De la biomejora moral a la IA para la mejora moral: asistentes morales artificiales en la era de los riesgos globales*

Pablo Neira Castro
Universidad Complutense de Madrid
pablonei@ucm.es

Fecha de recepción: 20/6/2023
Fecha de aceptación: 2/11/2023
Fecha de publicación: 19/3/2024

Resumen

En este artículo argumento que el empleo de tecnologías biomédicas para la mejora moral es inadecuado para evitar riesgos globales, debido a que puede provocar problemas similares a los que se pretenden resolver por medio de su uso. Por este motivo, defiendo la conveniencia de explorar otras tecnologías para la mejora moral que no supongan tantos peligros y argumento en favor de un modelo de mejora moral por medio de la inteligencia artificial. En concreto, defiendo el empleo de SocrAI, un asistente moral artificial diseñado para mejorar nuestra deliberación moral. Para ello, propongo tres criterios que permitan evaluar y aumentar su seguridad y eficacia. Asimismo, señalo la importancia de tener en cuenta las cuestiones estructurales e institucionales —*i. e.*, las normas o los incentivos políticos, económicos, sociales y culturales— en las propuestas de mejora moral, y muestro cómo SocrAI puede tener impacto en ellas.

Palabras clave: mejora moral; asistentes virtuales; deliberación moral; SocrAI

Abstract. *From moral bio-enhancement to AI for moral enhancement: Artificial moral assistants in the age of global risks*

In this article, I argue that the use of biomedical technologies for moral enhancement is inadequate for mitigating global risks, as this can lead to problems similar to those they seek to address. For this reason, I advocate for exploring alternative technologies for moral enhancement that entail fewer risks, and I make a case for a model of moral enhancement through artificial intelligence. Specifically, I support the use of SocrAI, an artificial moral assistant designed to improve our moral deliberation, and propose three criteria to assess and enhance its safety and efficacy. Additionally, I highlight the importance of considering structural and institutional issues – i.e., political, economic, social and cultural norms

* Me gustaría agradecer a los dos revisores anónimos sus valiosas críticas y sugerencias. Además, agradezco a Joan Llorca Albareda y a Blanca Rodríguez López sus comentarios a las primeras versiones de mi artículo. El trabajo fue realizado en el marco de las prácticas externas del máster en Ética Aplicada de la Universidad Complutense de Madrid, en colaboración con el proyecto de investigación SocrAI+ [Mejora Moral e Inteligencia Artificial. Aspectos Éticos de un Asistente Virtual Socrático (2021-2023) / Ref: B-HUM-64-UGR20] de la Universidad de Granada.

or incentives – in moral enhancement proposals, and demonstrate how SocrAI can have an impact on them.

Keywords: moral enhancement; virtual assistants; moral deliberation; SocrAI

Sumario

1. Introducción

En el siglo pasado los seres humanos alcanzamos una capacidad tecnológica sin precedentes que nos ha dotado del poder de alterar radicalmente el medio en el que vivimos y a los seres que en él habitan. Desde entonces hemos desarrollado instrumentos, armas y herramientas con un gran poder transformador. Ello nos ha permitido mejorar la calidad de vida de numerosas personas y satisfacer las necesidades de multitud de individuos, pero también ha provocado ciertos problemas y riesgos que no podemos obviar. Entre estos encontramos el cambio climático de origen antrópico, el uso de armas de destrucción masiva o el crecimiento de la desigualdad. A su vez, en el futuro es esperable que aparezcan nuevos riesgos y problemas relacionados con tecnologías tales como la ingeniería genética, la inteligencia artificial (IA), la nanotecnología o las interfaces cerebro-máquina.

Debido a su gran poder transformador, estas tecnologías y el uso que hagamos de ellas condicionarán la manera en que será el futuro, puesto que previsiblemente tendrán consecuencias a muy largo plazo. Esto es lo que ha facilitado que algunos filósofos y científicos, desde hace algunos años, hayan comenzado a hablar acerca de los riesgos globales que los seres humanos afrontamos como especie, de los riesgos existenciales y de la responsabilidad con las generaciones futuras (cf. Bostrom, 2002; Bostrom y Cirkovic, 2011; Boyd y Wilson, 2020). Según algunos, lo que hagamos con algunas de estas y otras tecnologías que aún estamos comenzando a utilizar determinará cómo será el futuro. El problema que algunas voces denuncian consiste en que no estamos preparados para el futuro (Persson y Savulescu, 2012: 1), dado que nuestra limitada capacidad moral es insuficiente para afrontar adecuadamente estos problemas y riesgos. Por este motivo, argumentan que deberíamos emplear

algunas neurotecnologías o ingeniería genética para aumentar nuestras capacidades de reflexión y decisión ética, lo que se conoce como *biomejora moral*.

Sin embargo, esta posición entraña graves problemas. Si asumimos que no estamos preparados moralmente para afrontar los riesgos que estas tecnologías nos presentan, no parece una idea muy prudente tratar de evitarlo modificándonos genéticamente por medio de tecnologías que pueden suponer grandes riesgos con nuestras (inadecuadas) capacidades morales. Además, se ha denunciado que aquellas personas que sugieren emplear algunas tecnologías con estos fines —especialmente aquellas a quienes se conoce como *transhumanistas*— pertenecen a un grupo demográfico muy homogéneo, que no representa los intereses ni las preocupaciones de buena parte de la población mundial (Gleiberman, 2023: 11). Y argumentan que esta retórica apocalíptica de riesgos y problemas globales provocados por las nuevas tecnologías está siendo utilizada por los mismos defensores de estas tecnologías para dar sentido a sus propios objetivos e influir en la toma de decisiones políticas. Por estos motivos, algunos de los medios propuestos para mejorarnos moralmente y poder enfrentarnos a estos problemas parecen poco razonables, debido a los riesgos que entraña su uso. Es por ello por lo que conviene definir nuevos medios para la mejora moral que no sean tan invasivos como las propuestas de biomejora.

En este trabajo crítico las propuestas de biomejora por ser inadecuadas para lograr la mejora moral necesaria para enfrentar los problemas y riesgos globales, y trato de defender un modelo de mejora moral por medio de la IA (IAmejora moral). Para ello, en primer lugar, expongo la problemática que dio origen a las propuestas de mejora moral. En segundo lugar, critico la adecuación y cuestiono la necesidad de la mayoría de las propuestas de biomejora para evitar riesgos globales. En tercer lugar, trato de defender un modelo de IAmejora a partir de un bot conversacional conocido como SocrAI. Por último, analizo algunas de las principales críticas a este modelo y trato de contestarlas.

2. La era de los riesgos globales: necesidad de una ética para pensar el futuro

2.1. Inadecuados para el futuro[1]

Los seres humanos hemos desarrollado y empleado tecnología durante toda nuestra historia para sobreponernos a nuestras necesidades y vulnerabilidades naturales. Desarrollamos herramientas, armas, edificaciones y sistemas de cultivo para obtener alimento y protegernos del frío y de los animales salvajes. A su vez, empleamos sistemas de transporte para movernos a mayores distancias de las que nuestra limitada fisiología nos permite. Nuestro poder, en relación con el medio en el que vivimos, siempre ha sido muy limitado. Ello llevó a que, al comienzo de la modernidad, filósofos como Francis Bacon (1620, libro I, aforismo 129) o Descartes (1637, sexta parte: 45) apuntasen al dominio de

1. Título tomado de la obra de Ingmar Persson y Julian Savulescu, *Unfit for the future.*

la naturaleza como objetivo de la técnica y la ciencia humanas. Así, el proyecto moderno se inició como el de desvelar los secretos de la naturaleza por medio de la ciencia y emplear su conocimiento para dominarla gracias a la tecnología, llevándonos a un avance sin precedentes en el conocimiento científico y en la invención de máquinas y artilugios con los que dominar la naturaleza.

Hoy, sin embargo, nos encontramos ante un nuevo problema. Nuestras capacidades tecnológicas han aumentado hasta tal punto que nuestro dominio sobre la naturaleza es —al menos tal y como lo concibieron los filósofos modernos— efectivo. Sin embargo, carecemos del poder de controlar algunos de los efectos que nuestras tecnologías tienen sobre la naturaleza, invirtiendo la situación con respecto a la modernidad: los problemas a los que nos enfrentamos no son ya de origen natural, ni se producen por nuestra falta de poder frente a la naturaleza, sino que son de origen antrópico, y se deben a nuestro enorme poder tecnológico, que excede nuestra capacidad de prever y evitar los perjuicios que provocan en el medioambiente y los animales humanos y no humanos que en él habitan[2].

Estos nuevos problemas, además, no parten únicamente de dificultades técnicas y epistémicas, sino que son fundamentalmente éticos y políticos. Muchos de los perjuicios que provocamos por medio del uso de la tecnología nos son conocidos, como el cambio climático y los derivados del posible uso de armas de destrucción masiva. Sin embargo, no estamos haciendo lo suficiente por evitarlos. Esto, para algunos filósofos como Ingmar Persson y Julian Savulescu (2012), es muestra de que nuestra capacidad tecnológica ha evolucionado más que nuestra psicología moral, y por ello ven en la modificación de esta última la posibilidad de sobreponernos a estos nuevos problemas. Es decir, creen que nuestras capacidades morales, determinadas por nuestra biología, son inadecuadas para afrontar estos riesgos y problemas. Por ello, ven en la transformación de nuestra biología y de la naturaleza humana la esperanza de poder adecuarnos a las nuevas circunstancias creadas por nuestro avance tecnológico.

Dada nuestra incapacidad para tomar medidas éticas y políticas con el fin de evitar estos riesgos y problemas, parece que nuestras actuales capacidades morales son insuficientes para afrontar los retos que nos plantea el uso de ciertas tecnologías. Por este motivo, necesitamos mejorar moralmente, en el sentido de aumentar nuestra confianza en los demás, altruismo, sentido de la justicia y deseo de colaborar para evitar estos problemas. Los seres humanos mejoramos nuestro comportamiento moral a lo largo de la historia por medio de la educación, la socialización y la comunicación intercultural. No obstante, estos medios de mejora moral tradicionales parecen demasiado lentos e inefi-

2. Günther Anders (2011: 31-32) describe este desfase o ritmo diferente entre nuestra capacidad de actuar y crear y nuestra capacidad de comprender y sentir aquello que provocamos por medio de nuestro actuar, llamando desnivel [o brecha] prometeica a la «a-sincronía del hombre con su mundo de productos».

caces, debido a la dificultad de transmitir el desarrollo moral que un individuo ha alcanzado a lo largo de su vida a las siguientes generaciones (Persson y Savulescu, 2012: 118) y al hecho de que siguen existiendo conductas morales cuestionables (*e. g.* corrupción, asesinatos, guerras, violaciones…) pese a los grandes esfuerzos que ha habido por eliminarlas (Lara y Deckers, 2020: 275). Esto ha llevado a algunos teóricos a defender formas de mejora moral no tradicionales a partir de diferentes tecnologías, como psicofármacos, hormonas u otras neurotecnologías (Douglas, 2015; Earp et al., 2017; Macpherson et al., 2019), ingeniería genética (Walker, 2009b; Persson y Savulescu, 2010), robots o inteligencia artificial (Gips, 1995; Savulescu y Maslen, 2015; Borenstein y Arkin, 2016; Giubilini y Savulescu, 2018; Lara y Deckers, 2020).

Estas formas de mejora moral no tradicionales podrían ayudarnos a evitar riesgos globales como el cambio climático antropogénico o el posible uso de armas de destrucción masiva, al promover:

a) La responsabilidad ambiental, impulsándonos a tomar medidas que reduzcan nuestra huella ecológica, como el uso de energía limpia, la reducción de residuos y la adopción de hábitos de consumo sostenibles.
b) El comportamiento altruista, ayudándonos a priorizar el bienestar de las generaciones futuras y de las sociedades que más se verán afectadas por estos problemas.
c) El consumo ético, llevándonos a optar por productos y servicios con un menor impacto ambiental e impulsando la demanda de alternativas sostenibles.
d) La cooperación y la colaboración transnacional, ayudando a las distintas sociedades a trabajar por el bien común.
e) La empatía y la resolución pacífica de conflictos, reduciendo la probabilidad de realizar carreras armamentísticas y escaladas bélicas.
f) La aversión a dañar a otras personas, impidiendo que ciertos individuos o grupos participen en actividades terroristas o traten de causar grandes daños.
g) La toma de decisiones éticas, haciéndonos menos propensos a participar o apoyar acciones que impliquen el uso de armas de destrucción masiva o un gran impacto ambiental.
h) La diplomacia internacional, favoreciendo la participación de los distintos individuos y sociedades en esfuerzos diplomáticos que pretendan resolver estos problemas.

2.2. La solución transhumanista

Las propuestas de mejora moral que más atención han recibido hasta el momento son las de biomejora moral, que defienden el empleo de tecnologías biomédicas como la ingeniería genética, hormonas, psicofármacos y otras neurotecnologías para alcanzar tal fin. Según sus defensores, estas tecnologías servirían para alterar nuestra psicología moral, haciéndonos menos propensos

a ciertas conductas inmorales y ayudándonos a alcanzar acuerdos y consensos para evitar riesgos y problemas a una escala global. De esta manera, la biomejora moral consiste en promover genes que influencien en la adquisición de virtudes (Walker, 2009b: 28); aumentar nuestra motivación moral para acelerar la internalización de las doctrinas morales (Persson y Savulescu, 2012: 107); dejar a los individuos con motivaciones o comportamientos moralmente mejores de los que habrían tenido de otra manera (Douglas, 2013: 162), o, de manera más general, mejorar nuestras capacidades morales (DeGrazia, 2014: 361).

Aunque no hay acuerdo acerca de qué características psicológicas son las que deberíamos tratar de modificar para mejorarnos moralmente, algunas de las que más atención han recibido son la empatía, la simpatía, el altruismo, el sentido de la justicia y las emociones morales (Douglas, 2008: 231; Persson y Savulescu, 2012: 108; DeGrazia, 2014: 363; Lara, 2017: 374-376). Los medios propuestos para hacerlo son de tipo biomédico, como el uso de hormonas, psicofármacos, estimulación transcraneal (magnética o por corriente directa), estimulación cerebral profunda o métodos de selección embrionaria e ingeniería genética (Persson y Savulescu, 2012: 107; DeGrazia, 2014: 361-362; Earp et al., 2017: 167).

Modificar algunas de las características psicológicas que condicionan nuestro comportamiento moral es, para algunos de sus defensores, la mejor o la única manera que tenemos de mitigar los riesgos agenciales provocados por la gran capacidad tecnológica que tenemos actualmente. Aunque en general suelen defender el uso voluntario de estas técnicas, también contemplan la posibilidad de que se apliquen a niños en escuelas sin su consentimiento —de igual modo que se les hace ir a la escuela para mejorarlos moralmente— (Persson y Savulescu, 2012: 113) o se emplee masivamente aplicando hormonas como oxitocina o serotonina al agua corriente, sumado a un aumento de los métodos de vigilancia de ciudadanos (Persson y Savulescu, 2012: 125). Todo ello con el objetivo de evitar que se empleen armas de destrucción masiva, a las que un creciente número de personas tendrá acceso; que sigamos propiciando el cambio climático antropogénico y la degradación ambiental, y que sigamos sin hacer nada para reducir la enorme desigualdad existente actualmente y la pobreza extrema de millones de personas.

3. Sobre la adecuación y la necesidad de la biomejora

3.1. La paradoja de la biomejora

Persson y Savulescu (2012: 130) defienden la biomejora debido a que creen que somos inadecuados para hacer frente a muchos de los riesgos que nos plantea el uso de algunas tecnologías y debido a que consideran que es más fácil dañar que beneficiar a alguien. Además, consideran que si desarrollamos medios efectivos para la mejora cognitiva posibilitaremos que un creciente número de personas posea más poder para causar grandes daños (Persson y

Savulescu, 2008)[3]. Por ello, argumentan que, si no mejoramos moralmente de una manera significativa, corremos un gran riesgo de que ciertos individuos empleen armas de destrucción masiva o generen daños a multitud de personas y de que no seamos capaces de frenar lo suficiente el cambio climático antropogénico. No obstante, es ciertamente controvertido que, si somos moralmente inadecuados para emplear prudentemente algunas de las nuevas tecnologías y es más fácil dañar que beneficiar, podamos usar las tecnologías de biomejora de una manera satisfactoria.

Es decir, que las dos premisas que emplean Persson y Savulescu para defender la biomejora sirven también para rechazar la mayoría de los medios para mejorarnos, debido a que su alcance es potencialmente enorme y sus efectos, ampliamente imprevisibles. Esto es a lo que llamo la *paradoja de la biomejora*, que podemos resumir así: somos moralmente inadecuados para afrontar los riesgos que provocan algunas de las nuevas tecnologías, por lo que necesitamos mejorarnos moralmente para evitar riesgos globales; pero el empleo de tecnologías para la biomejora previsiblemente provocará una gran cantidad de riesgos globales, debido a que somos moralmente inadecuados para usarlas. Entre ellos encontramos los siguientes: que supongan perjuicios injustificables para quienes se mejoren; que sirvan a aquellos individuos o grupos que pretenden causar grandes daños para lograrlo, o que su desarrollo provoque la difusión de armas de destrucción masiva más baratas y eficaces que las existentes actualmente.

Tal como la conciben sus defensores, la biomejora debe ser aplicada a buena parte de la población mundial[4], puesto que, aunque solo haya un pequeño porcentaje de la población con motivos extremadamente inmorales, podrían llegar a emplear armas de destrucción masiva (Persson y Savulescu, 2008: 174). Esto provoca que si un medio para la biomejora se muestra perjudicial tras ser empleado, los daños que puede suponer sean enormes. Sin embargo, aumentar la seguridad de las tecnologías para la biomejora moral supondrá previsiblemente aumentar el tiempo de su desarrollo, quizás más de lo que podemos demorarnos para evitar la catástrofe climática o el uso de armas de destrucción masiva. Además, nada garantiza que aquellas tecnologías desarrolladas para la biomejora moral —que tienen como fin alterar las emociones, la motivación y el comportamiento humanos— no sean empleadas por ciertos grupos de manera que perjudiquen a multitud de personas. Algunos gobiernos autoritarios podrían emplearlas, no para mejorar moralmente a sus ciudadanos, sino para hacerlos más dóciles y sumisos (Sparrow, 2014: 28). Objetivos similares podrían ser perseguidos por algunas sectas o grupos coercitivos, así como por grupos terroristas.

3. Sin embargo, hay quien defiende que, para lograr la mejora moral tal y como la conciben estos autores, es necesario mejorarnos cognitivamente, ya que consideran que hay una retroalimentación entre nuestras capacidades morales y cognitivas (véanse Harris, 2011: 106, y Gordon y Ragonese, 2023).
4. Este objetivo, además de ser moralmente problemático, supone múltiples desafíos de implementación que están lejos de resolverse, como los señalados por Jon Rueda (2020).

Del mismo modo, algunos agentes que pretenden causar grandes daños podrían emplear estas tecnologías para alcanzar sus objetivos. Esta idea es defendida por Phil Torres (2017) al analizar los diferentes tipos de agentes que podrían provocar riesgos existenciales si tuvieran los medios necesarios para ello. En su artículo argumenta que la biomejora moral podría aumentar la probabilidad de que algunos utilitaristas negativos fuertes o futuros ecoterroristas intenten extinguir a los seres humanos, debido a que con ello cumplirían sus objetivos morales, como la eliminación del sufrimiento o los perjuicios medioambientales que provocamos, respectivamente (Torres, 2017: 4-5). Esto podría suceder empleando las tecnologías de biomejora de la manera prevista por sus defensores, al aumentar su motivación moral, su sentido de la justicia o su altruismo. Pero, nuevamente, no hay garantías de que no sean usadas con objetivos distintos a los que sus defensores prevén. Si desarrollamos biotecnologías que permitan alterar el comportamiento humano de forma controlada, algunos grupos que pretenden causar grandes daños podrían emplearlos a su favor, reduciendo el altruismo y el sentido de la justicia de multitud de individuos o aumentando otras disposiciones que los hagan más proclives a ciertas conductas inmorales.

Finalmente, el desarrollo de algunas de las tecnologías necesarias para alcanzar la biomejora moral efectiva, como las de edición e ingeniería genética, previsiblemente tendrá como consecuencia la difusión y el abaratamiento de ciertas armas de destrucción masiva, como los virus modificados genéticamente. La mejora moral por medio de la ingeniería genética es defendida por algunos autores como la mejor manera de evitar nuestra extinción. El motivo de ello reside en que actualmente disponemos de armas que pueden destruir la civilización, y creen que la única manera de evitar su uso es por medio de la creación de una especie posthumana que tenga mayores capacidades morales que los seres humanos y pueda garantizar que nunca sean usadas (Walker, 2009a: 6-7; Persson y Savulescu, 2010: 668). Sin embargo, las tecnologías necesarias para lograrlo son las mismas que podrían servir para crear armas biológicas con un enorme potencial destructivo. Es decir, son tecnologías de doble uso: pueden ser empleadas tanto para fines benéficos (*e.g.* mejorarnos moralmente, aumentar nuestra capacidad cognitiva, etc.) como para fines ampliamente perjudiciales (*e.g.* guerra biológica, difusión de virus modificados genéticamente, creación de supersoldados).

Walker (2009a: 7) argumenta que esto no es un motivo para rechazar la biomejora genética, debido a que ya disponemos de armas que pueden acabar con la civilización, por lo que no hay nada mucho peor que el proyecto transhumanista pueda provocar. Sin embargo, su argumento, que se centra en el potencial destructivo de las armas de las que disponemos, obvia cuántos agentes tienen acceso a ellas en la actualidad, y cuántos podrían tener acceso a ellas en caso de desarrollar los medios necesarios para su propuesta, así como cuál es la probabilidad de que sean usadas en uno y otro escenario. La mayoría de las armas de destrucción masiva existentes en la actualidad —especialmente aquellas que pueden provocar nuestra extinción, como las armas nucleares o

los virus modificados genéticamente— requieren de una gran capacidad económica, técnica y cognitiva para ser desarrolladas y utilizadas. Por este motivo, únicamente algunos Estados y grandes organizaciones militares disponen de ellas[5]. No obstante, para conseguir la biomejora genética, será necesario aumentar las investigaciones en ingeniería genética y abaratar las tecnologías de edición genética existentes actualmente. Esto, antes de permitirnos crear una especie posthumana moralmente mejorada, lo que provocará será que se reduzcan drásticamente las capacidades necesarias para emplear armas biológicas. Las tecnologías necesarias para la creación de posthumanos mejorados moralmente tardarán previsiblemente mucho tiempo en desarrollarse, y todavía más en implementarse; mientras que el desarrollo y la difusión de tecnologías baratas y técnicamente accesibles para la creación de virus modificados genéticamente tardará previsiblemente mucho menos, y se verá retroalimentado por el desarrollo de la biomejora genética. Esto hace que el proyecto transhumanista pueda aumentar las probabilidades de que se produzca nuestra extinción, ya que antes de conseguir mejorarnos moralmente puede facilitar a un mayor número de agentes el acceso a armas con la capacidad de provocarla, aumentando la probabilidad de que sean usadas con respecto a la situación actual. Por tanto, si queremos evitar el empleo de armas de destrucción masiva, parece conveniente atender a otras formas de mejora moral que puedan implementarse más rápidamente y cuyo desarrollo no facilite la creación ni el acceso a armas biológicas.

3.2. ¿Necesitamos realmente la biomejora?

Recientemente ha aparecido en la literatura toda una corriente que critica los motivos, las aspiraciones y la retórica empleada por los defensores de la mejora humana y de algunos movimientos occidentales contemporáneos como una forma de neocolonialismo que, por medio de narrativas catastrofistas y apocalípticas, trata de influir en las instituciones globales y de imponer los valores occidentales (Boenig-Liptsin y Hurlbut, 2016; Gergan et al., 2020; Mitchell y Chaudhury, 2020; Simpson, 2020). Dentro de esta corriente se relacionan el transhumanismo y la mejora humana con ciertos movimientos filosóficos angloamericanos, como el largoplacismo o el altruismo eficaz, y se argumenta que, bajo la pretensión de estar evitando riesgos catastróficos globales, realmente se está impulsando una agenda tecnológica muy concreta (Gleiberman, 2023: 2). La idea que tratan de defender es que muchos de los riesgos que denuncian los transhumanistas, provocados por las nuevas tecnologías, son un caballo de Troya para defender el desarrollo de tecnologías enormemente con-

5. Häggström (2016: 4, traducción propia) argumenta en este sentido con respecto a las armas nucleares: «¿Por qué, casi 70 años después de Hiroshima y Nagasaki, la cifra total de personas muertas en una guerra nuclear es un número de seis dígitos, y no uno de diez? […] [U]n hecho especialmente vital es que la tecnología de las armas nucleares […] ha demostrado su lentitud a la hora de facilitar su acceso. […] No todas las tecnologías tienen esta característica».

trovertidas para solucionarlos. Creen que, por medio de la retórica de salvar a la humanidad, del Antropoceno, del cambio climático y de los riesgos existenciales, se está tratando de imponer una agenda occidentalista en la política global. Asimismo, acusan a los defensores de esta agenda de pertenecer a un grupo demográficamente muy reducido de hombres blancos, socioeconómicamente privilegiados y con predilección por la filosofía analítica, la computación y la tecnología (Mitchell y Chaudhury, 2020: 312; Gleiberman, 2023: 11). Por estos motivos argumentan que no representan en modo alguno los intereses de los seres humanos a nivel global, como ellos pretenden. Además:

1. Los problemas que denuncian son provocados precisamente por las sociedades en las que nacen estos movimientos, y no por todos los seres humanos (*anthropos*), como sugieren las narrativas del Antropoceno (Simpson, 2020: 64-65).
2. Las soluciones que proponen son una nueva forma de colonialismo para influir en las culturas no blancas, indígenas y no occidentales (Mitchell y Chaudhury, 2020: 311-312; Gleiberman, 2023: 11).
3. Los discursos acerca del Antropoceno y de los riesgos globales deberían abrirse a una mayor pluralidad de voces (Simpson, 2020: 68).

Estas críticas muestran hasta qué punto los problemas que conceptualizamos y las soluciones que entendemos como razonables están influidos por las narrativas culturales con las que construimos nuestra identidad. Mark Coeckelbergh (2022), en su libro *Self-improvement: Technologies of the soul in the age of artificial intelligence*, emplea la noción de identidad narrativa de Paul Ricoeur para señalar cómo las ansias de autosuperación y mejora occidentales están íntimamente ligadas a la manera en la que nos interpretamos a nosotros mismos y a nuestra sociedad. Por medio de esta defiende una concepción relacional del «yo», según la cual los seres humanos construimos nuestra identidad a partir de las narrativas sociales y culturales que nos rodean. Pero los seres humanos no somos solo socialmente relacionales, sino que defiende que también somos tecnológicamente relacionales, en el sentido de que nuestra identidad está condicionada por las tecnologías de las que disponemos, y que las tecnologías que imaginamos y tratamos de desarrollar están influidas por cómo entendemos nuestra identidad (Coeckelbergh, 2022: 86). Por ello, influir en las narrativas que dan forma a nuestra manera de entender el mundo puede tener un enorme impacto en la reducción de riesgos, y no deberíamos desestimar esta opción asumiendo que la inevitabilidad del desarrollo tecnológico es algo dado y que no está sujeto a ser cambiado.

3.3. Más allá de la biomejora

Si la biomejora parece tan problemática para resolver los problemas a los que nos enfrentaremos en el futuro cercano, ¿qué nos queda entonces? ¿Debemos resignarnos a aceptar sufrir las consecuencias de los riesgos provocados por el

uso de la tecnología? En primer lugar, como apunté arriba, conviene revisar los presupuestos con los que conceptualizamos los problemas y las soluciones que se están debatiendo. En particular, conviene revisar la tesis de la inevitabilidad del desarrollo tecnológico, puesto que si bien parece poco probable que en un corto período de tiempo las diferentes sociedades que desarrollan tecnología avanzada lleguen a acuerdos y establezcan regulaciones globales para evitar riesgos, también lo es que consigamos, como proponen los defensores de la biomejora, mejorar moralmente a buena parte de su población.

En segundo lugar, como han defendido algunos teóricos de los riesgos globales y de la mejora humana, como Nick Bostrom (2014: 229-237) o Toby Ord (2020: 200), deberíamos tratar de implementar un desarrollo diferencial de las tecnologías emergentes según su alcance y peligrosidad. El desarrollo diferencial es entendido por ellos como la aceleración del desarrollo de las tecnologías que nos protejan de los riesgos existenciales, ante la presunción de que frenar el desarrollo de tecnologías que generen estos riesgos será prácticamente imposible. Aunque deberíamos hacer esto, creo que no deberíamos asumir tan apresuradamente la tesis de la inevitabilidad, y pasar a entender el desarrollo diferencial de una manera más amplia, como la aceleración del desarrollo de tecnologías protectoras y la desaceleración del uso y la implementación de tecnologías peligrosas. En este sentido, la ingeniería genética aplicada a seres humanos es probablemente una de las tecnologías que deberíamos de considerar peligrosas, y con las que deberíamos extremar la prudencia[6]. Además, debido a que nuestros valores, preocupaciones e intereses condicionan aquello que valoramos positiva o negativamente, probablemente algo que ahora consideremos una mejora genética pueda volverse un empeoramiento en un corto espacio de tiempo[7]. Jonathan Glover (2006: 98, traducción propia) expresa esta preocupación: «John Mackie me dijo una vez que si la ingeniería genética humana hubiera estado disponible en la época victoriana, la gente podría haber diseñado a sus hijos para que fueran patriotas y piadosos». Esto, a mi modo de ver, muestra que el rápido cambio de nuestros valores hace que una mejora definitiva e irreversible, como a la que puede dar lugar la ingeniería genética, puede llegar a ser indeseable.

Por ello, creo que deberíamos apostar por mejoras morales que puedan ser afinadas conforme se implementan, y que no sean definitivas ni irreversibles. A su vez, dado que ha sido el empleo de algunas tecnologías el que ha provo-

6. Sin embargo, debido a que es probable que en los próximos años aparezcan nuevas tecnologías genéticas enormemente disruptivas y transformativas, es necesario pensar de antemano cuáles serán sus posibles consecuencias, como defiende Marcos Alonso (2024) en este volumen. Asimismo, debemos comenzar a considerar cómo desarrollar y aplicar respuestas institucionales a los posibles problemas derivados de estas tecnologías. Para ello, Rueda (2024) defiende en otro artículo de este número especial que la creación de instituciones de gobernanza global de las tecnologías de mejora genética es un paso necesario.
7. Sparrow (2015: 235-238) argumenta en un sentido similar al considerar que, debido a que muchas de las tecnologías para la biomejora son irreversibles, difíciles de actualizar y modificar, es probable que las mejoras conseguidas por medio de ellas queden obsoletas en un corto espacio de tiempo.

cado los riesgos que nos han llevado a necesitar mejorarnos moralmente, estas propuestas deben incluir medios para garantizar que su aplicación no provoque la paradoja de la mejora al generar nuevos riesgos. En el siguiente apartado trataré de argumentar en favor de un tipo de mejora diferente al visto hasta ahora, que no emplea la biotecnología, sino la inteligencia artificial.

4. Necesidad de asistentes morales artificiales: ¿Cómo podría la inteligencia artificial servir para mejorar moralmente a la humanidad?

En la literatura acerca de la mejora moral, recientemente ha aparecido una corriente que se aleja de la biomejora que inicialmente dominó el debate. En esta corriente se apunta al uso de tecnologías computacionales como los sistemas de IA para lograr la mejora moral de los seres humanos. Las propuestas para hacerlo son múltiples, pero lo común a todas ellas es la aceptación de los postulados siguientes:

1. La IA puede servir para evaluar situaciones moralmente conflictivas.
2. Puede tener la capacidad de decidir cursos de acción adecuados para una cierta problemática moral.
3. Puede servir para mejorar moralmente a los seres humanos.

Por ello, en este último apartado exploraré las principales formas de IAmejora propuestas y trataré de defender que algunas de ellas evaden las principales objeciones a la biomejora y pueden servir para la mejora moral de la humanidad.

4.1. ¿Qué es la IAmejora moral?

De igual modo que el uso de psicofármacos o de la ingeniería genética es propuesto para la mejora moral de los seres humanos, también ha sido sugerido para este fin el uso de robots autónomos y sistemas de inteligencia artificial. La idea que hay tras estas propuestas es que las limitaciones de nuestra psicología moral pueden ser mitigadas o suprimidas por medio de herramientas robóticas o computacionales que tengan una mayor capacidad de análisis de información y evaluación de diferentes cursos de acción; por lo que el uso de estas para mejorarnos moralmente puede ser éticamente permisible o incluso obligatorio (Gips, 1995; Dietrich, 2001; Savulescu y Maslen, 2015; Borenstein y Arkin, 2016). Francisco Lara (2021: 5-9) identifica tres modelos de IAmejora:

1. *Ética de las máquinas*: el objetivo de este modelo es desarrollar máquinas y robots autónomos que puedan servir para dirigir el comportamiento humano, bien como ayudantes o consejeros (Gips, 1995: 250) o bien sustituyéndonos completamente (Dietrich, 2001: 326-328).
2. *Empujoncitos (*nudges*) decisionales*: emplear empujoncitos, entendidos como «cualquier aspecto de una arquitectura de elección, o entorno de

toma de decisiones, que tiene como objetivo influir en las personas para que supuestamente tomen mejores decisiones para su bienestar, dejando siempre intacta su libertad de elección» (Lara, 2021: 6, traducción propia), para guiar a los seres humanos sin reducir la autonomía individual. La manera de hacer esto es múltiple, debido a la vaguedad del concepto de *empujoncito*, pero podría ser por medio de robots que promuevan ciertas conductas éticamente deseables en sus acompañantes humanos[8] o que sirvan de educadores y cuidadores de personas para promover su desarrollo moral (Klincewicz, 2019: 442-443).

3. *Asistentes morales artificiales (AMA)*: dentro de este modelo, Lara distingue dos propuestas, a las que podemos añadir la suya. La primera de ellas es la de Savulescu y Maslen (2015: 84-90), que consiste en implementar asistentes virtuales que, a partir de los valores y las circunstancias (estado de ánimo, calidad del sueño, etc.) del agente, sea capaz de ofrecer consejos morales para situaciones particulares o para alcanzar ciertos fines concretos. La segunda es la de Giubilini y Savulescu (2018), que se basa en utilizar un AMA que implemente una versión del observador ideal hipotético —omnisciente, imaginativo, desinteresado, desapasionado y consistente— que ayude al agente a alcanzar un equilibrio reflexivo con sus juicios morales y le asista para tomar decisiones. Por último, la propuesta de Lara (2021) consiste en desarrollar un bot conversacional, un agente moral artificial que cumpla con los criterios de interactividad (ser capaz de responder a inputs ambientales), autonomía (ser capaz de hacer juicios éticos por sí mismo) y adaptabilidad (ser capaz de formular preguntas y sugerencias acerca de nuevas situaciones). El AMA que propone para esta labor es SocrAI, el cual sigue una estrategia híbrida combinando principios de arriba abajo y aprendizaje de abajo arriba[9].

4.2. De la biomejora a la IA para la mejora moral: reversibilidad, implementación progresiva y escalabilidad

Al considerar las IA como un medio para la mejora moral, surgen inevitablemente dos preguntas: la primera de ellas es si el uso de estas herramientas podría mejorar la situación actual, ayudándonos a controlar los efectos perjudiciales de nuestras tecnologías y a evitar los riesgos y los peligros provocados

8. Borenstein y Arkin (2016: 36) ponen de ejemplo el aumento del deseo de una persona de ser más limpia tras adquirir una Roomba, y apuntan a la posibilidad de desarrollar robots que puedan transmitirnos aprobación y desaprobación —verbalmente o de otras formas— para guiarnos hacia un curso de acción deseable sin prohibirnos ninguna conducta concreta.
9. La estrategia híbrida consiste en combinar principios de arriba abajo (codificados en sus algoritmos) con aprendizaje de abajo arriba (a partir de los inputs de diferentes individuos) (Wallach y Allen, 2008: 117-124). Sin embargo, puesto que SocrAI está diseñado para mejorar nuestra deliberación moral, no tendrá integrados principios éticos sustantivos, sino únicamente principios generales de conducta ética que serán ajustados o interpretados para adaptarse a diferentes situaciones, así como guías generales sobre cómo razonar mejor (Lara, 2021: 10).

por estas, y la segunda, si evaden las principales objeciones a la biomejora y son una mejor opción que esta para tal fin. Por ello, ahora trataré de responder ambas cuestiones, centrándome en el último modelo de IAmejora, el de asistentes morales artificiales.

La mejora moral se propone, como vimos, para lidiar con las deficiencias de la psicología humana y con el alcance de las acciones posibilitadas por la tecnología. Otro de los argumentos empleados para defender la mejora moral es la enorme complejidad de nuestros entornos sociotécnicos (Rueda, 2023: 204), que, según algunos teóricos como Monasterio Astobiza (2021: 261), nos obliga a automatizar la vida y el trabajo. Para ello, los AMA pueden ser nuestra mejor opción (Savulescu y Maslen, 2015; Giubilini y Savulescu, 2018; Lara y Deckers, 2020; Lara, 2021). Por un lado, estos respetan y mejoran la autonomía individual, compensando las limitaciones de nuestra psicología moral e implementando la función positiva de las emociones en la moralidad humana sin sus desventajas, como sesgos y prejuicios (Giubilini y Savulescu, 2018: 185-186). Aparentemente podrían hacerlo preservando el pluralismo, sin necesidad de emplear, como en el caso de algunas formas de biomejora, unos valores únicos que dirijan la mejora moral de los seres humanos (Savulescu y Maslen, 2015: 91). Además, aunque fracasemos en nuestros primeros intentos de crear un AMA o aunque nuestros valores cambien en el futuro, previsiblemente ello no tendrá consecuencias tan desastrosas como la biomejora, ya que podremos afinar el software o incluir nuevas funcionalidades a través de actualizaciones. Por ello, el empleo de los AMA para la mejora moral parece más adecuado que el empleo de la biomejora para tal fin.

Por otro lado, algunos modelos de AMA ya están siendo desarrollados e implementados (Lara y Deckers, 2020: 282-283), por lo que es probable que en un corto periodo de tiempo dispongamos de asistentes virtuales efectivos —a diferencia de los principales tipos de biomejora, que, debido a su potencial alcance, deberán probar su seguridad antes de ser empleados en humanos. Estos podrán ser entrenados conforme los usen diferentes agentes, mejorando sus mecanismos de análisis, argumentación y deliberación. Por ello, aunque no podemos anticipar que el uso de estos será suficiente para mejorarnos moralmente en el sentido requerido, tenemos motivos para pensar que su rápida implementación y su adaptabilidad los convertirán en una mejor opción que los principales tipos de biomejora.

Sin embargo, estas propuestas tampoco están exentas de problemas. Probablemente no todos los modelos de AMA sean igualmente deseables. En primer lugar, si la automatización y el empleo de tecnología avanzada fueron los que originaron la necesidad de mejorarnos moralmente debido a los riesgos y problemas a los que nos enfrentamos, parece circular e insatisfactorio el presentar la automatización como la solución a ellos (Rueda, 2023: 9)[10]. En

10. Debe advertirse que esto puede llevarnos por otro camino a la paradoja de la mejora, ya que el desarrollo y la implementación de las IA avanzadas también pueden provocar consecuencias desastrosas y riesgos globales. Sin embargo, en este caso parece más sencillo

segundo lugar, una de las mayores preocupaciones al emplear las IA para cuestiones éticas es la posible existencia de sesgos decisionales y errores en las recomendaciones y predicciones (Rueda, 2023: 9). Si asumimos que nuestra psicología moral es insuficiente para evaluar las implicaciones éticas del empleo de ciertas tecnologías, difícilmente podremos evitar la aparición de sesgos en el desarrollo e implementación de los primeros AMA. Sin embargo, también en este caso tienen ventajas sobre la biomejora, debido a que podemos aspirar a una *implementación progresiva*, asumiendo que en las primeras versiones de un AMA habrá errores que deberán ser subsanados. Además, el uso de *machine learning* podría servir para resolver este problema, haciendo que las sucesivas versiones de un AMA sean entrenadas y auditadas por otras IA que tengan como fin asegurar la inexistencia de sesgos.

Con todo, parece que el uso de los AMA para la mejora moral presenta menos problemas que las propuestas de biomejora y es, por tanto, más adecuada para tal fin. Aunque es pronto para afirmar que el empleo de estas tecnologías provocará una mejora moral suficiente para evitar el cambio climático y el uso de armas de destrucción masiva, tenemos motivos para pensar que en un corto período de tiempo estas podrán, al menos, ayudarnos a frenarlos. Una de las diferencias radicales entre el uso de este tipo de IAmejora y todos los tipos de biomejora es la *escalabilidad* que puede llegar a tener. La naturaleza computacional de los AMA los hace actualizables y permite integrar nuevas tecnologías que mejoren su eficacia conforme sean desarrolladas. Es decir, si bien las primeras versiones de estos asistentes probablemente consistan en aplicaciones web que funcionen como bots conversacionales, sus sucesivas versiones irán incluyendo funcionalidades y tecnologías más avanzadas, como dispositivos de realidad aumentada o virtual que permitan al agente experimentar los efectos de sus posibles decisiones (Lara y Deckers, 2020: 285) y mejorar su empatía (Rueda y Lara, 2020; Lara, 2021: 15-16); sensores que permitan monitorear la salud, el ambiente o la psicología del agente que los emplee, con el fin de alertarle de factores que puedan perjudicar su juicio o su comportamiento moral (Savulescu y Maslen, 2015: 84-86), o bases de datos que permitan contrastar en tiempo real los sistemas de valores, preferencias y decisiones de diferentes agentes o grupos a partir de los inputs que realicen en sus conversaciones con los AMA, con el fin de encontrar puntos de unión entre los distintos sistemas de valores existentes para resolver problemas de cooperación[11]. A su vez, la IAmejora abre la posibilidad de extrapolar el apren-

evitarlo que con la biomejora, debido a que se podrían desarrollar mecanismos de seguridad y de *reversibilidad*, por si los softwares empleados para la mejora moral se mostrasen perjudiciales.

11. Un AMA podría recoger las diferentes respuestas de los distintos agentes que lo empleen, así como de distintas páginas webs (*e. g.* foros, redes sociales), categorizando los valores que se desprendan de ellas de forma anonimizada para elaborar bases de datos que contengan información acerca de los valores, las creencias y las preferencias compartidas por diferentes individuos y grupos. Esto serviría para adaptar las funcionalidades y las características de los distintos AMA a los diferentes entornos culturales en los que sean empleados. A su vez, nos permitiría contrastar y comparar los valores y las creencias de distintos individuos y

dizaje moral de un individuo al de otros muchos, mejorando enormemente la eficacia de los medios tradicionales para la mejora moral como la educación. Además, al incidir sobre nuestras ideas y valores, pueden servir para crear uniones transnacionales, contribuyendo a construir nuevas historias sobre nosotros mismos (Coeckelbergh, 2022: 127) que sirvan para unir a las distintas sociedades bajo un proyecto común[12].

4.3. Hacia un modelo de IAmejora moral interactivo: el caso de SocrAI

Para terminar, me gustaría aplicar las nociones que he ido destacando —reversibilidad, implementación progresiva y escalabilidad— para señalar las ventajas de la IAmejora frente a la biomejora al modelo de AMA descrito por Lara y Deckers (2020) o por Lara (2021), que además de salvaguardar la mayoría de las virtudes de la IAmejora evade numerosos inconvenientes de otras formas de esta.

Lara y Deckers (2020) defienden un modelo interactivo de IAmejora en el que la mejora moral del agente se produzca a través del cambio de sus valores por medio del diálogo. Para ello, proponen SocrAI, un bot conversacional (Lara, 2021: 9) que pueda recibir, a través de ordenadores, dispositivos de realidad virtual o interfaces cerebrales, información de diferentes bases de datos sobre ciencia, lingüística, lógica y sobre cómo la gente piensa y razona moralmente (Lara y Deckers, 2020: 284). Este bot monitoreará la biología y el ambiente del agente a través de sensores, procesando esta información para entablar una conversación con él a través de un asistente de voz virtual, en la que tratará de mejorar su razonamiento moral. Creen que su propuesta tiene dos ventajas fundamentales sobre otras formas de biomejora e IAmejora (Lara y Deckers, 2020: 282):

a) Que en ella la participación del agente es mayor que en el de otras propuestas, y argumentan que por ello aumenta su autonomía.
b) Que pone el énfasis en el rol formativo que la máquina tiene para el agente, en vez de en el resultado que provoca.

De esta manera, SocrAI debe ser desarrollado de forma que únicamente sirva para ayudar al agente procedimentalmente en su deliberación, sin tener ningún sistema de valores sustantivo integrado. Las funciones que proponen para dicho modelo son las siguientes (Lara y Deckers, 2020: 283-284):

---

sociedades por medio de otras IA y expertos humanos para buscar nuevos puntos de unión entre ellas, con el fin de alcanzar consensos transnacionales que nos permitan resolver problemas de cooperación, como evitar riesgos globales.

12. Esto podría contribuir al objetivo defendido por Javier Rodríguez-Alcázar y Lilian Bermejo-Luque (2024) en este volumen de que ciertos debates (como el de la mejora humana), que actualmente están dominados por los Estados, se abran a una comunidad política más amplia que incluya a toda la humanidad.

1. Proveer soporte empírico.
2. Mejorar la claridad conceptual.
3. Entender la lógica argumentativa.
4. Probar si un juicio tiene plausibilidad ética.
5. Crear conciencia sobre las limitaciones personales.
6. Asesorar sobre cómo ejecutar una decisión.

A estas funcionalidades Lara (2021: 15) añadirá un séptimo punto: poseer empatía cognitiva para tratar de mejorar la capacidad motivadora de SocrAI.

Para que una forma de mejora moral sea adecuada y suficiente, esta debe, en primer lugar, no provocar problemas de magnitud similar a los que pretende solventar y, en segundo lugar, servir para tratar los problemas que intenta resolver. A lo largo de este trabajo defendí que la mayoría de las formas de biomejora que es esperable que podamos emplear en un corto período de tiempo a escala global no cumplen la primera condición. En lo que resta de este apartado trataré de defender que SocrAI puede ser una mejor alternativa a ellas.

En principio, SocrAI evade los principales peligros de la biomejora por no ser tan invasiva fisiológica y psicológicamente. Pero hay otros dos motivos para confiar en que este AMA no supondrá grandes riesgos en su implementación. El primero de ellos es que esta IA puede implementarse en sus primeras versiones de manera reversible. Es decir, que una vez sea puesta en marcha pueda ser desconectada o inutilizada si se probase perjudicial en algún sentido. El segundo motivo es que su implementación puede (y debe) ser progresiva, aumentando sus funcionalidades y capacidades conforme se vaya probando su seguridad.

La pregunta que surge entonces es si previsiblemente este AMA servirá para mejorarnos moralmente de manera suficiente para afrontar el nuevo abanico de riesgos y problemas éticos que se nos presentan. Mientras que los defensores de la biomejora han enfatizado el componente motivacional de la mejora moral, Lara y Deckers (2020) enfatizan el aspecto cognitivo de la deliberación ética, aunque Lara (2021) también defiende que SocrAI podría motivar a la acción a un agente aún más que otro agente humano. Estos dos aspectos parecen claramente relevantes para la acción ética, pero creo que no tienen en cuenta un aspecto que condiciona enormemente la acción humana: los incentivos sociales y culturales. El conjunto de prácticas, comportamientos y actitudes vistos como deseables o indeseables en un entorno cultural concreto, así como las instituciones y las estructuras sociales aceptadas, condicionan enormemente la manera en la que los seres humanos actuamos.

Sin embargo, los AMA como SocrAI pueden llegar a tener efectos en los incentivos sociales y en las cuestiones institucionales. Aunque un bot conversacional por sí solo quizás no pueda hacer más que mejorar moralmente a algunos individuos, la implementación de varias IA de este tipo tendrá efectos enormemente positivos a nivel estructural gracias a la escalabilidad que permite este tipo de tecnología. Si se logra implementar un AMA como SocrAI

que mejore la deliberación ética de los seres humanos a nivel individual, es probable que su uso pase a ser visto como socialmente deseable. Ello generará un gran incentivo para utilizar asistentes éticos de este tipo y compartir el aprendizaje moral del que nos provean con nuestros allegados, a través de redes sociales y otros medios, por lo que es esperable que su uso escale rápidamente una vez estén disponibles[13]. Por otro lado, las tecnologías que servirán para aumentar el potencial de este AMA irán desarrollándose progresivamente, facilitando que este modelo sea también tecnológicamente escalable. Si bien las primeras versiones serán asistentes virtuales que podamos utilizar desde nuestros ordenadores y teléfonos móviles, las subsiguientes versiones irán empleando tecnología cada vez más inmersiva, bien sea realidad aumentada o virtual o interfaces cerebro-máquina. Por ello este modelo de mejora moral permite, por un lado, realizar una rápida implementación, ya que no plantea grandes problemas éticos ni requiere un avance tecnológico significativo, y, por el otro, un gran alcance futuro, conforme su uso se vaya extendiendo y se desarrollen tecnologías que doten a estas máquinas de más capacidad.

Un ejemplo de cómo SocrAI u otros AMA pueden tener efectos en las cuestiones estructurales que dificultan la reducción de riesgos consiste en la verificación de noticias falsas (*fake news*) y desinformación, difundidas por redes sociales y medios de comunicación. La inacción frente a problemas globales como el cambio climático no proviene únicamente de fallos morales o motivacionales. En buena medida, esta está relacionada con las creencias aceptadas en los diferentes entornos políticos, sociales y culturales (Hornsey y Fielding, 2020: 5-10). Pese al amplio consenso científico existente acerca de los perjuicios que puede provocar el cambio climático, actualmente hay multitud de individuos, organizaciones políticas y agencias de información que tratan de negar su existencia, sus implicaciones o su relación con la acción humana (López, 2021: 289-291). La desinformación acerca de estas cuestiones impide alcanzar consensos para resolver problemas de cooperación como los que suponen la reducción de riesgos globales. Por este motivo, el empleo de diversos AMA como SocrAI para verificar información puede contribuir enormemente a facilitar que los gobiernos de las sociedades democráticas tomen medidas más fuertes para tratar de mitigar problemas que tendrán consecuencias a muchos años vista, como la degradación ambiental. El hecho de proveer soporte empírico en tiempo real facilita que desestimar la importancia de

13. No obstante, como argumenta Lydia Feito (2024) en este mismo número con respecto a la biomejora, la aparición de tecnologías de perfeccionamiento humano puede generar tantos incentivos por emplearlas que termine provocando presión social para mejorarnos o suponiendo la exclusión y discriminación de quienes rechacen la mejora, lo que podría convertirse en una forma de deshumanización y de tiranía. Este problema, que ella señala respecto a las propuestas de mejora humana por medios biotecnológicos, podría aparecer también en las propuestas de mejora moral por medios computacionales, haciendo que su implementación provoque presión social, menoscabando nuestra autonomía o dando lugar a nuevas formas de discriminación. Por este motivo considero que, de igual modo que con las propuestas de biomejora moral, también en las propuestas de IAmejora moral debemos priorizar la seguridad de su uso frente a la velocidad de su implementación.

tomar medidas frente a estas cuestiones sea una estrategia muy perjudicial para los diferentes actores políticos y medios de comunicación, puesto que cambiaría los incentivos estructurales en las sociedades democráticas para tomar medidas políticas frente a estos problemas.

5. Posibles problemas de un asistente socrático virtual

Evidentemente, aunque SocrAI parece más seguro que muchas de sus alternativas para la mejora moral, es esperable que este y otros tipos de AMA generen ciertos riesgos y problemas. A su vez, aunque defiendo que esta y otras formas de IAmejora darán lugar a un avance moral relevante, hay quienes piensan que este no será suficiente, por lo que ahora pasaré a analizar las críticas más relevantes a este modelo y trataré de contestarlas.

En primer lugar, Monasterio Astobiza (2021: 271-279) señala tres problemas generales que deberán enfrentar los modelos de IAmejora. El primero de ellos es el *problema del pluralismo axiológico*, que se refiere a la dificultad de programar una IA para que tenga en cuenta una rica variedad de valores presentes en la moralidad humana; el segundo, el *problema de la evitabilidad*, que se refiere a la dificultad de impedir que la IA nos manipule o engañe en la toma de decisiones, y el tercero es el *problema de la atrofia moral*, que indica el riesgo de que el uso habitual de estos asistentes artificiales pueda perjudicar nuestras habilidades morales a largo plazo (Rueda, 2023: 6). Aunque considero que no tenemos manera de garantizar que no se produzcan estos problemas, SocrAI, a diferencia de otros modelos de IAmejora, mitiga enormemente todos ellos. El problema del pluralismo axiológico es mucho más relevante para aquellas IA que deben tomar decisiones autónomas u ofrecer guías de acción a un agente humano. En el caso de las IA como SocrAI, que únicamente tiene como objetivo poner en cuestión los valores del propio agente y mejorar su capacidad de deliberación ética, probablemente este no sea un gran problema, debido a que no necesitará amoldarse completamente a los valores de ningún agente particular, sino únicamente a una moralidad común[14]. El problema de la evitabilidad también es mitigado en la medida en que SocrAI no tratará de convencer positivamente al agente para que tome una u otra decisión. Al ser simplemente un asistente para la reflexión moral, el riesgo de que lleve a un agente a realizar una acción perjudicial para él mismo o para otros es previsiblemente bajo. Aun así, es posible que SocrAI se exceda en sus funciones al buscar formas de optimizar la educación moral que provee y consecuentemente termine ofreciendo consejos o guías que en un principio no estaba previsto que brindara. Para resolver este problema creo que la mejor opción que tenemos es establecer mecanismos sólidos de reversibilidad, a fin de poder deshabilitar una versión concreta de la IA o todas ellas, si se muestran perju-

14. Esta moralidad común consistiría en principios ampliamente aceptados por diversas sociedades humanas, como los Derechos Humanos, el principialismo bioético o la Regla de Oro.

diciales. Por último, el problema de la atrofia moral quizás no afecte sustancialmente a SocrAI, ya que en principio este AMA tiene como objetivo mejorar las vías de aprendizaje moral tradicionales como la educación. Al depender la mejora moral del agente humano, y no de seguir una serie de consejos predeterminados por la máquina, su capacidad de reflexión, deliberación y acción éticas puede verse, antes que atrofiada, mejorada.

En segundo lugar, aunque para algunos racionalistas como Harris (2011) o Gordon y Ragonese (2023) las razones son motivadoras en sí mismas, para otros teóricos como Douglas (2008) o DeGrazia (2014) estas no son suficientes. Por este motivo, una propuesta como la de SocrAI, que únicamente pretende mejorar nuestra deliberación moral, parece insuficiente para muchos defensores de la biomejora. Lara y Deckers (2020: 285) aceptan este problema, aunque creen que SocrAI podría motivar indirectamente al modificar los valores del agente, haciendo que este tenga motivos para cambiar su manera de actuar. Además, Lara (2021: 20-22) añade que su capacidad motivadora podría verse aumentada incluyendo la funcionalidad de la empatía cognitiva y a través del empleo de sistemas inmersivos como la realidad virtual, que permitan al agente experimentar los efectos de sus decisiones. A esto podemos añadir lo expuesto arriba acerca de cómo los incentivos sociales condicionan la acción humana. Si SocrAI u otros AMA llegan a ser socialmente aceptados, su uso pasará a ser visto como algo digno de aprobación moral. Gracias a la escalabilidad de esta y otras tecnologías, sería relativamente fácil implementar bonificaciones como sistemas de reputación virtuales que promuevan beneficios a quienes las empleen para mejorar su deliberación y su comportamiento ético[15]. Ejemplos de estos podrían ser descuentos y reducción de precios en comercios y empresas que promuevan su uso estableciendo consorcios para beneficiar a aquellas personas que los usen, reducciones fiscales o de impuestos gubernamentales, o exposición social y mediática para aquellas personas que realicen acciones meritorias. Todo ello, aunque no incide directamente en los mecanismos de motivación intrínseca del agente, modifica el contexto y los incentivos sociales para la acción, creando una motivación extrínseca para comportarse de manera moralmente adecuada.

Por último, es probable que, conforme desarrollemos e implementemos esta y otras tecnologías para la mejora moral, aparezcan nuevas críticas a estas

15. Inversamente, podrían emplearse para desincentivar a los individuos o a las organizaciones que realizaran acciones moralmente cuestionables, como difundir desinformación o provocar grandes perjuicios medioambientales. Si se integraran estos asistentes en las páginas web de medios de comunicación, redes sociales o empresas, servirían para contrastar información en tiempo real. Esto haría más transparente la información en la red, puesto que permitiría a los distintos agentes tomar decisiones más informadas, al provocar que la desinformación y las noticias falsas fueran detectadas al momento, lo que generaría incentivos para quienes las difundieran por contrastar la información que compartieran, ya que de otro modo perderían credibilidad. Asimismo, se podrían emplear para reducir la asimetría de información entre empresas y consumidores, lo que permitiría una mayor trazabilidad de los productos contaminantes y la detección de estrategias de marketing engañosas, como la ecoimpostura (*greenwashing*).

formas de IAmejora[16]. Y ello probablemente sea positivo, ya que nos ayudará a aumentar la seguridad de estas tecnologías y su eficacia. Pero, por otro lado, muchos de los aspectos que ahora consideramos problemáticos de esta y otras formas de mejora moral (como que reducen la autonomía o pueden provocar atrofia moral) quizás en un futuro dejen de serlo. El cambio de valores necesario para enfrentarnos a los problemas provocados por algunas tecnologías, así como el cambio socioeconómico y político necesario para lograrlo, previsiblemente harán que en las próximas décadas y siglos pasemos a aceptar valores nuevos y a rechazar otros viejos. Además, a esto contribuirá el desarrollo de esta y otras tecnologías, en la medida en que, como afirma Coeckelbergh (2022), estas influyan en las narrativas con las que creamos nuestra identidad y entendemos nuestro mundo. Por ello, antes que elaborar un programa completo para evitar riesgos como el cambio climático o el uso de armas de destrucción masiva, considero que puede ser más fructífero ir dando pequeños pasos, ya que a cada uno de ellos se abrirán más posibilidades para evitarlos. A medida que avancemos, lo primordial será garantizar la seguridad de las tecnologías que empleemos para ello, con el fin de evitar la paradoja de la mejora. En este sentido, SocrAI evade los principales problemas de otras formas de mejora moral, y su implementación evade muchos de los grandes problemas de las tecnologías para esta empresa.

6. Conclusión

Los seres humanos, gracias a nuestra capacidad tecnológica, hemos alcanzado un poder sin precedentes de alterar la naturaleza y de afectar a multitud de individuos, lo que ha sido enormemente beneficioso para la calidad de vida de un gran número de personas. Sin embargo, también ha provocado nuevos riesgos y problemas globales que no estamos afrontando adecuadamente, como el cambio climático o el posible uso de armas de destrucción masiva. Esto ha hecho que se haya comenzado a defender la mejora moral de los seres humanos por medios tecnológicos. Una de las propuestas más notorias en este sentido es la que trata de emplear medios biotecnológicos como neurotecnologías o ingeniería genética para tal fin.

En este trabajo he tratado de mostrar que, si asumimos que somos inadecuados para resolver muchos de los problemas generados por las nuevas tecnologías, no parece que sea adecuado emplear biotecnologías con un gran alcance (como la ingeniería genética o el uso masivo de psicofármacos) para resolverlos, ya que pueden tener efectos ampliamente perjudiciales. Esto es lo que llamo la «paradoja de la biomejora», que consiste en que las mismas biotecnologías que parecen necesarias para mejorarnos moralmente pueden servir a ciertos grupos para provocar grandes perjuicios y riesgos globales, debido a que carecemos de las capacidades morales necesarias para evitar que sean mal

16. En Formosa y Ryan (2021) puede encontrarse una revisión sistemática de las críticas a los AMA y una respuesta a estas.

empleadas. Por ello, he criticado la mayoría de las propuestas de biomejora, por ser aparentemente inadecuadas para lograr evitar problemas como el cambio climático o el uso de armas de destrucción masiva.

He tratado de argumentar que hay formas de mejora moral que evaden este problema. Entre ellas, algunas formas de IAmejora, como el uso de asistentes morales artificiales, suponen menos riesgos que la biomejora. Estos servirán para optimizar nuestra deliberación y (en menor medida) motivación moral y, a su vez, defendí que podrían servir para alcanzar mayores consensos transnacionales y alterar los incentivos estructurales que impiden afrontar ciertos riesgos globales. Gracias a la reversibilidad, la implementación progresiva y la escalabilidad de las tecnologías para este tipo de mejora moral, es posible que sirvan para mejorar la conducta humana sin provocar riesgos ni problemas de gran envergadura. Por ello, he defendido una forma de IAmejora como medio para progresar moralmente y poder evaluar con mejor criterio otras formas de mejoramiento moral más problemáticas. Esta es conocida como SocrAI, un bot conversacional concebido para mejorar las capacidades de deliberación moral humanas. A su vez, he defendido que este modelo puede ser ampliado y complementado con otros avances tecnológicos que mejoren su eficacia. Por último, he tratado de responder a las principales críticas a este modelo de IAmejora.

Referencias bibliográficas

Almeida, Mara y Diogo, Rui (2019). «Human enhancement: Genetic engineering and evolution». *Evolution, medicine, and public health*, 1, 183-189. <https://doi.org/10.1093/emph/eoz026>

Alonso, Marcos (2024). «Post genetic revolution dynamics. How will modified and unmodified humans coexist?». *Enrahonar. An International Journal of Theoretical and Practical Reason*, 72, 35-54. <https://doi.org/10.5565/rev/enrahonar.1527>

Anders, Günther (2011). *La obsolescencia del hombre: Sobre el alma en la época de la segunda revolución industrial*. Vol. 1. Traducido por Josep Monter Pérez. Valencia: Pre-Textos.

Bacon, Francis (1620). *Novum Organum*. Traducido por Cristóbal Litrán. Barcelona: Orbis, 1984.

Boenig-Liptsin, Margarita y Hurlbut, J. Benjamin (2016). «Technologies of Transcendence at Singularity University». En: Hurlbut, J. Benjamin y Tirosh-Samuelson, Hava (eds.). *Perfecting Human Futures: Transhumanist Visions and Technological Imaginations*. Nueva York: Springer, 239-268.

Borenstein, Jason y Arkin, Ron (2016). «Robotic nudges: The ethics of engineering a more socially just human being». *Science and Engineering Ethics*, 22, 31-46. <https://doi.org/10.1007/s11948-015-9636-2>

Bostrom, Nick (2002). «Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards». *Journal of Evolution and Technology*, 9. Recuperado de <https://jetpress.org/volume9/risks.html>

— (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bostrom, Nick y Cirkovic, Milan M. (2011). *Global catastrophic risks*. Oxford: Oxford University Press.

Boyd, Matt y Wilson, Nick (2020). «Existential Risks to Humanity Should Concern International Policymakers and More Could Be Done in Considering Them at the International Governance Level». *Risk Analysis*, 40, 2303-2312.
<https://doi.org/10.1111/risa.13566>

Coeckelbergh, Mark (2022). *Self-Improvement: Technologies of the soul in the age of artificial intelligence*. Nueva York: Columbia University Press.

DeGrazia, David (2014). «Moral enhancement, freedom, and what we (should) value in moral behaviour». *Journal of Medical Ethics*, 40 (6), 361-368.
<https://doi.org/10.1136/medethics-2012-101157>

Descartes, René (1637). «Discurso del método para dirigir adecuadamente la razón e investigar la verdad en las ciencias. En: *Discurso del método, dióptrica, meteoros y geometría*. Traducido por Guillermo Quintás Alonso, 2.ª ed., 1-57. Madrid: Alfaguara, 1987.

Dietrich, Eric (2001). «Homo Sapiens 2.0: Why we should build the better robots of our nature». *Journal of Experimental and Theoretical Artificial Intelligence*, 13 (4), 323-328.

Douglas, Thomas (2008). «Moral Enhancement». *Journal of Applied Philosophy*, 25, 228-245.
<https://doi.org/10.1111/j.1468-5930.2008.00412.x>

— (2013). «Moral enhancement via direct emotion modulation: A reply to John Harris». *Bioethics*, 27 (3), 160-168.
<https://doi.org/10.1111/j.1467-8519.2011.01919.x>

— (2015). «The Morality of Moral Neuroenhancement». En: Clausen, Jens y Levy, Neil (ed.). *Handbook of Neuroethics*. Nueva York: Springer, 1227-1249.

Earp, Brian D.; Douglas, Thomas y Savulescu, Julian (2017). «Moral Neuroenhancement». En: Johnson, L. Syd y Rommelfanger, Karen S. (ed.). *The Routledge Handbook of Neuroethics*. Londres: Routledge, 166-185.

Feito, Lydia (2024). «La libertad de elegir o la tiranía de la mejora». *Enrahonar. An International Journal of Theoretical and Practical Reason*, 72, 91-111.
<https://doi.org/10.5565/rev/enrahonar.1531>

Formosa, Paul y Ryan, Malcolm (2021). «Making moral machines: Why we need artificial moral agents». *AI & Society*, 36, 839-851.
<https://doi.org/10.1007/s00146-020-01089-6>

Gergan, Mabel; Smith, Sara y Vasudevan, Pavithra (2020). «Earth beyond repair: Race and apocalypse in collective imagination». *Environment and Planning D: Society and Space*, 38 (1), 91-110.
<https://doi.org/10.1177/0263775818756079>

Gips, James (1995). «Towards the ethical robot». En: Ford, Kenneth M.; Glymour, Clark y Hayes, Patrick (ed.). *Android epistemology*. Cambridge, MA: MIT Press, 243-252.

Giubilini, Alberto y Savulescu, Julian (2018). «The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence». *Philosophy & Technology*, 31, 169-188.
<https://doi.org/10.1007/s13347-017-0285-z>

Gleiberman, Mollie (2023). «Effective Altruism Doing transhumanism better (Working Paper)». *Institute of Development Policy*. Recuperado de <https://medialibrary.uantwerpen.be/files/8518/eb53f30f-c469-400e-a9c1-8fb43a52bcc7.pdf>

Glover, Jonathan (2006). *Choosing Children: The Ethical Dilemmas of Genetic Intervention*. Oxford: Oxford University Press.

Gordon, Emma C. y Ragonese, Viola (2023). «Cognitive and Moral Enhancement: A Practical Proposal». *Journal of Applied Philosophy*, 40 (3), 474-487.
<https://doi.org/10.1111/japp.12619>

Häggström, Olle (2016). *Here be dragons: Science, technology and the future of humanity*. Oxford: Oxford University Press.

Harris, John (2011). «Moral enhancement and freedom». *Bioethics*, 25 (2), 102-111.
<https://doi.org/10.1111/j.1467-8519.2010.01854.x>

Hornsey, Matthew y Fielding, Kelly (2020). «Understanding (and Reducing) Inaction on Climate Change». *Social Issues and Policy Review*, 14, 3-35.
<https://doi.org/10.1111/sipr.12058>

Jonas, Hans (1984). *The imperative of responsibility: In search of an ethics for the technological age.* Chicago: University of Chicago Press.

Klincewicz, Michal (2019). «Robotic nudges for moral improvement through stoic practice». *Techné: Research in Philosophy and Technology*, 23 (3), 425-455.
<https://doi.org/10.5840/techne2019122109>

Lara, Francisco (2017). «Oxytocin, Empathy and Human Enhancement». *THEORIA: An International Journal for Theory, History and Foundations of Science*, 32 (3), 367-384.
<https://doi.org/10.1387/theoria.17890>

— (2021). «Why a Virtual Assistant for Moral Enhancement When We Could have a Socrates?». *Science and Engineering Ethics*, 27 (4).
<https://doi.org/10.1007/s11948-021-00318-5>

Lara, Francisco y Deckers, Jan (2020). «Artificial Intelligence as a Socratic Assistant for Moral Enhancement». *Neuroethics*, 13, 275-287.
<https://doi.org/10.1007/s12152-019-09401-y>

López, María Ángela (2021). «El cambio climático: Negacionismo, escepticismo y desinformación». *Tabula Rasa*, 37, 283-301.
<https://doi.org/10.25058/20112742.n37.13>

Macpherson, Ignacio; Roqué, María Victoria y Segarra, Ignacio (2019). «Moral enhancement, at the peak of pharmacology and at the limit of ethics». *Bioethics*, 33, 992-1001.
<https://doi.org/10.1111/bioe.12613>

Mitchell, Audra y Chaudhury, Aadita (2020). «Worlding beyond 'the' 'end' of 'the world': White apocalyptic visions and BIPOC futurisms». *International Relations*, 34 (3), 309-332.
<https://doi.org/10.1177/0047117820948936>

Monasterio Astobiza, Aníbal (2021). «Automatizando la toma de decisiones morales: Inteligencia artificial y mejora humana». En: Lara, Francisco y Savulescu, Julian. *Más (que) humanos: Biotecnología, inteligencia artificial y ética de la mejora*. Madrid: Tecnos, 255-283.

Ord, Toby (2020). *The Precipice: Existential Risk and the Future of Humanity*. Londres: Bloomsbury Publishing.

Paulo, Norbert y Bublitz, Jan Christoph (2019). «How (not) to Argue For Moral Enhancement: Reflections on a Decade of Debate». *Topoi*, 38, 95-109.
<https://doi.org/10.1007/s11245-017-9492-6>

Persson, Ingmar y Savulescu, Julian (2008). «The perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity». *Journal of Applied Philosophy*, 25 (3), 162-177.
<https://doi.org/10.1111/j.1468-5930.2008.00410.x>

— (2010). «Moral Transhumanism». *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, 35 (6), 656-669.
<https://doi.org/10.1093/jmp/jhq052>

— (2012). *Unfit for the future: The need for moral enhancement*. Oxford: Oxford University Press.

Rodríguez-Alcázar, Javier y Bermejo-Luque, Lilian (2024). «Human enhancement technologies and the arguments for cosmopolitanism». *Enrahonar. An International Journal of Theoretical and Practical Reason*, 72, 15-33.
<https://doi.org/10.5565/rev/enrahonar.1489>

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

— (2023). «¿Automatizando la mejora moral humana?: La inteligencia artificial para la ética». [Nota crítica sobre Lara, Francisco y Savulescu, Julian (eds.) (2021). «Más (que) humanos: Biotecnología, inteligencia artificial y ética de la mejora». Madrid: Tecnos.] *Daimon Revista Internacional de Filosofía*, 89, 199-209.
<https://doi.org/10.6018/daimon.508771>

— (2024). «The global governance of genetic enhancement technologies: Justification, proposals, and challenges». *Enrahonar. An International Journal of Theoretical and Practical Reason*, 72, 55-71.
<https://doi.org/10.5565/rev/enrahonar.1519>

Rueda, Jon y Lara, Francisco (2020). «Virtual Reality and Empathy Enhancement: Ethical Aspects». *Frontiers in Robotics and AI*, 7 (506984). <https://doi.org/10.3389/frobt.2020.506984>

Savulescu, Julian y Maslen, Hannah (2015). «Moral enhancement and artificial intelligence: moral AI?». En: Romportl, Jan; Zackova, Eva y Kelemen, Jozef (ed.). *Beyond artificial intelligence: The disappearing human-machine divide*. Nueva York: Springer, 79-95.

Simpson, Michael (2020). «The Anthropocene as colonial discourse». *Environment and Planning D: Society and Space*, 38 (1), 53-71. <https://doi.org/10.1177/0263775818764679>

Sparrow, Robert (2014). «Better Living Through Chemistry?: A Reply to Savulescu and Persson on "Moral Enhancement"». *Journal of Applied Philosophy*, 31 (1), 23-32. <https://doi.org/10.1111/japp.12038>

— (2015). «Enhancement and Obsolescence: Avoiding an "Enhanced Rat Race"». *Kennedy Institute of Ethics Journal*, 25 (3), 231-260. <https://doi.org/10.1353/ken.2015.0015>

Torres, Phil (2017). «Moral bioenhancement and agential risks: Good and bad outcomes». *Bioethics*, 31 (9), 1-6. <https://doi.org/10.1111/bioe.12389>

Walker, Mark (2009a). «Ship of fools: Why transhumanism is the best bet to prevent the extinction of civilization». *The Global Spiral*, 9 (9). Recuperado de <https://metanexus.net/h-ship-fools-why-transhumanism-best-bet-prevent-extinction-civilization/>

— (2009b). «Enhancing genetic virtue: A project for twenty-first century humanity?». *Politics and the Life Sciences*, 28 (2), 27-47. <https://doi.org/10.2990/28_2_27>

Wallach, Wendell y Allen, Colin (2008). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Pablo Neira Castro es candidato predoctoral en la Universidad de Granada y contratado con cargo al proyecto AutAI (Inteligencia artificial y autonomía humana. Hacia una ética para la protección y mejora de la autonomía en sistemas recomendadores, robótica social y realidad virtual). Sus principales líneas de investigación son la ética de la mejora humana, la ética de la inteligencia artificial y la reducción de riesgos globales.

Pablo Neira Castro is a predoctoral candidate at the University of Granada, working on the AutAI project (Artificial intelligence and human autonomy. Towards an ethics for the protection and enhancement of autonomy in recommender systems, social robotics and virtual reality). His main lines of research are the ethics of human enhancement, the ethics of artificial intelligence, and global risk reduction.