso de Inteligencia Artificial no ha parado de crecer – y no solo en China, obviamente. ¿Es posible programar las máquinas de acuerdo a criterios morales? Parece difícil que eso suceda por razones incluso de definición conceptual, como nos recuerda Wittgenstein. Pero

la necesidad de una gestión ética de la información es obvia. El juicio ético que tiene en cuenta el contexto de las reglas es básico para evitar que la aplicación mecánica de las reglas, ciegas por ellas mismas, acaben conduciendo al totalitarismo tecnológico.