Daimon. Revista Internacional de Filosofía (en prensa): trabajo aceptado para publicación tras revisión por pares doble ciego.
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico)
http://dx.doi.org/10.6018/daimon.508771

¿Automatizando la mejora moral humana? La inteligencia artificial para la ética. (Nota crítica sobre LARA, F. y J. SAVULESCU (eds.) (2021), *Más (que) humanos. Biotecnología, inteligencia artificial y ética de la mejora*. Madrid: Tecnos)

Automating human moral enhancement? Artificial intelligence for ethics. (Critical note on LARA, F. y J. SAVULESCU (eds.) (2021), *Más (que) humanos. Biotecnología, inteligencia artificial y ética de la mejora*. Madrid: Tecnos).

JON RUEDA[*]

Resumen: ¿Puede la inteligencia artificial (IA) hacernos más morales o ayudarnos a tomar decisiones más éticas? El libro *Más (que) humanos. Biotecnología, inteligencia artificial y ética de la mejora*, editado por Francisco Lara y Julian Savulescu (2021), puede inspirarnos filosóficamente sobre este debate contemporáneo. En esta nota crítica, contextualizo la aportación general del volumen y analizo los dos últimos capítulos de Monasterio-Astobiza y de Lara y Deckers, quienes argumentan a favor del uso de la IA para hacernos mejores agentes morales. El objetivo es ampliar y matizar críticamente algunas cuestiones clave sobre cómo la IA puede asistirnos para tomar decisiones más éticas.

Recibido: 27/01/2022. Aceptado: 21/02/2022.

[*] Jon Rueda es investigador predoctoral en filosofía, La Caixa INPhINIT Fellow, en la Universidad de Granada. Su investigación se sitúa en la intersección entre la bioética, la ética de las tecnologías emergentes y la filosofía de las innovaciones biomédicas. El autor agradece los comentarios de Belén Liedo a una versión previa del manuscrito. Esta investigación ha sido financiada por los proyectos ETHAI+3 ("Ética digital. La mejora moral desde un uso interactivo de la inteligencia artificial") de la Agencia Estatal de Investigación (PID2019-104943RB-I00) y SOCRAI3 ("Mejora moral e inteligencia artificial. Aspectos éticos de un asistente virtual socrático") del programa FEDER/Junta de Andalucía (B-HUM-64-UGR20), y un contrato La Caixa INPhINIT Retaining (LCF/BQ/DR20/11790005). Publicaciones recientes: Rueda, J. (2021). Socrates in the fMRI Scanner: The Neurofoundations of Morality and the Challenge to Ethics. *Cambridge Quarterly of Healthcare Ethics*, 30 (4), 606-612. https://doi.org/10.1017/S0963180121000074. Rueda, J. (2021). From self-determination to offspring determination? Reproductive autonomy, procrustean parenting, and genetic enhancement. *Theoria (Stockholm)*. https://doi.org/10.1111/theo.12349. Correo electrónico: ruetxe@ugr.es.

Palabras clave: Mejora moral, inteligencia artificial, inteligencia artificial para la ética, mejora humana, ética, ética de la inteligencia artificial.

Abstract: Can artificial intelligence (AI) make us more moral or assist us in making more ethical decisions? The book *Más (que) humanos. Biotecnología, inteligencia artificial y ética de la mejora*, edited by Francisco Lara and Julian Savulescu (2021), can provide philosophical inspiration on this contemporary debate. In this critical note, I contextualize the overall contribution of the volume and analyse the last two chapters by Monasterio-Astobiza, and Lara and Deckers, who argue for the use of AI to make us better moral agents. The objective is to expand and critically qualify some key questions about how AI can assist us in making more ethical decisions.

Key-words: Moral enhancement, artificial intelligence, AI for ethics, human enhancement, ethics, AI ethics.

1. Introducción

Hay dos razones por las que el concepto mismo de 'mejora humana' puede parecer presuntuoso. Por un lado, tilda de novedad algo que llevamos haciendo desde los albores de la humanidad: mejorarnos como especie. Imaginemos que un *Homo sapiens* del Pleistoceno y una lectora de esta nota crítica pudiesen contemplarse mutuamente. ¿Pensarían ambos que están ante alguien de su misma especie? Supongamos que nuestra sapiens contemporánea es una persona alfabetizada que lee estas páginas mediante gafas que corrigen las deficiencias de su visión y desde la pantalla de su ordenador portátil —un aparato que funciona como la extensión de su mente (cf. Clark y Chalmers, 1998)—. Nuestra lectora es capaz de comunicarse en varios idiomas, toca instrumentos musicales complejos, potencia su físico mediante la dieta y el entrenamiento y tiene una esperanza de vida cuya longevidad la convertiría en una Matusalén del Pleistoceno. Podemos sacar dos conclusiones de esta comparación odiosa: (*a*) los cambios intra-específicos (algunos de ellos cambios *a mejor*) pueden ser abrumadores (Jones, 2006; Liedo y Rueda, 2021), y, (*b*) la mejora humana es tan antigua como la humanidad misma (Harris, 2007; Buchanan, 2011).

Por otro lado, el término 'mejora humana' genera recelos porque está cargado de valor. Parece dar por bueno a nivel normativo aquello que está describiendo. ¿Quién no quiere ir a mejor? ¿Por qué no deberíamos mejorarnos más? ¿Quienes proponen la mejora no están haciendo, al final y al cabo, una oferta que no podemos rechazar? Quien quiera vislumbrar posibilidades de respuesta a estos interrogantes, debería leer el libro *Más (que) humanos: biotecnología, inteligencia artificial y ética*, editado por

Francisco Lara y Julian Savulescu. En su lectura podrá darse cuenta de que hay una novedad genuina en las tecnologías actuales de mejora humana y de que no toda mejora es deseable. Si bien es cierto que los humanos nos caracterizamos por una "aspiración a superarnos" (Lara y Savulescu, 2021, p. 15) —como dicen los editores en la introducción—, es innegable que las tecnologías emergentes abren un abanico de posibilidades sin precedentes en la historia de nuestra especie. Asimismo, estas mejoras no convencionales suscitan una miríada de cuestiones éticas que reclaman atención filosófica aguda. Una de las razones que más anima a la lectura de este volumen colectivo es, precisamente, que en él se encuentran algunos de los aspectos éticos más controvertidos del debate.

La otra carta de invitación a su lectura es, igualmente, la pericia de los autores que contribuyen a esta obra. Los doce capítulos que lo componen —o bien traducciones de artículos en inglés, o bien publicaciones inéditas— son resultado del trabajo de algunos de los filósofos y filósofas más curtidos en el debate de la mejora humana, tanto a nivel nacional como internacional. Sus editores son el puente de colaboración entre el único grupo de investigación español (liderado por Francisco Lara en la UGR) que ha cursado varios proyectos financiados específicos sobre la ética de la mejora humana y el centro académico más puntero del mundo sobre este tema, el *Uehiro Centre for Practical Ethics* (dirigido por Julian Savulescu) de la Universidad de Oxford. Así, este libro viene a satisfacer una demanda nacional en alza sobre esta problemática que es cada vez más apremiante, pero que cuenta con escasos volúmenes en castellano para introducirse al debate general (Diéguez, 2017; Diéguez, 2021) y a los asuntos éticos en particular (Savulescu, 2012; Ortega Esquembre et al., 2015; Bostrom y Savulescu, 2017).

El libro está divido en dos temáticas principales. La primera parte versa sobre la caracterización de la mejora humana y de su relación con debates éticos particulares. Voy a concentrar mi nota crítica en la segunda parte del libro, dedicada a la mejora moral, ya que en ella se encuentran algunas contribuciones que merecen una mayor discusión. El itinerario es el siguiente. Comenzaré con una brevísima introducción al debate de la mejora moral y resumiré las aportaciones de los dos últimos capítulos en los que Aníbal Monasterio-Astobiza, por un lado, y Francisco Lara y Jan Deckers, por otro, defienden el uso de la Inteligencia Artificial (IA) para mejorarnos como agentes morales. En la siguiente sección, desarrollaré tres cuestiones que reclaman más

atención: (*i*) la taxonomía sobre la relación entre la ética y la IA, (*ii*) la justificación de la necesidad de utilizar la IA para la ética y (*iii*) otros potenciales innovadores de la IA para ayudarnos a tomar mejores decisiones de corte normativo.

2. La mejora moral humana

La segunda parte del libro está centrada en la *mejora moral*. Como se puede comprobar en el volumen de Lara y Savulescu, la mejora moral es, sin duda, un debate filosóficamente desafiante (cf. Rueda, 2020). Hay varias razones para tal consideración. En primer lugar, la mejora moral ha suscitado reticencias incluso entre algunos de los defensores más vehementes de otras biomejoras (Agar, 2010, 2015; Harris, 2016). En segundo lugar, hay un desacuerdo genuino sobre en qué consistiría una mejora moral, esto es, sobre qué es la *moral* en la mejora moral (Harris, 2016; Rueda, 2020). Los intentos de definir de manera neutral la mejora moral acaban desenterrando disputas filosóficas sobre qué es la moralidad (o qué debería ser), oscilando entre la vertiente descriptiva y normativa del concepto (Raus et al., 2014). En tercer lugar, se podría añadir una tercera particularidad, sugerida en la introducción de esta obra (Lara y Savulescu, 2021, p. 20), que no ha sido lo suficientemente atendida. ¿Qué es lo que convierte tan significativa a esta mejora respecto a las demás? ¿Por qué genera tantos reparos? Una respuesta plausible es que esta mejora tiene como objetivo potenciar un rasgo fundamental (aunque quizás no exclusivo) de nuestra especie: nuestra condición distintiva de agentes morales (cf. Ayala, 2010). Al fin y al cabo, muchas de las otras biomejoras (como las físicas, las cognitivas, el aumento de la longevidad o el de resistencia a enfermedades) ya las hemos aplicado en animales de laboratorio y algunas también en la ganadería industrial. Lo que es singular en la mejora moral es la particularidad de que los sapiens somos los *principales* animales candidatos a esta mejora —reconociendo, eso sí, que hay quienes afirman que ciertas especies no humanas también podrían contar con cierta moralidad proto-moralidad (Monsó et al., 2018).

Así, como hemos dicho anteriormente, si la historia humana es la historia de la mejora humana, ¿qué lugar ocupa en ella la mejora moral? ¿Es la historia humana también la historia de la mejora moral? Una respuesta afirmativa a esta pregunta no cabe duda que resulta más controvertida. Sin embargo, es cierto que siempre hemos

intentado hacernos más morales por diferentes medios (Rueda y Lara, 2020). La educación (formal e informal), el derecho, el arte y la literatura, la religión, las normas sociales o los incentivos negativos (como los castigos, las prisiones, las multas, etc.) son ejemplos de cómo hemos tratado de influir y condicionar el comportamiento moral humano. En este sentido, en el libro editado por Lara y Savulescu se presentan diversas opciones no tradicionales para ahondar dichas posibilidades y sus respectivas consideraciones ético-filosóficas.

Lara dedica el primer capítulo de este segundo apartado al potencial de la hormona de la oxitocina para mejorar el comportamiento moral humano (traducción de Lara, 2017). Después, Antonio Diéguez y Carissa Véliz (traducción de Diéguez y Véliz, 2019) se preguntan, por otra parte, si la mejora moral podría socavar la libertad humana, en especial "la libertad para caer" que había sugerido Harris (2016). Argumentan que, a pesar de que un agente moralmente mejorado tenga menos opciones para caer (es decir, para obrar moralmente mal), ello no conlleva que la mejora moral haga que los sujetos mejorados sean menos libres. Asimismo, Paloma García Díaz publica una contribución inédita sobre "El neurofeedback y la mejora de la agencia moral". La autora plantea la posibilidad de optimizar neurotecnológicamente la toma de decisiones morales mediante esta interfaz cerebro-ordenador que se basa en entrenar el control de las ondas cerebrales.

Como se ha mencionado, los dos últimos capítulos son los que más nos interesan en esta nota crítica. Ambos suponen un giro innovador, ya que se propone una vía alternativa a la bio- y la neuro-mejora moral, defendiendo el uso la IA para influir de manera proactiva en los procesos de deliberación y decisión de la moralidad. Resumiré brevemente el argumento principal de cada contribución para, en la siguiente sección, añadir ciertos matices que pueden ser enriquecedores para el debate.

Monasterio-Astobiza publica la contribución inédita "Automatizando la toma de decisiones: Inteligencia Artificial y mejora humana". En ella defiende que la idea de crear herramientas computaciones para mejorar la toma de decisiones ha estado presente desde los desarrollos históricos iniciales de la IA. Aplicado al ámbito de la moralidad, la IA podría asistirnos de manera providencial debido a que la racionalidad humana es limitada y abundante en sesgos cognitivos. Las tecnologías computacionales basadas en el aprendizaje algorítmico automático serían especialmente valiosas, además, cuando haya que tomar decisiones en contextos de ingente información y con resultados

inciertos y probabilísticos. Si bien el autor defiende con entusiasmo la aplicación de la IA para la mejora moral humana, reconoce que este campo emergente se enfrenta a tres desafíos. En primer lugar, el "problema del pluralismo axiológico" constata la difícil programación de la IA para que tenga en cuenta la rica variedad de valores presentes en la moralidad humana. En segundo lugar, el "problema de la evitabilidad" refiere a la dificultad de impedir que la IA nos manipule o engañe en la toma de decisiones. En tercer y último lugar, el "problema de la atrofia moral" señala el riesgo de que el uso habitual de estos asistentes artificiales pueda perjudicar nuestras habilidades morales a largo plazo.

Lara y Deckers, por otro lado, ponen a disposición del público castellanoparlante su artículo "La inteligencia artificial como asistente socrático para la mejora moral", originalmente publicado en *Neuroethics* (2020). Este asistente consistiría en un bot conversacional que, mediante el diálogo socrático, mejoraría la toma de decisión de los usuarios. El proceso dialógico ayudaría a aportar mayor apoyo empírico, aumentar la claridad conceptual, comprender la lógica argumentativa, determinar el grado de plausibilidad de un juicio moral particular, generar conciencia de las limitaciones personales y asesorar sobre cómo ejecutar las decisiones. Este modelo socrático garantizaría la autonomía de los usuarios dándoles un rol protagonista en su propio proceso de mejora moral, a diferencia de propuestas anteriores que les relegaban a un lugar secundario (Savulescu y Maslen, 2015; Giubilini y Savulescu, 2018). Esta idea del asistente moral socrático ha sido recientemente ampliada en Lara (2021).

En definitiva, tanto Monasterio-Astobiza como Lara y Deckers consideran la IA como una candidata prometedora para la mejora moral, sumándose a esta posición ya defendida, aunque con diferentes modelos, por autores previos (Savulescu y Maslen, 2015; Klincewicz, 2016; Giubilini y Savulescu, 2018). Creo, eso sí, que hay al menos tres cuestiones relacionadas con ambos capítulos que necesitan cierta discusión.

3. IA, ética y mejora moral

Una vez presentada la posibilidad de usar métodos no convencionales como la IA para mejorar la moralidad humana, conviene ahora, en primer lugar, clarificar las posibles relaciones entre la ética y la IA.

3.1. *La taxonomía sobre la relación entre la ética y la IA*

Aníbal Monasterio-Astobiza ofrece una triple taxonomía: la ética de la IA, la ética para máquinas y IA moral (todos términos provenientes de la literatura académica en inglés). En primer lugar, la *ética de la IA* trata sobre los aspectos éticos que surgen en el uso, desarrollo o implementación de sistemas de IA. Se basa en análisis normativos realizados por humanos o instituciones y puede proponer principios, guías o reglas para el desarrollo y aplicación ética de estas tecnologías computacionales y algoritmos de aprendizaje automático (Floridi et al., 2018; Coeckelbergh, 2021; Tsamados et al., 2021).

En segundo lugar, la *ética para máquinas* consiste en desarrollar máquinas o sistemas autónomos artificiales que tomen decisiones morales. Es decir, estas máquinas basadas en IA considerarían aspectos morales a la hora de proponer sus recomendaciones o de ejecutar ciertas acciones. Un ejemplo famoso es el de los coches autónomos. Los vehículos autónomos pueden enfrentarse a colisiones inevitables en las que deben decidir quién vive y quién muere, esto es, ante dilemas sacrificiales tipo el problema del tranvía (Keeling, 2020; Rueda, 2022). El desarrollo de los algoritmos para decidir en tales contextos tiene en cuenta explícitamente la consideración de factores morales. Más ejemplos podrían darse en la implementación de robots sociales y otras máquinas que tomen decisiones por sí mismas con repercusiones morales.

En tercer lugar, *IA moral* consiste, según Monasterio Astobiza, en "cómo aprovechar el potencial de la IA para tomar mejores decisiones en contextos morales" y "conseguir la cooperación máquina-ser humano para resolver problemas complejos e importantes" (Lara y Savulescu, 2021, p. 267). En esta última caracterización es donde el autor inscribe la cuestión de la mejora moral mediante la IA. Esta caracterización terminológica, proveniente de Maslen y Savulescu (2015), es desafortunada y confusa. El concepto de 'IA moral' da literalmente a entender, más bien, un sistema artificial inteligente que se comporta moralmente. Este no es el objetivo de usar la IA para la mejora moral humana. El fin de esta última es que los humanos se comporten más éticamente o que tomen mejores decisiones morales, no las máquinas. Los sistemas de IA serían simplemente el medio de mejora. Por lo tanto, sería conveniente no utilizar el término 'IA moral' para referirnos a las posibilidades de mejora moral mediante la IA.

En contrapartida, creo que puede haber otras formulaciones simples más precisas en las que incluir la discusión de la mejora moral a través de la IA. En particular, propongo que el término 'IA para la ética' podría ser un concepto más adecuado. La *IA*

para la ética establece de manera clara que la IA es el medio que tiene como finalidad apoyar en la toma de decisiones éticas. Incluiría, obviamente, la mejora moral mediante la IA. Sin embargo, al ser un paraguas conceptual más amplio, podría incluir otras opciones (que se comentarán más adelante) que no tengan que ver con la mejora particular de las habilidades morales de individuos, sino todo tipo de decisiones asistidas mediante la IA que tengan el objetivo de considerarse como éticas. En consecuencia, la 'IA para la ética' podría ser un avance terminológico que ofrezca un acomodo más preciso a las futuras propuestas de mejora moral mediante la IA y también a otras incipientes aplicaciones.

3.2 *Sobre la necesidad de utilizar la IA para mejorar las decisiones morales*

¿Por qué necesitamos la IA para la ética? Hay dos justificaciones principales para defender el uso de la IA para mejorarnos moralmente: (*1*) la complejidad de nuestros entornos sociotécnicos y (*2*) las deficiencias de la psicología moral humana. Estos dos argumentos, como veremos, también pueden solaparse. Monasterio-Astobiza se apoya en los dos, mientras que Lara y Deckers parten implícitamente del segundo. Conviene hacer ciertas matizaciones respecto a ambas razones.

El primer argumento señala que el mundo contemporáneo es realmente complejo. Consideremos el siguiente fragmento de la contribución de Monasterio-Astobiza:

> Como una *trampa-22*, un callejón sin salida lógico, la humanidad solo puede escoger seguir mecanizando y automatizando la vida y el trabajo porque cualquier alternativa a escoger es perjudicial. Si continuamos la senda de la automatización, tal y como nos advirtiera Giedion o Wiener, podemos deshumanizarnos. Pero es que si no continuamos la senda de la automatización y mecanización de la vida y el trabajo, la complejidad de un mundo hiperconectado con problemas globales (cambio climático, desigualdades, pandemias, etc.) nos desbordará dado que estos problemas solo pueden gestionarse con la ayuda de máquinas y sistemas socio-técnicos. La construcción de un mundo automatizado y digital y los problemas que entraña solo puede responderse desde un marco que contemple la automatización y digitalización de las soluciones (Lara y Savulescu, 2021, p. 261).

Este argumento parte de un hecho indudable: el actual mundo globalizado es altamente complejo, interconectado y con desafíos colectivos sumamente difíciles de coordinar. Sin embargo, Monasterio-Astobiza presenta la automatización (incluida la de la mejora moral) como algo inevitable. Esta visión demasiado determinista debería atemperarse. Hay alternativas a la automatización y sus mismos desarrollos están sometidos a

incertidumbres y diferentes grados de intensidad. Entender la automatización como un proceso unívoco o como una tendencia unificada a escala planetaria es demasiado simplista e ignora la gran variabilidad de factores que condicionan sus distintas implementaciones. Asimismo, incluso aunque quisiéramos auxiliarnos en métodos computacionales para tomar mejores decisiones morales cabría una gran diversidad de modelos entre los que habría que decidir. Además, es curioso que presente la automatización como una solución al problema mismo de la automatización. Esta argumentación circular es paradójica e insatisfactoria.

El segundo argumento señala que los seres humanos tenemos deficiencias en nuestra psicología moral y que estos están en parte condicionados por "la influencia de la biología en la moralidad" (Lara y Savulescu, 2021, p. 283). Estas limitaciones se hacen más evidentes en nuestro mundo global actual con aspiraciones éticas exigentes y universalistas. Rasgos fundamentales de la psicología moral humana se forjaron evolutivamente en el Pleistoceno, cuando se vivía en pequeñas comunidades interdependientes y con lazos cercanos (Schaik et al., 2014; Burkart et al., 2018; Tomasello, 2018). La idea de que muchas de esas características han quedado desfasadas en el mundo contemporáneo ha sido, de hecho, uno de los argumentos principales para demandar la biomejora moral (Persson y Savulescu, 2012). Lara y Deckers, por su parte, argumentan que, a pesar de la sobrada evidencia respecto a la influencia biológica en moralidad, los métodos de biomejora son problemáticos y que sería más deseable corregir nuestras limitaciones de manera externa mediante la IA. Este propósito es en cierto modo laudable, ya que las evidencias neurocientíficas recientes apoyan su premisa de que la moralidad humana tiene una base biológica importante y algo deficiente (cf. Rueda, 2021a). Sin embargo, cabe advertir que la IA también tiene sesgos y limitaciones considerables. De hecho, los posibles sesgos discriminatorios de la IA son de las mayores preocupaciones (Coeckelbergh, 2021; Pot et al., 2021; Starke et al., 2021). La IA, además, no es infalible y puede equivocarse en las tareas que realiza o errar en sus predicciones y recomendaciones. Estos casos crean graves desafíos para la atribución de la responsabilidad (Martin, 2019; Felder, 2021; Tigard, 2021). Esto ha sido ampliamente ignorado por los defensores de la mejora moral mediante la IA y creo que debería cobrar más peso en el debate futuro. Bien es cierto que, más recientemente, Lara (2021) ha reconocido el que la falibilidad humana puede derivar también en errores de programación que serían problemáticos para la mejora moral.

3.3. *Más allá de la mejora moral: otros potenciales innovadores de la IA para la ética*

Restringir los potenciales de la IA para la ética al ámbito de la mejora moral individual sería desacertado. Mejorar las habilidades morales de las personas es *prima facie* algo positivo (Rueda, 2021b). Sin embargo, muchas veces de lo que se trata es de tomar decisiones más éticas, independientemente de si esto mejora o no las capacidades para decidir moralmente. La IA podría ayudar en muchos ámbitos a tomar decisiones más morales con independencia de si eso mejora las cualidades morales de los decisores. Por lo tanto, hay otros potenciales que no deberían pasar desapercibidos. Voy a exponer brevemente dos de ellos.

Por un lado, la IA puede asistir decisiones en contextos profesionales donde hay implicados una gran cantidad de datos y cuyas actividades tienen claras implicaciones normativas, como el ámbito sanitario o empresarial. Los datos masivos son difíciles de gestionar para los humanos y las herramientas computacionales basadas IA pueden ser muy valiosas en su procesamiento (Colmenarejo Fernández, 2017; Calvo, 2021). Un ejemplo de aplicación a nivel sanitario sería el uso de la IA para asistir en la distribución de recursos médicos escasos. En estas situaciones, se deben tener en cuenta múltiples variables para predecir los beneficios de pacientes candidatos y priorizar en base a criterios éticos. Ya se ha propuesto un modelo, por ejemplo, para la distribución ética de órganos a partir de la IA (Sinnott-Armstrong y Skorburg, 2021), aunque hay todavía que superar serios desafíos como la inexplicabilidad del funcionamiento de los algoritmos (Rueda et al., 2022).

Por otro lado, la IA también puede ser de gran ayuda para la toma de decisiones más allá de la agencia individual adscrita a personas concretas. Las instituciones y las compañías también pueden aprovechar este potencial, ya sean del sector público o del privado. El gran potencial de la IA para la ética en el contexto institucional reside en el diseño de políticas públicas o estrategias de actuación en las que no solo están involucrados una gran cantidad de datos y variables factuales, sino también múltiples preferencias morales de la ciudadanía. La IA podría procesar las actitudes morales de la población respecto a una controversia concreta para informar (no dictaminar) las decisiones públicas correspondientes. Esto sería particularmente interesante en el diseño de políticas y regulaciones respecto a tecnologías emergentes especialmente polémicas

(cf. Savulescu et al., 2021), sobre todo en las que ya ha habido encuestas masivas sobre los aspectos morales en juego (Awad et al., 2018). En definitiva, la idea consistiría en que las instituciones usen la IA para diseñar políticas públicas basadas tanto en evidencias como en valores. Ello ayudaría a poner en marcha políticas más éticas. Esto desafortunadamente parece haber pasado desapercibido.

4. Conclusión

Esta nota crítica ha ampliado y discutido algunas cuestiones confusas o desatendidas en el debate emergente sobre el uso de la IA para la mejora moral, a raíz de dos contribuciones del libro de editado por Lara y Savulescu. En cuanto a la relación entre la IA y la ética, se ha mostrado, que la investigación filosófica no solo debe centrarse en la ética de la IA, sino que también debemos encaminarnos a reflexionar detenidamente sobre el uso de la IA para la ética. La IA para la ética no solo recoge esa "aspiración a superarnos" que mencionaban Lara y Savulescu en la introducción, sino que también nos muestra que a veces nos sentimos *superados* por el entorno altamente complejo y cada vez más "datificado" en el que se enmarcan nuestras aspiraciones normativas. Quizás la IA nos ayude a sentirnos menos sobrepasados como humanos y, en el mejor de los casos, nos ayude a tomar decisiones más éticas.

Referencias:

Agar, N. (2010). Enhancing Genetic Virtue? *Politics and the Life Sciences*, 29(1), 73-75.

Agar, N. (2015). Moral Bioenhancement is Dangerous. *Journal of Medical Ethics*, 41, 343-345.

Ayala, F. J. (2010). The difference of being human: Morality. *Proceedings of the National Academy of Sciences of the United States of America*, 107(SUPPL. 2), 9015–9022. https://doi.org/10.1073/pnas.0914616107.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., y Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Buchanan, A. (2011). *Beyond humanity? The ethics of biomedical enhancement*. Oxford: Oxford University Press.

Bostrom, N. y Savulescu, J. (2017). *Mejoramiento humano*. Teell.

Burkart, J. M., Brügger, R. K., y Van Schaik, C. P. (2018). Evolutionary origins of morality: Insights from non-human primates. *Frontiers in Sociology*, 3, 17.

Calvo, P. (2021). Una propuesta de diseño de sistema de gobernanza ética de datos masivos para la investigación e innovación responsable. *DILEMATA*, 34, 31-49.

Clark, A. y Chalmers, D. (1998). The extended mind. *Analysis* 58(1), 7-19.

Coeckelbergh, M. (2021). *Ética de la inteligencia artificial*. Madrid: Cátedra.

Colmenarejo Fernández, R. (2017). *Una ética para big data*. *Introducción a la gestión ética de datos masivos*. Barcelona: Editorial UOC.

Diéguez, A. (2017). *Transhumanismo: La búsqueda tecnológica del mejoramiento humano.* Barcelona. Herder.

Diéguez, A. (2021). *Cuerpos inadecuados: El desafío transhumanista a la filosofía*. Barcelona. Herder.

Diéguez, A. y Véliz, C. (2019). Would Moral Enhancement Limit Freedom? *Topoi*, 38(1), 29-36.

Felder, R. M. (2021). Coming to Terms with the Black Box Problem: How to Justify AI Systems in Health Care. *Hastings Center Report*, 51(4), 38–45. https://doi.org/10.1002/hast.1248.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., y Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Giubilini, A., y Savulescu, J. (2018). The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence. *Philosophy and Technology*, 31(2), 169–188. https://doi.org/10.1007/s13347-017-0285-z.

Harris, J. (2007). *Enhancing Evolution*. Princeton: Princeton University Press.

Harris, J. (2016). *How to be Good. The Possibility of Moral Enhancement*. Oxford: Oxford University Press.

Jones, D. G. (2006). Enhancement: Are ethicists excessively influenced by baseless speculations? *Medical Humanities*, 32(2), 77–81. https://doi.org/10.1136/jmh.2005.000234.

Keeling, G. (2020). Why Trolley Problems Matter for the Ethics of Automated Vehicles. *Science and Engineering Ethics*, 26(1), 293–307. https://doi.org/10.1007/s11948-019-00096-1.

Klincewicz, M. (2016). Artificial intelligence as a means to moral enhancement. *Studies in Logic, Grammar and Rhetoric*, 48(1), 171–187. https://doi.org/10.1515/slgr-2016-0061.

Lara, F. (2017). 'Oxytocin, Empathy and Human Enhancement.' *Theoria*, 32(3), 367-384.

Lara, F., y Deckers, J. (2020). Artificial Intelligence as a Socratic Assistant for Moral Enhancement. *Neuroethics*, 13(3), 275–287. https://doi.org/10.1007/s12152-019-09401-y.

Lara, F. (2021). Why a Virtual Assistant for Moral Enhancement When We Could have a Socrates ? *Science and Engineering Ethics*, 1–27. https://doi.org/10.1007/s11948-021-00318-5.

Lara, F. y Savulescu, J. (eds.) (2021). *Más (que) humanos. Biotecnología, inteligencia artificial y ética de la mejora*. Madrid: Tecnos.

Liedo, B. y Rueda, J. (2021). In Defense of Posthuman Vulnerability. *Scientia et Fides*, 9(1), 215-239. http://dx.doi.org/10.12775/SetF.2021.008.

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, 160(4), 835–850. https://doi.org/10.1007/s10551-018-3921-3.

Monsó, S., Benz-Schwarzburg, J., y Bremhorst, A. (2018). Animal Morality: What It Means and Why It Matters. *Journal of Ethics*, 22(3–4), 283–310. https://doi.org/10.1007/s10892-018-9275-3.

Ortega Esquembre, C., Richart Piqueras, A., Páramo Valero, V. y Ruíz Rubio, C. (2015). *Mejoramiento humano. Avances, investigaciones y reflexiones éticas y políticas*. Granada: Comares.

Persson, I., y Savulescu, J. (2012). *Unfit for the future: The need for moral enhancement*. Oxford: Oxford University Press.

Pot, M., Kieusseyan, N., y Prainsack, B. (2021). Not all biases are bad: equitable and inequitable biases in machine learning and radiology. *Insights into Imaging*, 12(1). https://doi.org/10.1186/s13244-020-00955-7.

Raus, K., Focquaert, F., Schermer, M., Specker, J., y Sterckx, S. (2014). On Defining Moral Enhancement: A Clarificatory Taxonomy. *Neuroethics*, 7(3), 263-273.

Rueda, J. (2020). Climate Change, Moral Bioenhancement and the Ultimate Mostropic. *Ramon Llull Journal of Applied Ethics*, 11, 277-303. https://www.raco.cat/index.php/rljae/article/view/368709.

Rueda, J. (2021a). Socrates in the fMRI Scanner: The Neurofoundations of Morality and the Challenge to Ethics. *Cambridge Quarterly of Healthcare Ethics*, 30(4), 606-612. https://doi.org/10.1017/S0963180121000074.

Rueda, J. (2021b). Enhancing Virtue without Becoming Ned Flanders? *AJOB Neuroscience*, 12 (2-3), 121-124. https://doi.org/10.1080/21507740.2021.1904051

Rueda, J. (2022). Hit by the virtual trolley: When is experimental ethics unethical? *Teorema*, 41(1), en prensa.

Rueda, J. y Lara, F. (2020). Virtual Reality and Empathy Enhancement: Ethical Aspects. *Frontiers in Robotics and AI*, 7: 506984. https://doi.org/10.3389/frobt.2020.506984.

Rueda, J., Delgado Rodríguez, J. Parra Jounou, I., Hortal Carmona, J., Ausín, T. y Rodríguez-Arias, D. (2022). "Just" accuracy? Procedural fairness demands explainability in AI-based medical resource allocations. *Hasting Center Report*, en revisiones menores.

Savulescu, J. (2012). *¿Decisiones peligrosas? Una bioética desafiante*. (B. Rodríguez López y E. Bonete Perales, Eds. y Trads.). Madrid: Tecnos.

Savulescu, J., y Maslen, H. (2015). Moral enhancement and artificial intelligence: Moral AI? En J. Romportl, E. Zackova, y J. Kelemen (eds.), *Beyond artificial intelligence. The disappearing humanmachine divide* (pp. 79–95). Springer.

Savulescu, J., Gyngell, C., y Kahane, G. (2021). Collective Reflective Equilibrium in Practice (CREP) and controversial novel technologies. *Bioethics*, 35(7), 652–663. https://doi.org/10.1111/bioe.12869.

Sinnott-Armstrong, W. y Skorburg, J. A., (2021). How AI Can Aid Bioethics. *Journal of Practical Ethics*, 9(1). doi: https://doi.org/10.3998/jpe.1175.

Starke, G., De Clercq, E., y Elger, B. S. (2021). Towards a pragmatist dealing with algorithmic bias in medical machine learning. *Medicine, Health Care and Philosophy*, 24(3), 341–349. https://doi.org/10.1007/s11019-021-10008-5.

Tigard, D. W. (2021). Responsible AI and moral responsibility: a common appreciation. *AI and Ethics*, 1(2), 113–117. https://doi.org/10.1007/s43681-020-00009-0.

Tomasello, M. (2018). Precís of a natural history of human morality. *Philosophical Psychology*, 31(5), 661-668.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., y Floridi, L. (2021). The Ethics of Algorithms: Key Problems and Solutions. *AI & Society*. https://doi.org/10.1007/s00146-021-01154-8.

Van Schaik, C., Burkart, J. M., Jaeggi, A. V., y von Rohr, C. R. (2014). Morality as a biological adaptation–an evolutionary model based on the lifestyle of human foragers. En M. Christen, C. van Schaik, J. Fischer, M. Huppenbauer, y C. Tanner (eds.). *Empirically informed ethics: Morality between facts and norms* (pp. 65-84). Springer, Cham.