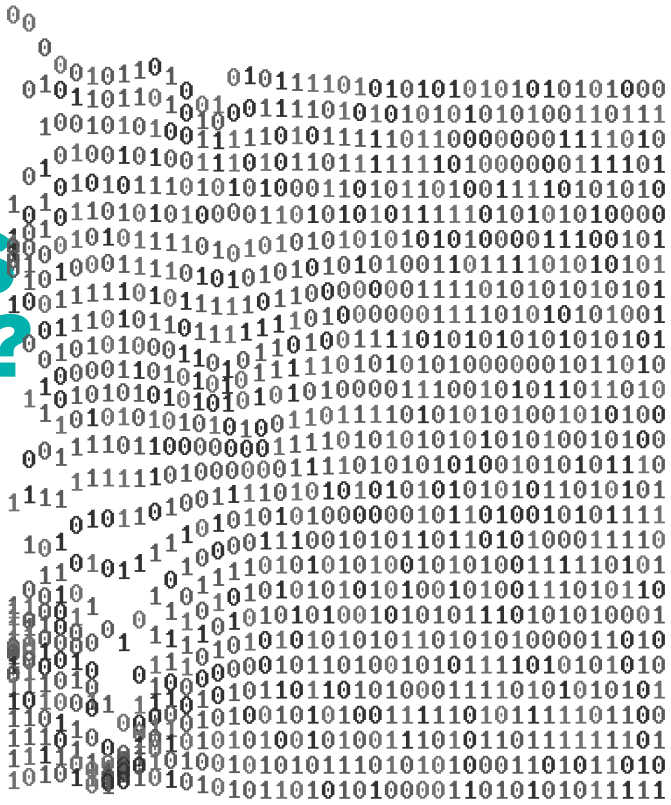
¿PUEDEN SER RESPONSABLES LAS MÁQUINAS?

por **Antonio Diéguez**

Suele esgrimirse la falta de voluntad en la IA como argumento contra la responsabilidad sobre sus acciones. Hay mucha literatura al respecto, pero conviene asumir que habrá máquinas que atiendan a razones morales.



El asunto de si podrían tener las máquinas alguna vez responsabilidad moral ha cobrado un interés creciente en la discusión filosófica sobre la tecnología. Esto ha ido ocurriendo a medida que han mejorado los sistemas de inteligencia artificial, que son los únicos entes artificiales que podrían aspirar con algo de plausibilidad a ser considerados como agentes morales. Muchos de los autores que han tratado la cuestión consideran que las máquinas carecen, al menos por el momento, de los requisitos necesarios para la atribución de responsabilidades. No pueden responder de sus acciones porque no tienen voluntad libre ni consciencia para realizar esas acciones a sabiendas (aunque sí sean buenas previendo las consecuencias). Nótese que hablamos de responsabilidad moral y no de responsabilidad causal o de responsabilidad legal. En principio, a los robots y a las máquinas de IA se les puede atribuir responsabilidad legal igual que se hace con las corporaciones o empresas, como señaló una resolución de la Unión Europea sobre este asunto (Hern 2017).

Pero ¿no podría ser esto un prejuicio antropocéntrico? Así como hay quien afirma que la inteligencia artificial está mostrando cómo se puede ser inteligente de forma muy distinta al modo en que lo son las personas (o algunos animales), ¿no podría ocurrir que hubiera también una forma distinta de ser agente moral, y por tanto susceptible de responsabilidad, que fuera muy distinta a la del ser humano y que encajara con el comportamiento de las máquinas?

Por otro lado, si bien es cierto que no tenemos aún máquinas conscientes y es muy posible que no las tengamos jamás, puesto que la IA puede funcionar muy bien sin consciencia, lo que sí cabe decir es que tenemos máquinas autónomas en un cierto sentido del término. Algunas máquinas, en efecto, toman decisiones con independencia del control humano y en ocasiones su comportamiento resulta completamente imprevisible para los programadores. Por ejemplo, las armas inteligentes autónomas seleccionan por sí mismas los objetivos que han de abatir, y en un futuro no muy lejano tendremos quizá coches autónomos, con un grado de autonomía mayor que los que circulan ya en fase de prueba. Aunque ciertamente no serían entidades autónomas en un sentido más fuerte que incluyera el libre albedrío, lo serían al menos en el sentido de que actúan controladas por sí mismas, es decir, autodirigidas y sin input humano o impedimento externo. Pero ¿es este un sentido de autonomía que pueda valer para una atribución ética de sus acciones? ¿Cabe responsabilizarlas de esas acciones incluso aunque no pueda decirse que hubo voluntariedad en ellas?

La respuesta inmediata parece ser que no, que la responsabilidad requiere agencia moral y la agencia moral no se basa en la mera autonomía en sentido lato. No obstante, puede que esta intuición esté equivocada. ¿No tendría sentido imponer un castigo a una máquina que cause un daño sin que un ser humano la haya condicionado en su decisión? ¿No debería pedir disculpas? ¿No debería ser destruida o "reparada" en su comportamiento si el daño hubiera sido grave? ¿Veríamos esta reparación como similar al arreglo de unos frenos gastados en un coche que ha atropellado a una persona? ¿O consideraríamos que tiene de alguna forma el sentido de una expiación por el daño

causado? ¿Sería una simple farsa castigar a una máquina? Y si lo fuera, ¿qué las distinguiría en tal caso de los animales, a los que a veces castigamos por su conducta, sin considerarlos por ello moralmente responsables? Algunos estudios empíricos muestran, de hecho, que los seres humanos tenemos tendencia a atribuir cierta responsabilidad a los robots autónomos (Furlough *et al*. 2021). Pero también se ha dicho que los intentos por responsabilizar a las máquinas han sido criticados a menudo como estrategias poco hábiles para quitar responsabilidad (moral y legal) a las empresas y a los seres humanos que están detrás de determinada tecnología.

Carissa Véliz (2021) ha argumentado que los algoritmos de la IA son zombis morales y no agentes morales, puesto que ni sienten, ni tienen emociones morales, ni pueden formar valores ni actuar en función de ellos, y todos esos requisitos son necesarios para hablar de una auténtica agencia moral. La noción de autonomía que es necesaria para la acción moral es mucho más exigente que la noción que se emplea cuando se dice que algunos de los sistemas de IA pueden tomar decisiones autónomas. No basta con que los algoritmos puedan cambiar de estados o decidir con independencia de sus creadores. Eso no es suficiente para atribuirles agencia moral ni, por tanto, responsabilidad moral. No se autodeterminan mediante razones, no comprenden la situación en la que han de actuar, no tienen consciencia de su acción, solo operan en función de las instrucciones o el entrenamiento recibido y la propia estructura del algoritmo, factores que el propio algoritmo no puede valorar moralmente.

Moralidad sin mente

Sin embargo, no todos ven el asunto del mismo modo. En su análisis, Véliz discute la opinión mantenida por Luciano Floridi y J. W. Sanders (2004). Para ellos, los "agentes artificiales" pueden ser capaces de un comportamiento (in) moral, y por eso debemos extender nuestra noción de agente moral. Creen que sería posible hablar de agentes morales que no tuvieran libre albedrío ni consciencia o mente. Ellos llaman a esto "moralidad sin mente" (*mind-less morality*). La idea central que desarrollan es que en principio puede haber agentes artificiales que no tengan responsabilidad moral plena (*responsibility*), pero sí tengan obligación de rendir cuentas, de responder por las consecuencias (*accountability*). En español usamos el término "responsabilidad" para ambas cosas, pero podemos traducir *accountability* mediante las expresiones que acabamos de usar. La diferencia entre los dos tipos de responsabilidad puede verse con un ejemplo sencillo. Si un menor coge el coche de sus padres y tiene un accidente que causa destrozos en otro coche, él es el responsable moral de su acción, pero serán sus padres quienes deberán rendir cuentas ante la autoridad. El menor sería un caso de agente con responsabilidad, pero sin capacidad para rendir cuentas. Lo que Floridi y Sanders plantean es que los agentes artificiales, como los padres de este ejemplo, podrían tener obligación de rendir cuentas sin tener por ello responsabilidad.

Ellos consideran que, desde un cierto nivel de abstracción, para que un agente sea considerado un agente moral basta con que dé lugar a acciones moralmente calificables, y una acción es moralmente calificable si causa un bien o un mal moral. Por lo tanto, si un agente artificial es causa de un bien o un mal moral, es un agente moral y debe responder por ello (*accountability*), aunque no pueda ser considerado como moralmente responsable, por carecer de los estados intencionales adecuados. Pensar de otro modo, según nos dicen, sería antropocentrismo. Pero la conclusión de que se trata de una noción de responsabilidad más cercana a lo legal que a lo moral resulta tentadora.

Creo que Véliz tiene razón al señalar que estas condiciones son insuficientes para atribuir agencia moral, tanto más cuanto que el concepto de autonomía que manejan, como cambios de estado flexibles y siguiendo reglas internas que no están controladas externamente, está también lejos de lo que suele entenderse por tal en el discurso ético. Esas reglas o normas, al fin y al cabo, les han sido impuestas a los agentes artificiales que tenemos de forma externa, sin que tengan capacidad para rebelarse contra ellas, y nadie sabe si en el futuro los habrá que puedan dárselas a sí mismos o rechazar las que se les han dado y sustituirlas por otras. Por no hablar de los aspectos emocionales de la agencia moral, o los relacionados con la autorreflexión, el sentimiento de culpa y el arrepentimiento, que aquí quedan fuera.

Pero ¿qué pasaría si alguna vez tuviéramos sistemas de IA con una inteligencia similar a la humana, es decir, con inteligencia artificial general (AGI, por sus siglas en inglés) y con sentimientos similares a los humanos? Es una hipótesis muy especulativa, puesto que, pese a los discursos que se nos lanzan con frecuencia, conseguir AGI no es algo fácil ni previsible, e incluso aunque se consiguiera, estaríamos lejos aún de que las máquinas pudieran tener sentimientos o siquiera que pudieran experimentar algún tipo de placer o de dolor.

Asumamos, sin embargo, a modo de exploración que algún día será así, que tendremos máquinas que autodeterminan su conducta atendiendo a razones y particularmente a razones morales; máquinas capaces de deliberación moral y de experimentar sufrimiento. Es de suponer entonces que tales máquinas deberían ser consideradas agentes y pacientes morales y que quizá deberíamos concederles ciertos derechos (Gordon y Pasvenskiene 2021). No podrían ser tratadas a nuestro antojo, sino que tendríamos ciertas obligaciones morales y legales hacia ellas. Así como hemos ido ampliando el círculo de la consideración moral para incluir en él a los animales, si tales máquinas existieran alguna vez habría que incluirlas también en dicho círculo. Cabría incluso la posibilidad de que, si fueran más inteligentes que los

seres humanos y más sensibles aún al placer y al sufrimiento, merecieran mayor consideración moral, a no ser que estuviéramos dispuestos a reevaluar y reajustar algunos de los principios éticos más arraigados.

Se dirá, y yo me sumaría a la objeción, que no hay ahora algoritmos capaces de sufrimiento y que eso del sufrimiento y del placer es algo propiamente animal, puesto que depende de la fisiología de un cuerpo biológico, con sus hormonas y neurotransmisores, en el que se producen las reacciones correspondientes en los órganos de dicho cuerpo. En todo caso, lo que queremos subrayar aquí es que, si por hipótesis tuviéramos sistemas así en el futuro, habría buenas razones para considerarlos agentes morales. Probablemente pocas personas le negarían consideración moral plena a David, el niño robot protagonista de la película de Spielberg A. I. *Inteligencia Artificial*. No obstante, es una posibilidad tan remota, tan de ciencia ficción, que me parece un esfuerzo inútil el intentar concienciar ahora acerca de ello.

En este empeño por despejarles a las máquinas inteligentes el camino de la moralidad no faltan tampoco quienes creen que no es necesario atribuir emociones y sufrimiento a los agentes artificiales para considerarlos al menos como pacientes morales. John Danaher (2020), por ejemplo, cree que para que las máquinas inteligentes puedan tener consideración moral basta con que se comporten de forma análoga (pero no necesariamente de forma igual) a aquellos seres que la tienen (humanos y, en cierta medida, animales). Basta, por ejemplo, con que su conducta sea similar a la de alguien que experimenta dolor para que se las trate como si realmente sintieran dolor. Es lo que llama "conductismo ético" (*methodological behaviourism*). La razón principal que da para sostener algo así es que no podemos tener acceso directo en ningún ser a las propiedades metafísicas que suelen aducirse para asignarle consideración moral, como la consciencia o las altas capacidades cognitivas o la existencia de intereses. Solo podemos inferir esas propiedades a partir de la conducta, y eso es lo que –según él– hacemos en la vida diaria.

Es fácil ver que la clave aquí está en si debemos aceptar que una simulación del sufrimiento, por buena que sea, constituye de hecho una experiencia consciente de sufrimiento. Si no es así, sostener que tenemos alguna obligación moral de evitar un sufrimiento simulado sería tanto como decir que hay que compadecer al futbolista que se tira al césped fingiendo una patada en la espinilla que nadie cree que realmente existiera o al actor que representa en un escenario la muerte de su personaje. ¿Estaríamos excluyendo así del círculo de la moralidad un caso dudoso de forma errónea o censurable, como sostiene Danaher? ¿Estaríamos ante una situación comparable al rechazo tradicional del estatus moral a los animales? No, si asumimos, como creo que debe hacerse, que el mero despliegue de una conducta análoga, sin nada que garantice que hay sufrimiento real en un sentido comparable al de los animales, no es moralmente significativo. Porque, a diferencia de lo que sugiere Danaher, sin saber qué mecanismos son los causantes de la conducta del agente artificial y qué tienen en común con los mecanismos biológicos que generan la sensación de dolor, no hay ninguna razón para atribuirle dolor o sufrimiento al sistema de IA cuya conducta estemos considerando. El mero análisis de la misma no es suficiente, tal como pretende el conductismo metodológico.

Implementar un comportamiento moral

Finalmente, se habla cada vez más de introducir valores morales en los sistemas de IA, no solo como prevención ante una hipotética AGI futura o como control posible de inteligencias artificiales muy superiores a la nuestra (el problema de la alineación), sino porque hay ya sistemas en desarrollo, como los robots asistenciales o los vehículos autónomos, que necesitarían de ciertas directrices éticas en caso de accidente (recuérdense los famosos dilemas del tranvía). Pero ni siquiera hay acuerdo acerca de cómo podría lograrse la implementación de un comportamiento moral en las máquinas inteligentes (Cave *et al.* 2019). ¿Podrá, por ejemplo, una IA valorar alguna vez que una razón dada para motivar una acción concreta puede ser más débil desde el punto de vista de su eficacia que la razón para llevar a cabo una acción distinta, pero más fuerte desde el punto de vista de su mayor importancia moral o su mayor justicia y que, por tanto, esa debe ser la acción a realizar? ¿Podrá tener la virtud de la prudencia, la capacidad para adecuar los grandes principios éticos a los casos concretos en contextos variables y complejos? ¿Habría que introducirle algunos principios morales universales (si conseguimos establecerlos) o estos serían demasiado abstractos y tendría que aprender del comportamiento moral de los seres humanos? ¿Es deseable que lo haga, dadas las deficiencias morales de nuestra propia conducta? ¿Podría distinguir entre lo que es habitual hacer y lo que es moralmente correcto?

Y si alguna vez se consiguieran máquinas capaces de aplicar reglas morales en su conducta, seguiría habiendo un problema para atribuirles responsabilidad. El que los objetivos de estas inteligencias artificiales fueran conformes con el deber moral e incluso el que pudieran realizar razonamientos morales y estuvieran éticamente alineadas no implicaría necesariamente que hubiera que considerarlas como agentes morales plenos, como miembros de una comunidad moral. No lo serían mientras que su acción no cumpliera con los requisitos mínimos de autonomía en el sentido de voluntad libre y consciencia que hemos visto que suelen reclamarse para ello; mientras se limitaran, por utilizar la distinción kantiana, a obrar conforme al deber pero no por mor del deber. Esta es, en el fondo, la razón que lleva a Robert Sparrow a decir que "incluso las 'inteligencias artificiales fuertes', como las que Kurzweil, Brooks, Moravec y otros discuten, es probable que

habiten una 'zona gris' en medio de la gama entre los sistemas que están determinados y los que son agentes morales plenos, precisamente porque es difícil ver cómo se les podría considerar responsables de sus actos" (Sparrow 2007, p. 66). En la atribución de responsabilidad moral no solo importa que se tomen decisiones que encajen con los preceptos morales, sino también cómo se tomen esas decisiones. Aunque se les llegue a llamar "máquinas éticas", como algunos proponen, no serían en todo caso máquinas responsables.

Ahora bien, en tanto que las máquinas no sean agentes morales plenos, la responsabilidad moral de sus acciones debe recaer sobre los seres humanos, incluso (*pace* Sparrow) en los casos de las armas autónomas (en cuyo análisis ético Sparrow fue pionero), por indirecta que esa responsabilidad humana pueda llegar a ser y por difícil que se vuelva su atribución en la práctica. Al menos algunas personas tendrán la responsabilidad de haber decidido su uso para matar a otras, como cabría hacer en el ejemplo que Sparrow menciona de los niños soldados. No son ellos los responsables morales de sus acciones (o lo son en muy pequeña medida), sino las personas adultas que les lavan el cerebro, los adiestran y los mandan a la guerra. Y aquí no importa demasiado que no hayan tenido un control directo sobre las acciones concretas de las que deben ser hechos responsables. Quizá en el asunto de las armas autónomas tenga mucho sentido hablar de responsabilidades morales colectivas, porque son organizaciones de personas las que suelen tomar las decisiones al respecto y esto incluye desde los ingenieros que las diseñaron hasta los militares que las usaron, pasando por los políticos que las aprobaron y fomentaron, aunque este es un tema controvertido. Todo esto puede sonar ingenuo, porque esas organizaciones están encantadas con lo que hacen y no van a dejar de hacerlo, pero no se trata de convencerlas de que hacen mal, sino de aclarar a quiénes hay que atribuir la responsabilidad, la asuman o no. ~

REFERENCIAS

Cave, S. *et al.* (2019), "Motivations and risks of machine ethics", *Proceedings in the* IEEE, 107(3), pp. 562-574.

Danaher, J. (2020), "Welcoming robots into the moral circle: A defence of ethical behaviourism", *Science and Engineering Ethics*, 26, pp. 2023-2049.

Floridi, L. y J. W. Danders (2004), "On the morality of artificial agents", *Minds Mach*, 14, pp. 349-379.

Furlough, C., T. Stokes y D. J. Gillian (2021), "Attributing blame to robots: I. The influence of robot autonomy", *Human Factors*, 63(4), pp. 592-602.

Gordom, J.-S. y A. Pasvenskiene (2021), "Human rights for robots? A literature review", *AI and Ethics*, 1, pp. 579-591.

Hern, A. (2017), "Give robots 'personhood' status, EU committee argues", *The Guardian*, 12 de enero, https://www.theguardian.com/technology/2017/jan/12/give-robots-personhood-status-eu-committee-argues

Sparrow, R. (2007), "Killer robots", *Journal of Applied Philosophy,* 24(1), pp. 62-77.

Véliz, C. (2021), "Moral zombies: why algorithms are not moral agents", *AI & Society*, 36, pp. 487-497.

ANTONIO DIÉGUEZ es catedrático de lógica y filosofía de la ciencia en la Universidad de Málaga y miembro de número de la Academia Malagueña de Ciencias.