teorema
Vol. XLI/1, 2022, pp. 141-149
ISNN 0210-1602
[BIBLID 0210-1602 (2022) 41:1; pp. 141-149

Retos y desafíos éticos ante la inteligencia artificial

Germán Massaguer Gómez

Ética de la Inteligencia artificial, de MARK COECKELBERGH, MADRID, CÁTEDRA, 2021, 183 pp.

1. INTRODUCCIÓN

La presente transformación, paulatina y a veces imperceptible, de nuestro mundo a raíz del desarrollo y la integración de la Inteligencia Artificial (IA) ha generado numerosos debates. Estos se hacen todavía más acuciantes debido a la omnipresencia de la IA en los ámbitos sociales, laborales, económicos, políticos o judiciales, entre otros. La creciente importancia de la IA ha desencadenado una serie de efectos (positivos y negativos) con suficiente trascendencia para que contemos ya con una enorme producción de literatura académica centrada en sus implicaciones éticas. Pero no solo la integración de esta tecnología en nuestra vida cotidiana hace que los debates sean múltiples: también la promesa de lo que la IA puede llegar a ser causa desasosiego, anunciando dilemas y controversias morales para los que parecemos carecer de respuestas. Para muchos, las potencialidades de la IA plantean un riesgo tan elevado para la humanidad que, independientemente de las probabilidades de que conlleve una catástrofe irreversible, debemos plantear con mucha anterioridad y precaución cómo prevenir dichos riesgos.

En el libro que comentamos, Mark Coeckelbergh subraya reiteradamente que las cuestiones que rodean a la IA son muy complejas y que no sabemos, ni podemos saber a ciencia cierta, en qué va a derivar una sociedad en la que la ubicuidad de esta tecnología se sume al creciente desarrollo científico. Existen dos escenarios futuros contrapuestos fácilmente imaginables: por un lado, la IA podría ayudarnos a solucionar problemas actuales (ajenos a esta tecnología), como el hambre, la guerra, o la crisis climática; por otro lado, no solo podría agravar dichos problemas, sino que podría generar nuevos si, por ejemplo, su desarrollo no se

hace de manera sostenible o sus beneficios solo se hacen patentes para unas pocas personas [p. 154]*. La pregunta a la que nos enfrenta el desarrollo de la IA no es otra que la pregunta por el tipo de sociedad que queremos construir. Preguntar cómo queremos que sea el mundo no es, desde luego, una pregunta nueva. Sin embargo, parece que estamos en una de esas disyuntivas históricas en la que la respuesta resulta insoslayable. En esta nota crítica nos centraremos en algunos problemas a los que se enfrentan las iniciativas que pretenden aplacar las posibles consecuencias negativas de la IA, atendiendo a las dificultades y a algunas soluciones que podrían minimizas los riesgos asociados a la IA. Estos problemas están en el centro del libro que comentamos, en el que Coeckelbergh busca, con una narrativa original y agradable, introducir al lector en los conceptos y problemas éticos de la IA.

II. Este libro

El libro de Coeckelbergh reúne prácticamente todas las cuestiones que conectan ética e IA. De hecho, se trata en gran parte de un libro introductorio al tema, que analiza cada una de las discusiones más relevantes en este ámbito. En palabras del autor:

> [Este libro] busca dar al lector una buena visión general de los problemas éticos que surgen en conexión con la IA entendida de forma amplia, desde las influyentes narrativas sobre el futuro de la IA y las cuestiones filosóficas sobre la naturaleza y el futuro de lo humano, hasta las preocupaciones éticas sobre la responsabilidad, el prejuicio y el modo de lidiar con cuestiones prácticas del mundo real que surgen de la tecnología mediante la aplicación de políticas (preferiblemente antes de que sea demasiado tarde) [p. 22].

Antes de centrarnos en aspectos concretos, conviene tener una panorámica general de la estructura de este libro. En el primer capítulo se introduce el tema analizando la ya existente permeabilidad de la IA en nuestra cotidianeidad. En el capítulo segundo se nos introduce al fenómeno de la superinteligencia artificial. En el tercer capítulo se discute si realmente es posible que llegue a existir una IA general (equivalente a una mente humana), trayendo a colación las diferentes posiciones enfrentadas respecto a este debate. En el capítulo cuarto se habla sobre el estatus moral de las máquinas, preguntando si se puede (o podría llegar a) hablar de agencia moral y conciencia en entes artificiales. En el capítulo quinto se define la IA como "una inteligencia desplegada o simulada por un código (algorítmico) o por máquinas" [p. 61]. Esta definición continúa en el capítulo si-

guiente, en el que se habla sobre aprendizaje automático y el potencial impacto de esta vertiente de la IA. En el capítulo séptimo se tratan los problemas vinculados a la privacidad, en el octavo a la responsabilidad y en el noveno a los sesgos. Estos capítulos son ciertamente descriptivos y aunque la posición del autor se deja entrever en algunos lugares, el objetivo de estos capítulos es exponer de forma fidedigna las diferentes perspectivas predominantes en cada discusión. Pero el libro no se detiene en un mero estado de la cuestión, sino que, a partir del análisis de las políticas de actuación que ya han planteado gobiernos, corporaciones, universidades, etc., expone una serie de propuestas acerca de cómo debemos abordar el presente y el futuro inmediato que nos espera. Estos temas se tratan en los tres últimos capítulos.

En esta nota, nos centraremos en comentar los debates planteados precisamente en estos últimos tres capítulos. Esto no sólo supone evaluar cómo son las políticas en sí, ni cómo deberían ser, sino que requiere pensar en la implicación y efectividad de las mismas y en la naturaleza de los problemas que pretenden resolver. Alrededor de todas las discusiones filosóficas sobre la IA se encuentra la pregunta de cuál es la labor de la ética. Los retos que plantea esta tecnología, que inciden en cuestiones tan fundamentales como qué es la moral o qué es la mente, parecen desbordar en ocasiones los métodos de la ética práctica al viejo estilo. Por tanto, preguntarnos dónde está el lugar de la ética es equivalente a preguntarnos asimismo cómo debemos enfocar los problemas que surgen al hilo de la IA. Existe un riesgo, e incluso un miedo, a que el análisis ético de la IA acabe derivando en unas meras listas de principios, normalmente aplicados desde diversos 'comités de buenas prácticas'. Esto implicaría el olvido de la reflexión sobre la vida, el trabajo, las relaciones humanas, nuestras necesidades y nuestros miedos. Estas reflexiones deben estar en el centro de los debates sobre la IA, pues, como hemos mencionado, esta tecnología cohabita con nosotros en nuestro día a día, influyendo sobre nuestras maneras de actuar, pensar y tomar decisiones – desde las más nimias a las más trascendentes.

¿Cómo debemos entonces enfocar la evaluación ética de la IA? En palabras del autor, "[las] cuestiones que necesitan responderse proponiendo políticas de actuación" no sólo son "*qué* debe hacerse, sino también *por qué, cuándo, cuánto, por parte de quién*, y cuáles son la *naturaleza, extensión y urgencia del problema*" [p. 123]. Gran parte de las propuestas del autor conciernen a dichas políticas de actuación, de las cuales varias son analizadas y resumidas en los mencionados últimos capítulos. Analizaremos a continuación uno de los problemas que considero de especial re-

levancia a la hora de programar propuestas de actuación política y cuya solución es tremendamente compleja: *la prevención de daños*. Como veremos, para discutir este dilema, debemos volver sobre gran cantidad de cuestiones presentes en el libro de Coeckelbergh.

III. PREVENCIÓN DE DAÑOS: DOS MODELOS *EX ANTE*

Cuando la discusión se centra en la naturaleza de las propuestas de actuación política frente a la IA, conviene tener presente la necesidad de elaborar dichas propuestas anticipando las posibles consecuencias negativas. Esta necesidad de planificación y prevención, en sí positiva en cualquier ámbito de la vida, se hace evidente en un caso como el que nos ocupa, puesto que los riesgos potenciales son altísimos. Para prevenir estos riesgos y garantizar la seguridad, desde la filosofía de la tecnología se incide en tomar precauciones [Hansson (2018)]. Es importante, por tanto, tener en cuenta el principio de precaución, que nos conmina a reflexionar sobre los riesgos posibles de una tecnología cuando no tenemos suficiente conocimiento científico respecto a dicha tecnología [Schneider (2020), p. 453]. Las consecuencias posibles no siempre son imaginables de antemano. Además, las repercusiones en la sociedad son en muchos casos imprevisibles y pueden ser negativas a pesar de que las intenciones de quienes las diseñan, producen o comercializan sean buenas. Por lo tanto, el desafío consiste en legislar antes de que ocurran los daños [p. 124]. La complejidad reside en que prever los daños no es una tarea siempre posible [p.139]. Es prácticamente imposible acertar con completa exactitud las implicaciones reales que en la práctica pueda tener la IA.

Tenemos, por lo tanto, la tarea de pensar y actuar *ex ante*. Este discurso, que se vincula con el que alerta sobre los problemas que acarrea la IA y que podría acarrear en el futuro, está presente en gran parte de la literatura que trata sobre la ética de la IA. Hay, por lo tanto, un consenso sobre la necesidad de investigar para garantizar la seguridad[1] – consenso que, evidentemente, no se extiende a las compañías que quieren sacar sus productos al mercado cuanto antes y sin trabas legales[2].

Ante el escenario de prever lo imprevisible, ¿con qué estrategias contamos? Coeckelbergh apunta que "una forma de mitigar este problema es construir escenarios hipotéticos en torno a futuros conflictos de índole ética" [p. 139]. Esta estrategia se queda, no obstante, en un plano excesivamente teórico y no se aclara muy bien cómo podría funcionar. Evidentemente, es necesario pensar futuros escenarios hipotéticos, pero,

de nuevo, parece que las consecuencias previsibles alcanzarán únicamente tanto como nuestra imaginación.

Existen, sin embargo, algunas propuestas prácticas para combatir la incertidumbre. Una de ellas es instaurar escenarios de experimentación. Semejante a lo que ocurre en el campo de la biomedicina y la farmacia, puede que las nuevas tecnologías tengan que pasar por una serie de pruebas antes de ser comercializadas. Esto permitiría discernir qué impacto tienen sobre los seres humanos y, subsecuentemente, preparar una serie de leyes que aseguren que las personas no verán violados sus derechos. Un ejemplo ilustrativo de cómo esto podría funcionar ha sido propuesto por John Danaher en relación con los robots sexuales: ante los plausibles daños que su uso podría generar en la sociedad [por ejemplo, reforzar la desigualdad de género y la cultura de la violación [Gutui (2012)], se debe llevar a cabo una aproximación experimental, haciendo pruebas con un limitado grupo de personas y recabando información [Danaher (2017), pp.120-122]. Desde luego, a pesar de que se instaurasen unos comités de evaluación, nunca es posible entrever al 100% los efectos que una tecnología pueda tener a largo plazo sobre la sociedad en su conjunto. Sin embargo, considero que la estrategia de Danaher mitigaría sin lugar a dudas posibles daños.

Evidentemente, este tipo de estrategias tendrán una fuerte oposición. Mencionemos al menos dos problemas. El primero es el hecho de que esto supone una desaceleración del avance tecnológico. Los intereses de una compañía cuyo trabajo se centra, retomando el caso de Danaher, en desarrollar robots sexuales, entrarían en conflicto con esta propuesta, por una parte, si tiene que posponer su comercialización debido a la obligación de pasar una serie de fases de experimentación; por otra parte, si luego tiene que enfrentarse a la aprobación por parte de un comité – que también podría, sin duda, obligar a efectuar cambios en el diseño[3].

El segundo problema es la narrativa competitiva o, como lo ha bautizado Zuboff, el 'Síndrome China' [Zuboff (2019), p. 388]: desde el punto de vista occidental se ha empezado a considerar que este tipo de precauciones son trabas que nos harán perder la carrera tecnológica frente a otras potencias que dan menos importancia a las consideraciones éticas, como China o Rusia [p.131]. En otras palabras, se valora que no pongamos restricciones al avance tecnológico porque los citados países lo van a hacer igualmente – esto es especialmente relevante en cuestiones como las que conciernen al armamento autónomo.

Estos argumentos, sin embargo, carecen de fuerza suficiente como para desechar la idea de una experimentación preventiva. Con respecto al

primero, hay que considerar que los intereses comunes de la sociedad están por encima de los intereses económicos de las empresas. Con respecto al segundo, si el hecho de que otra persona haga algo no le otorga a ese acto ninguna validez moral, no parece que la agencia de entidades más complejas pueda validar en sí misma la dimensión moral de sus políticas. El desarrollo de armamento autónomo, si fuese moralmente incorrecto, lo sería independientemente de quién lo desarrollase.

Por lo tanto, la propuesta es emplear estrategias similares a las presentes en el campo de la biomedicina a la hora de permitir que ciertas tecnologías se introduzcan en nuestra sociedad. Esto puede pasar por fases de experimentación que nos permitan analizar con más detalle los efectos que tienen las tecnologías sobre los individuos, su forma de interrelacionarse entre ellos y con el mundo. A partir de esta información y su subsecuente interpretación estaremos más y mejor capacitados para elaborar propuestas de políticas de actuación que aseguren que los derechos de los ciudadanos y las ciudadanas no sean violados. Esta solución, aunque difícil de implementar por las cuestiones ya expuestas, nos ayudará a poder controlar y prevenir más eficazmente escenarios no deseados. Hemos visto dos modelos de actuación ex ante: los escenarios hipotéticos y la experimentación previa. Aunque ambos pueden ser eficaces, el segundo atiende mejor al principio de precaución, puesto que nos puede aproximar más a certezas científicas que eviten riesgos.

IV. Conclusión

El libro que comentamos trata una cantidad enorme de cuestiones concernientes a la ética de la IA, entre los cuales se encuentra la que hemos explorado y analizado con detenimiento. También incluye una serie de propuestas sobre cómo encarar los desafíos que van a marcar nuestro presente y futuro. Considero oportuno finalizar esta nota crítica resaltando algunas de estas propuestas, tremendamente relevantes y probablemente decisivas para afrontar los problemas que suscita la IA.

La principal propuesta es metodológica. Quizás la forma en la que hagamos ética de la IA pase por replantear el enfoque con el que la concebimos, trayendo a colación cuestiones que no se suelen tener en cuenta. Estas son las propuestas del autor que se dejan entrever a lo largo de la obra: en primer lugar, abogar por la gestión inclusiva. Esto quiere decir que a la hora de encarar cómo diseñar una tecnología (un algoritmo, un robot, etc.) se deben tener en cuenta los puntos de vista de todas las partes implicadas, poniendo especial énfasis en aquellas personas cuyas vidas

van a ser efectivamente alteradas [p.140]. Si la meta de la ética de la IA es mejorar el mundo, debemos asegurarnos de que sus propuestas beneficien a todas las personas. Y no sólo se habla de las personas *humanas*. Coeckelbergh va más allá y aboga por una ética de la IA no antropocéntrica, que tenga en cuenta su impacto sobre otros seres vivos y el medio ambiente [p. 152]. Esto pasa, entre otras cosas, por apostar por una IA sostenible.

Por último, me gustaría exponer la visión del autor respecto al papel de los filósofos y las filósofas. Coeckelbergh apuesta por una ética positiva e interdisciplinar. Esto implica que no se trata de elaborar una serie de principios éticos inalterables que ingenieros e informáticos tengan que cumplir. Nuestra misión no es poner trabas, restricciones y dedicarnos a distraer. Debemos ser capaces de elaborar una visión positiva de cómo queremos que sea el mundo, no a partir de prohibiciones sino de aspiraciones. "Al reducir los riesgos, la ética y la innovación responsable apoyan el desarrollo sostenible a largo plazo de las empresas y las sociedades. Sigue siendo un reto convencer a todos los actores en el campo de la IA […] de que este es ciertamente el caso" [p. 144]. Para ello, uno de los mecanismos cuya necesidad es cada vez más patente, es la interdisciplinaridad. Es imprescindible que haya comunicación entre los diferentes campos que trabajan la IA. Los filósofos y las filósofas debemos de acercarnos al mundo científico y comprender mejor cómo funcionan los mecanismos tecnológicos para poder aportar de una manera proactiva y positiva, no restrictiva y autoritaria. Yo añadiría, para concluir, que el ancestral conocimiento que nos otorga la filosofía sobre nuestra manera de comprender el mundo, relacionarnos entre nosotros, expresar y reaccionar ante emociones, puede ayudarnos a comprender mejor las necesidades que tenemos y de qué forma la IA puede ayudarnos a satisfacerlas y mejorar, por tanto, nuestra existencia en un planeta igualitario y sostenible.

Departamento de Humanidades
Universidad Carlos III
Calle Madrid 126, 28903 Getafe, Madrid
E-mail: germanmassaguer@gmail.com

NOTAS

* Las indicaciones entre corchetes de número de página sin ningún otro añadido se refieren al libro de Coeckelbergh objeto de esta nota.

[1] En enero de 2015 se reunieron en Puerto Rico grandes personalidades del campo de la IA como Margaret Boden, Nick Bostrom, Francesca Rossi o

Stuart Russell, entre otros muchos. En esta conferencia se incidió de manera reiterada precisamente sobre esta necesidad de reflexionar y actuar *ex ante* [Tegmark (2017), pp.35-37].

[2] Shoshana Zuboff, en su libro *The Age of Surveillance Capitalism,* muestra innumerables ejemplos sobre cómo compañías como Google o Facebook muestran explícita y abiertamente su descontento con las leyes e incluso la democracia, ya que las consideran una imposición que impide el libre progreso y avance de las tecnologías y el mercado. Paradójicamente, estas compañías también proponen una serie de políticas de actuación para salvaguardar la seguridad y los derechos de sus usuarios.

[3] Sin embargo, debemos tener cierta seguridad sobre los potenciales problemas y asegurar que no se basan en falsos miedos injustificados (provenientes, por ejemplo, de narrativas de ciencia ficción que se alejan de la realidad del avance tecnológico), puesto que esto podría desembocar, entre otras, en la dedicación de recursos, tiempo y atención a situaciones que no lo requieren.

REFERENCIAS

BODEN, M, (2018), *Artificial Intelligence. A Very Short Introduction*; Oxford: Oxford University Press.

DANAHER, J., (2017), 'The Symbolic-Consequences Argument in the Sex Robot Debate'; en *Robot Sex: Social and Ethical Implications,* Danaher, J., McArthur, N. (eds.) Cambridge: MIT Press, pp. 103-132.

GUTIU, S. (2012), "Sex Robots and Roboticization of Consent", *We Robot Law Conference Miami*; disponible en: <*http://robots.law.miami.edu/wp-content/uploads/2012/01/Gutiu Roboticization_of_Consent.pdf*>.

HANSSON, S.O. (2018), "Risk", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), disponible en <*https://plato.stanford.edu/entries/risk/*>.

LIAO, M., (2020), 'A Short Introduction to the Ethics of Artificial Intelligence'; en *Ethics of Artificial Intelligence*, Matthew S. Liao (ed.) Nueva York: Oxford University Press.

MÜLLER, V., (2020), 'Ethics of Artificial Intelligence and Robotics'; *The Stanford Encyclopedia of Philosophy*; Edward N. Zalta (ed.), disponible en <*https:// plato.stanford.edu/entries/ethics-ai/*>.

SCHNEIDER, S. (2020), 'How to Catch an AI Zombie: Testing for Consciousness in Machines'; en *Ethics of Artificial Intelligence*, Matthew S. Liao (ed.) Nueva York: Oxford University Press.

TEGMARK, M., (2017), *Life 3.0.*; Great Britain: Penguin Random Books.

ZUBOFF, S., (2019), *The Age of Surveillance Capitalism,* Londres: Profile Books.

ABSTRACT

Artificial Intelligence gives rise to some of the most pressing ethical debates of our time, concerning responsibility, agency, or autonomy, amongst other issues. This critical note discusses the way in which ethics should deal with the new moral dilemmas arising from this technology.

KEYWORDS: *Artificial Intelligence, Positive ethics, precautionary principle, metaethics*.

RESUMEN

La inteligencia artificial es uno de los fenómenos más relevantes de nuestro tiempo. En torno a ella hay numerosos debates (sobre responsabilidad, agencia o autonomía, entre otros). En esta nota crítica se discute cómo debe actuar la ética ante los nuevos dilemas morales que surgen de esta tecnología.

PALABRAS CLAVE: *inteligencia artificial, ética positiva, principio de precaución, metaética*.