ISSN: 2695-480X
SLMFCE

Revista de la Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

Número 68 **Febrero de 2024**

Contenido:

ISSN: 2695-480X
SLMFCE

Revista de la Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

General Número 68 Febrero de 2024

Editorial

Este número de la Revista SLMFCE incluye contenidos que creemos de mucho interés. En el marco del acuerdo de colaboración entre nuestra Sociedad y SWIP-Analytic España, publicamos la entrevista que María José Frápolli ha realizado a Francisca Pérez Carreño, catedrática de Estética y Teoría de las Artes en la Universidad de Murcia. La entrevista da expresión a la *Distinción a una Trayectoria* que SWIP-A España otorga como reconocimiento y homenaje a una trayectoria excepcional. Las investigadoras jóvenes, en particular, pueden encontrar en la profesora Pérez Carreño (Paca, para quienes la conocemos desde hace tiempo) un modelo inspirador para sus propias carreras. Además, sus respuestas y reflexiones ayudan a conocer mejor la vida académica española en un periodo de dificultad pero también motivador. Y se hace posible, sobre todo, conocer mejor la importante contribución de Francisca Pérez Carreño al pensamiento y al impulso de la investigación sobre Estética y Teoría de las Artes.

Este número incluye también, por iniciativa de nuestro editor David Pérez Chico, una sección monográfica sobre las consecuencias de la Inteligencia Artificial y la singularidad tecnológica. En los últimos años ha habido un interés creciente por reflexionar sobre lo que se conoce como la *singularidad* de la IA: el momento en que, en un futuro no lejano, la inteligencia artificial sería superior a la inteligencia humana. Ese momento se alcanzaría cuando los sistemas de AI fueran capaces de desarrollar a su vez sistemas de AI cada vez más inteligentes. Estos sistemas "superinteligentes", de acuerdo con los pronósticos más catastrofistas, quedarían fueran del control humano y las consecuencias finales serían impredecibles. El conjunto de trabajos que componen esta sección ofrecen una información muy actualizada del estado del debate; analizan con atención y criterio los argumentos disponibles, y ofrecen puntos de vista reflexivos y equilibrados que ayudan a conocer mejor lo que podemos esperar en el futuro.

La Revista incluye además las crónicas de los eventos académicos en los que han tomado parte jóvenes investigadores e investigadoras que pudieron optar a las ayudas de viaje que concede la SLMFCE. En conjunto, ponen de manifiesto la amplitud de intereses y el alto grado de internacionalización de la investigación más joven en nuestro ámbito. Y es posible encontrar, finalmente, las reseñas de algunos libros de interés para nuestra comunidad de especialistas.

Desde la Junta directiva de la SLMFCE agradecemos muy sinceramente la generosa participación de quienes han contribuido a este número con trabajos de mucha calidad y gran interés. Agradecemos también la colaboración eficiente de nuestra maquetista, Patricia García Rodríguez. Y confiamos en que la lectura de la Revista, además de resultar interesante, sirva de inspiración para nuevos estudios e investigaciones.

Cristina Corredor Lanas
Presidenta de la SLMFCE

www.solofici.org

SOCIEDAD DE LÓGICA, METODOLOGÍA Y FILOSOFÍA DE LA CIENCIA EN ESPAÑA

DISTINCIÓN A UNA TRAYECTORIA

La Sociedad Española de Filósofas Analíticas (SWIPA), (http://swipa.ugr.es/), instauró en 2022 la *Distinción a una Trayectoria* con el fin de homenajear y poner en valor la carrera académica de filósofas extraordinarias. En su tercera edición, la distinción ha recaído en Francisca Pérez Carreño, Catedrática de Estética y Teoría de las Artes de la Universidad de Murcia. Le damos desde aquí la enhorabuena por esta distinción, absolutamente merecida.

Desde SWIPA creemos que es esencial para entender la historia reciente de la filosofía en España conocer el trabajo de algunas mujeres que, sin el apoyo social e institucional que ahora nos parece normal, consiguieron llevar a cabo carreras brillantes, formarse en el extranjero y tirar adelante con sus vidas. Si eso ahora es una tarea titánica, recomendamos que se dedique un momento a pensar las dificultades que estas filósofas enfrentaron cuando el concepto *conciliación familiar* no existía, cuando el beso no consentido a Jenny Hermoso se hubiera visto como una gracieta entrañable y cuando, como Paca Pérez Carreño dice en su entrevista, un filósofo era un señor con gafas de pasta fumando en pipa (los "modernos", añadiríamos, porque los clásicos fumaban cigarrillos). En esas condiciones, algunas mujeres llegaron a ser referentes en sus temas, como es el caso de nuestra *distinguida* de 2024, hacerse con un nombre en la esfera internacional y llegar a catedráticas de universidad rompiendo con unas reglas no escritas que parecían inamovibles.

Francisca Pérez es Decana de la Facultad de Filosofía de la Universidad de Murcia. Desde 2001 es Investigadora Principal del Grupo ARESMUR (http://www.um.es/aresmur/home/).

Fue presidenta de la Sociedad Europea de Estética (2015-2022) y miembro de su Comité Ejecutivo (2008-2012), y miembro del Comité Ejecutivo de la Sociedad Española de Estética y Teoría de las Artes (2011-2018).

Pérez Carreño trabaja principalmente en estética filosófica y teoría del arte contemporáneo. Ha publicado una monografía sobre la filosofía de la imagen (*Los placeres del parecido. Icono y representación*) y varios artículos sobre representación pictórica y expresión ("Looking at metaphors", "Two routes to expressio in painting", "La Percezione Espressiva della Natura e dell'Arte"). Ha escrito sobre filosofía de la escultura ("La escultura. Un arte del espacio", "Teatralidad y la Escultura como un arte") y es autora del libro *El arte minimal. Objeto y sentido*. Ha escrito sobre filosofía del arte contemporáneo: "Significado y acción. Notas sobre Arte Conceptual en España", "Memoria, arte contemporáneo e identidad: Ilya Kabákov", "Experiencia y teoría estética en el arte conceptual", "Estetización y autonomía estratégica", "Institución-arte e intencionalidad artística", "El artista como reportero. Los mitos del fotoperiodismo artístico" y "La concepción de la imagen en el arte contemporáneo". También ha publicado sobre pintoras e historia feminista del arte: "Naturaleza y sujeto en Georgia O'Keefee", "Drama y espectador en Artemisia Gentileschi", "Estrategias conceptuales del arte feminista", y el libro *Artemisia Gentileschi*. Ha trabajado en la historia de la estética, editando a los teóricos del arte formalista Konrad Fiedler (*Escritos sobre arte*) y Adolf von Hildebrand (*El problema de la forma en la obra de arte*).

Ha trabajado sobre la relación del valor estético y otros valores en arte: "El valor moral del arte y la emoción", y "Sentimentality as an Ethical and Aesthetic Fault". En la actualidad su investigación se centra en el análisis de la experiencia estética, la percepción y el juicio y ha publicado: "Theatricality and Everyday Aesthetics", "Aesthetic Normativity and the Expressive Perception of Nature", y "The aesthetic value of the unnoticed".

Ha editado o coeditado los siguientes volúmenes: *En torno al arte. Estética. historia y crítica* (2023), *El valor del arte* (2017), *Estética después del fin del arte. Ensayos sobre A. Danto* (2005), *Expression in the Performing Arts* (2010), *Significado, emoción y valor. Ensayos sobre Filosofía de la Música* (2012) y *Estética* (2013).

Ha sido Investigadora Visitante en las universidades de Berkeley, Cambridge y Pompeu Fabra. Es miembro del comité editorial de *Estetika. European Journal of Aesthetics, Estetica. Studi e richerce, Daimon. Revista internacional de Filosofía, Revista de Filosofía* y *Enrahonar*. Edita la Serie de Filosofía (La balsa de La Medusa) en la editorial Antonio Machado.

La entrevista que viene a continuación permite conocer mejor a la persona, la académica y la filósofa que es Paca Pérez Carreño, y es un documento muy iluminador para conocer mejor nuestra historia reciente. Esperamos que la disfrutéis.

Distinción a una trayectoria: ENTREVISTA a FRANCISCA PÉREZ CARREÑO

(Distinción concedida por SWIP(A), Society for Women in Philosophy, http://swipa.ugr.es/)

¿Por qué decidió estudiar filosofía? ¿Qué recuerdos tiene de sus años de estudiante en la Universidad? ¿Qué profesores tuvieron una influencia mayor en su carrera posterior? ¿Tuvo algún referente femenino?

Como creo que es habitual, mis profesores de bachillerato influyeron decisivamente en la elección de carrera. Tuve dos buenos profesores de Filosofía, pero yo dudaba entre estudiar Clásicas o Filosofía. Finalmente me decidí a estudiar lo segundo en parte, precisamente, por consejo de mi profesor de Griego, José García Blanco. Estudié Filosofía en la Universidad Autónoma de Madrid. Recuerdo especialmente como magníficos docentes a Carlos Solís, de Historia de la Ciencia y a Ludolfo Paramio, de Sociología y Sociología del Conocimiento. Las clases de ambos eran de lo más divertidas, contagiaban entusiasmo y nos hicieron leer lo que entonces era novedoso: Hanson, Kuhn, Lakatos y Berger y Luckmann. Otros profesores importantes para mi desarrollo fueron José Hierro, que influyó en mi interés por la filosofía del lenguaje y Carlos Thiebaut, que me hizo leer a Habermas y conocer la Teoría Crítica y con quien después he trabado una amistad duradera. Finalmente decidí hacer la tesina y la tesis doctoral con Valeriano Bozal, en Estética, que juntaba mi interés por la filosofía del lenguaje y del arte.

En aquellos años este grupo de profesores entre otros daba un aire moderno al departamento, enseñando autores y temáticas contemporáneos. El ambiente entre los estudiantes era colaborativo y nada competitivo, leíamos mucho más que ahora y teníamos amistad fuera de las aulas. Los años en los que estudié eran los primeros de la transición (1979-85) y el optimismo general y el ambiente político contagiaban a todos. Éramos también bastante críticos con algunos profesores que no considerábamos a la altura y de los que pedíamos la expulsión. Creo recordar que conseguimos que alguno dimitiera.

Indudablemente la influencia de Valeriano Bozal durante toda mi vida académica y también personal es la mayor. Desde muy temprano demostró su confianza en mi capacidad. Primero apoyó mi entrada como Ayudante en el Departamento de Filosofía de la UAM, donde acababan de entrar entonces mis amigos del curso anterior, Huberto Marraud y Guillermo Solana. En aquellos años se aprobó la LRU (Ley de Reforma Universitaria) y terminamos rápidamente nuestras tesis para convertirnos en Ayudantes (LRU). Bozal creó también la Revista *La balsa de La Medusa,* de la que me hizo Secretaria de Redacción. Entonces todavía se recibían los artículos a máquina, en la editorial los „picaban" y nosotros los componíamos: cada página con su texto, sus notas, sus imágenes, a las que añadíamos los pies de foto. Eran horas de trabajo manual que recuerdo con nostalgia. Entonces adquirí mi afición por el trabajo editorial, al que he dedicado horas de corrección de traducciones y pruebas.

La balsa de La Medusa se acabó cuando las publicaciones empezaron a hacerse más académicas. Nos llegaban muchísimos más artículos, pero su carácter era diferente al de los quince o veinte primeros años. Claramente se empezaba a publicar para hacer carrera académica (yo la primera) y no por un interés más general en el análisis crítico de la cultura y la sociedad.

Por desgracia, en la carrera solo tuve una profesora, en primero, Pilar Castrillo, de Lógica. Era una extraordinaria profesora, pero perdí contacto con ella muy temprano. Desde luego en clase éramos bastantes mujeres y me acuerdo de mis amigas, con alguna de las cuales sigo teniendo contacto. Pero ahora percibo cómo el ambiente era totalmente masculino. Lo era incluso más que ahora.

¿Ha completado su formación en el extranjero? Si es así, ¿dónde? ¿qué recuerdos tiene de esas estancias? ¿qué aprendió en ellas?

Muy joven gané la plaza de Profesora Titular en la Universidad de Murcia y fue entonces cuando realicé una estancia de un año en la Universidad de California en Berkeley. Mi tutor allí era Richard Wollheim, que es junto a Valeriano Bozal, la otra gran influencia filosófica en mi vida. En Berkeley aquel año coincidían con Wollheim, Donald Davidson, Marcia Cavell, Bruce Vermazen y Barry Stroud. Yo solía ir a los seminarios de Wollheim, Davidson y Michael Baxandall, un gran historiador del arte. Me di cuenta de que el trabajo y la enseñanza filosóficos eran básicamente igual que en España. El seminario de Davidson era sobre Quine, incluía en sus clases una grabación

DISTINCIÓN A UNA TRAYECTORIA: FRANCISCA PÉREZ CARREÑO

de conversaciones con el propio Quine, y explicaba su propia superación del empirismo quineano. Aunque se trataba básicamente de seminarios, los estudiantes de doctorado sentían respeto si no temor a intervenir y hasta el final del curso Davidson se quejaba de que no le habían entregado aún los temas de sus ensayos. La experiencia me es familiar. Wollheim enseñaba de forma más magistral su teoría de la representación pictórica. Sus análisis de obras eran impresionantes. Igual sucedía con Baxandall que además introducía análisis psicológicos sobre la visión de las pinturas. La cultura de aquel grupo de académicos era impresionante. No solo la cultura filosófica, sino también artística. Recuerdo oír tocar a Davidson el piano con Bruce Vermazen al saxo. Y ver en casa del primero una de las *Pinturas ciegas* de Robert Morris con cita del propio Davidson.

El contacto con Wollheim y la colaboración con él hizo que tuviera una visión amplia de lo que era la filosofía analítica, ya que su trabajo está fuertemente influido por el psicoanálisis, en lo que coincidía con Marcia Cavell. Wollheim era muy generoso con su tiempo, leía mi artículo sobre metáfora visual y lo corregía cuidadosamente, me citaba a discutirlo en el despacho o en su casa o en un restaurante indio que le encantaba. Después le vi en España en tres ocasiones, primero vino a Murcia a la primera edición del Art, Mind, and Morality. Dio tres charlas y discutió, incansable, todas las ponencias. Después lo invitó Tomás Llorens para hablar sobre realismo en el Thyssen y sobre formalismo en un curso de verano en Santander. Murió muy poco después.

Antes del año en California, que además disfruté enormemente por los paisajes y los viajes que realizamos, había pasado medio año en la Universidad de Hull, en el Reino Unido. Aprovechando una prórroga del permiso de maternidad, acompañé en un intercambio de Erasmus a mi pareja que daba algunas clases allí. El departamento de Filosofía de la UMU tenía una estrecha relación de Erasmus con el de Hull. Ya nada más llegar a Murcia, Luis Valdés me envió para allá a supervisar la estancia de algunos estudiantes y dar una charla. Coincidí entonces con estudiantes como Ángel García, que luego sería colega aquí, y con Estela González-Arnal, que sigue siendo profesora allí. En Hull disfruté de la compañía y el trabajo en seminarios con filósofos y filósofas como Paul Gilbert, Kathleen Lenon o Gerry Wallace. A diferencia de los cursos a los que acudí en Berkeley, la enseñanza en Inglaterra era mucho más participativa para los alumnos. Escribí entonces (1993) un librito sobre Artemisia Gentileschi, cuando había poca historia del arte feminista en España, y que fue muy divulgado. También preparé el de John Constable, aprovechando la estancia, pero a mis amigos ingleses les parecía mucho menos interesante. De hecho, fue mucho menos leído.

Realicé posteriormente una estancia en Cambridge (RU). Ya con muchas obligaciones en España tuve que volver un par de veces y no fue tan productiva. Disfrute del ambiente de la ciudad y del Moral Sciences Club además de la amistad de Derek Matravers y Rob Hopkins.

¿Cómo recuerda sus primeros años de formación? ¿Ha tenido problemas para conciliar los distintos aspectos de su vida?

Fui ayudante al terminar la carrera en la UAM, pero en cinco años gané la plaza de Profesora Titular en la Universidad de Murcia. Era impaciente y cuando surgió la oportunidad me presenté y gané la plaza, aunque supusiera irme de Madrid. En aquellos años, ahora lo veo, trabajé muchísimo, en la elaboración de la Tesis sobre Semiótica y Estética de la Imagen, pero también en la edición de autores importantes del Formalismo alemán, como Fiedler y Hildebrand, y en artículos que desarrollaban mi tesis. Cuando me establecí como Profesora Titular en Murcia decidí tener a la que es mi única hija. Lo cierto es que no tuve problemas para conciliar la vida laboral y personal, en parte también porque mi marido era colega y porque vivíamos en una ciudad amable y mucho más fácil de dominar que Madrid. Como dije antes, realicé una estancia en Hull cuando mi hija era una bebé de cinco meses. La ayuda de amigos allí y la colaboración con mi pareja facilitó que participara en la vida del departamento. Como sucedió después en Berkeley.

Durante mi estancia en Cambridge dejé a mi hija con su padre en Murcia, pero me pareció algo natural hacerlo. Ahora veo aquellos años como productivos y felices y creo que realmente lo fueron, pero entonces sentía una verdadera obsesión por no dejarme aprisionar por la vida familiar y provinciana. En ocasiones pensaba que los cuidados no me iban a dejar volver a escribir o producir como antes. Quizá por eso también salí tantas veces de Murcia y quizá por eso no pude cumplir el deseo personal de tener más hijos. Quizá tampoco sea significativo que entre las seis profesoras que han pasado por el departamento de filosofía de Murcia durante estos años solo hayamos tenido tres hijos. Pero pienso también en muchas de mis conocidas filósofas o académicas en general que no los han tenido. No digo que la tarea académica sea la que en cada caso concreto lo haya impedido, ni que siempre se sienta como una pérdida, pero desde luego tiene que haber alguna relación entre la dedicación académica de las mujeres y el número de hijos.

En Murcia en pocos años me sentí responsable del desarrollo del Área de Estética y fui capaz de reunir en el área grandes colegas, amigos (Salva Rubio) y amigas (María José Alcaraz y Matilde Carrasco), con los que he logrado disfrutar de mi trabajo. Además, hasta la crisis de 2008 el crecimiento de la Universidad de Murcia me permitió disfrutar de medios materiales y de permisos que probablemente no hubiera disfrutado de seguir en Madrid. El último, en Barcelona, durante un par de años, que me permitió vivir cerca de mi hija ya doctoranda, de la ciudad y de mis amigas y amigos filósofos.

Aunque no en mis años de formación, últimamente sí he sentido que la labor de cuidados entorpece algo la vida tal como la vivía antes.

DISTINCIÓN A UNA TRAYECTORIA: FRANCISCA PÉREZ CARREÑO

¿Ha considerado alguna vez continuar su carrera fuera de España? ¿Por qué?

No, no lo he considerado nunca. La razón quizá sea que conseguí estabilidad laboral muy pronto. Además, mi grupo ARES-MUR trabaja bien y a un nivel internacional muy bueno. Siempre hemos tenido mucha colaboración con académicos europeos y americanos. Por último, desde finales de los noventa del siglo pasado el desarrollo en calidad de la Filosofía en España me parece innegable. Creo que he contribuido algo, aunque sea mínimamente, por ejemplo, en la fundación de la Sociedad Española de Filosofía Analítica y los Seminarios Interuniversitarios sobre Arte, Mente y Moralidad.

En esta carrera me he sentido acompañada por amigos y amigas en un periodo bastante bueno para la filosofía española. Además, mis compañeros y compañeras de generación han impulsado a los jóvenes filósofos y filósofas actuales y creo que podemos sentirnos orgullosos por ello.

¿Qué filósofos y filósofas considera Vd. que han marcado más claramente su desarrollo filosófico? ¿En qué puntos concretos de su pensamiento percibe Vd. la influencia de es@s filósof@s?

Como he dicho antes han sido dos filósofos varones, Valeriano Bozal y Richard Wollheim. Del primero he aprendido muchas cosas, pero principalmente y a través del análisis del gusto, la idea de que penetración cognitiva e inmediatez perceptual no están reñidas. Parece una idea sencilla, pero para mí, de joven, abrió un mundo. He aprendido muchas más cosas de él: por ejemplo, una determinada concepción de mímesis o representación, o la noción de "lucidez" como el valor máximo de la obra de arte. La capacidad del arte para hacer ver lo que está oculto. He admirado en él además otras virtudes ligadas al estudio, como el esfuerzo, la honestidad y la ambición académica.

De Wollheim sin duda el aprendizaje principal fue sobre el significado experiencial de las obras de arte, la idea de que la interpretación artística consiste en tener una experiencia adecuada de la obra. Esto permite entender el carácter cognitivo y afectivo del arte y su influencia en nuestra vida. También me influyó en aceptar cierto intencionalismo sobre la interpretación de la obra de arte, aunque no tan robusto como el suyo. Al principio su obra me sirvió para refinar mi idea sobre representación y expresión pictóricas, pero finalmente entendí también de su obra cómo ligaba el arte y la vida. Entre sus virtudes admiré de él su tolerancia, su actitud comprensiva hacia las debilidades humanas y su desconfianza de la perfección y el idealismo morales.

Ahora que lo pienso, el último texto de Bozal se tituló "La invención de lo humano" y tenía que ver también con la representación de lo oscuro y poco edificante y su valor contra la imagen sublime e idealizada de humanidad. La verdad es que el contacto con ambos, de los que recibí siempre críticas pertinentes y constructivas y apoyo personal, ha sido muy gratificante.

¿Cuál era su relación con las filósofas de su generación?

Entre mis compañeras en la UAM tenía un grupo de amigas con quienes no he tenido mucho contacto posterior. Solo continúo mi amistad con Adelina Sarrión Mora, que es una gran especialista española en Inquisición. Su especialidad está alejada de la mía, pero sus estudios sobre solicitación, beatas, etc. tiene una perspectiva clara de género. Otra amiga del grupo era Marina Garzón, Catedrática de Filosofía del Derecho en la Universidad de Castilla - La Mancha.

Fuera de la UAM, mi amistad con María José Frápolli se fraguó muy temprano, en los años noventa del siglo pasado en un Congreso sobre Lógica y Lenguajes Naturales, creo recordar. Otras filósofas de mi generación son Pepa Toribio, Marga Vázquez, Cristina Corredor, María Cerezo… Con ellas he ido coincidiendo en la SEFA y en otros contextos por afinidad generacional y filosófica. Siento por todas ellas admiración y cariño y recurro a sus escritos cuando necesito fundamentación en teoría del conocimiento o en filosofía del lenguaje.

Aunque son más jóvenes es mi deber mencionar también a mis compañeras de área, Mª José Alcaraz y Matilde Carrasco, con quienes la colaboración diaria es fácil, enriquecedora y divertida. María Cerezo completaba hace unos años el minúsculo grupo de mujeres en el Departamento de Filosofía de la UMU. También, y aunque ha desarrollado su carrera fuera, fue mi estudiante y es buena amiga Esa Díaz-León, cuyo trabajo en filosofía de la mente y del lenguaje feministas es para mí iluminador. Otra vez somos solo cuatro mujeres en el departamento de Filosofía. Éramos y seguimos siendo pocas, ¡pero muy buenas!

¿Cómo ve Vd. el papel de las filósofas en la universidad española?

He nombrado solo unas cuantas filósofas a las que me une la amistad y la pertenencia a la SEFA, es decir, también cierta afinidad filosófica, pero creo que filósofas de mi generación han jugado, y juegan todavía, un papel muy importante en la filosofía española. En particular, y no sorprendentemente, en los estudios feministas y de género. Sin filósofas que nos antecedieron como Celia Amorós, Amelia Valcárcel, Victoria Camps, Eulalia Pérez Sedeño, Alicia Puleo, y muchas otras, no habría filosofía feminista en España. Hoy en día las jóvenes filósofas juegan un papel fundamental en la revisión de la tradición filosófica, pero también en los estudios sobre identidad sexual y de género, el análisis de los prejuicios y estereotipos, de la filosofía del lenguaje y de la mente feminista, etc. Es decir, en temas directamente relacionados con el género. En algún momento u otro, muchas de nosotras nos vemos aludidas y responsables de participar en el estudio de estos temas. Esto mejora la filosofía en general. Además, veo estupendas filósofas en mi área y en la filosofía española en general.

También creo que en comparación con el pequeño número que somos en el total de filósofos en niveles universitarios, las mujeres ocupamos muchos puestos de gestión, en especial en

DISTINCIÓN A UNA TRAYECTORIA: FRANCISCA PÉREZ CARREÑO

asociaciones filosóficas, en comités de revistas y en gestión universitaria. Eso me hace pensar en que nos ocupamos más de tareas colaborativas y nos implicamos más en la creación de comunidades de trabajo que nuestros compañeros. Naturalmente esto es una generalización, pero creo que no muy atrevida.

Sobre la visibilidad de la investigación hecha por mujeres quizá todavía haya que insistir en que debemos apoyarnos citándonos, leyéndonos, discutiendo entre nosotras. Mi labor como jurado en las primeras ediciones del Premio SWIP-Analytic España Premio Lex Academic de Ensayo me dio la oportunidad de conocer el trabajo riguroso, actualizado, interesantísimo de jóvenes filósofas ocupadas en la interpretación y la explicación filosófica de problemas de actualidad.

De acuerdo con los datos, la presencia de mujeres en filosofía en todo el mundo es similar a la de las ingenierías, el número de filósofas en todos los niveles está muy por debajo del 50%, ¿cuál cree que es la razón? ¿cómo podría mejorarse la situación?

No sé cuál es la razón específica más allá de la obvia de que los sesgos machistas perviven en todos los ámbitos. Incluso y aún más en aquellos como la filosofía en los que se presume de rigor analítico y de pensamiento crítico. También me parece innegable que la filosofía pertenece al género de estudios de estereotipo masculino. Quizá tenga que ver con las imágenes de filósofos con gafas de pasta fumando en pipa, también con su atracción por problemas abstractos y alejados de la vida cotidiana o, por el contrario, con su aspiración a rey - filósofo.

Recuerdo un par de artículos de Jennifer Saul y Sally Haslanger, en los que se mencionan cuestiones psicológicas más concretas, sobre la persistencia de los estereotipos machistas en Filosofía. Ciertamente, los filósofos y filósofas solemos ser muy *judgmental,* muy amigos de juzgar a las personas a la primera con predicados como "muy inteligente", "agudo", "bueno dialécticamente" o sus contrarios. Esa tendencia es perjudicial cuando se trata de juzgar a nuestras estudiantes. Los estudios muestran que la evaluación del CV o el trabajo de las mujeres tiende a ser peor que el de los colegas masculinos similares. Ya desde en el aula universitaria la intervención de los varones es mucho más frecuente y merece una mayor atención y valoración que la de sus compañeras. No solo es que existan sesgos de género también entre las profesoras, sino que las propias estudiantes los tienen. Por un lado, existe el peligro del sesgo implícito en varones y mujeres, profesores y estudiantes, pero además de los estereotipos de género, por otro lado, las mujeres sentimos también la amenaza del estereotipo (estereotype threat). En cierta medida, tendemos a interiorizar los sesgos, y aunque no lo hagamos, el mero conocimiento de su existencia nos produce una ansiedad que nos puede hacer actuar y expresarnos peor en público.

Naturalmente los problemas del sesgo implícito y la amenaza del estereotipo se producen también fuera del aula, en contextos como las publicaciones científicas o los encuentros científicos.

Especialmente las profesoras tendríamos que ser más sensibles a la invisibilización de nuestras alumnas y favorecer su expresión y desarrollo público. Aunque no estamos a salvo del problema de cómo ser justas en la evaluación de las mujeres una vez conocido el sesgo (es decir, corriendo el peligro de sobrevalorarlas y corriendo el peligro de infravalorarlas para no sobrevalorarlas). Quizá deberíamos anonimizar las pruebas escritas, aunque la valorización de la actuación pública de las estudiantes todavía estaría sujeta a los sesgos.

De todos modos, y aunque sé que esto es muy discutible, creo que la situación es tan grave que debería estudiarse la posibilidad de ofrecer plazas solo para mujeres en aquellos departamentos en los que su presencia fuera menor del 25%, por ejemplo.

¿Cómo ve la situación de la filosofía en España? ¿Cree que es comparable con la situación de los países de nuestro entorno?

Sí, sinceramente, creo que el nivel medio es comparable. Sin embargo, los grandes nombres de la filosofía contemporánea (al menos analítica) suelen seguir siendo anglosajones. Que la mayoría escriba en su lengua materna y que la comunidad filosófica sea anglosajona favorece este hecho. Es muy difícil entrar en esa conversación sin pertenecer a esa academia. Ahora bien, la comunicación de los filósofos españoles con el exterior eleva la calidad de nuestra investigación.

¿Cuál diría Vd. que es su aportación a la filosofía contemporánea y a la universidad española?

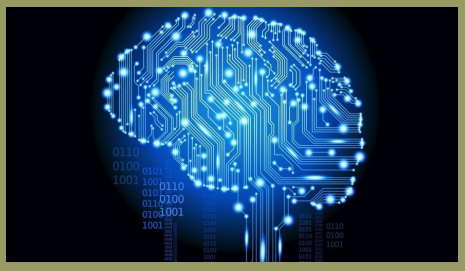
No sé exactamente cuál es mi aportación teórica, mi estudio sobre iconismo todavía se lee, así como mi análisis del Arte Minimal desde un punto de vista filosófico. También mi estudio de la expresión artística. Es decir, en general, la utilización de la filosofía y la estética filosófica para el análisis de fenómenos artísticos.

Institucionalmente he contribuido a la profesionalización académica de la Estética en España: primero con la creación del grupo Aresmur en la UMU, allí con la tutorización de doctorandos y post-doc.; colaborando con otros grupos trabajando en Estética y en Filosofía Analítica; fui una de los miembros fundadores de la Sociedad Española de Filosofía Analítica; también participé en la fundación de la Sociedad Española de Estética y Teoría de las Artes; por último, me encuentro orgullosa de mi colaboración en la fundación de la Sociedad Europea de Estética (ESA en sus siglas en inglés), de la que he sido presidenta durante siete años. Es decir, también he contribuido a la internacionalización de la Estética española.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

ÍNDICE

www.solofici.org



SLMFCE
Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

¿Es la idea de superinteligencia artificial realmente pensable?

Marcos Alonso
Universidad Complutense de Madrid
Facultad de Medicina
Departamento de Salud Pública y Materno-Infantil
marcos.alonso@ucm.es

Resumen: Los recientes avances en inteligencia artificial (IA), especialmente los relacionados con los llamados *Large Language Models* (LLM), han reavivado las preocupaciones éticas sobre el avance acelerado de estas tecnologías. En concreto, las discusiones sobre el advenimiento de la superinteligencia artificial, y lo que esto supondría para la humanidad, ha llegado a ser una preocupación de gran importancia. En este artículo se intenta abordar un problema previo: no tanto si la superinteligencia artificial es algo posible o próximo en el tiempo, sino si la mera idea de «superinteligencia» es algo realmente pensable. Para ello primero se exponen algunas ideas básicas del pensamiento transhumanista, corriente filosófica en la que suelen encuadrarse las discusiones sobre la superinteligencia artificial. Tras esto, se analiza más detalladamente la propuesta concreta de N. Bostrom, quien hasta la fecha ha abordado el problema de la superinteligencia artificial con mayor precisión y detalle. Por último, se desarrolla una crítica a este mismo concepto, desde el punto de vista de una comprensión biológica de la inteligencia. Se concluye que esta labor de aclaración conceptual respecto de la idea de superinteligencia artificial es crucial para abordar los problemas bioéticos asociados a este tema.
Palabras clave: Superinteligencia, transhumanismo, inteligencia, bioética, ética de la IA.

1. Introducción

Hace casi un siglo el filósofo español José Ortega y Gasset dejó dicho que "uno de los temas que en los próximos años se va a debatir con mayor brío es el del sentido, ventajas, daños y límites de la técnica" (Ortega, 2006, 553). Estas proféticas palabras no han hecho más que cumplirse, hasta el punto de que una proporción enorme de las investigaciones éticas actuales tienen como tema central el de las implicaciones sociales, éticas y políticas de las nuevas tecnologías. En este sentido, la preocupación por las máquinas inteligentes no es algo completamente nuevo. Desde la segunda mitad del siglo pasado se ha convertido en una parte central de nuestro imaginario colectivo, con numerosas novelas y películas dedicadas al tema. La posibilidad de un gobierno mundial automatizado (recordando a clásicos como *Un mundo feliz*, de Aldous Huxley o *1984*, de George Orwell); la plausible catástrofe para la humanidad derivada del malfuncionamiento de las máquinas (un tema recurrente en el cine, como atestiguan *2001 Odisea en el espacio*, de Stanley Kubrick, *Terminator*, de James Cameron o *Matrix*, de Lilly y Lana Wachowski); o las posibilidades y limitaciones de las inteligencias artificiales con valores humanos (reflejadas en obras como *Yo, robot*, de Isaac Asimov o en la reciente *Her*, de Spike Jonze), son solo algunos de los temas que han ido apareciendo de manera cada vez más recurrente en las producciones culturales de las últimas décadas. Podría decirse que las creaciones artísticas anticipan el mundo que está por venir, si bien en el caso de la inteligencia artificial (IA) los avances científico-técnicos suceden de una manera tan vertiginosa que la imaginación humana tiene dificultades para mantener el ritmo.

Ante esta situación, abordar de manera seria y rigurosa problemas que hasta hace no mucho parecían meras ocurrencias de ciencia ficción se convierte en una empresa especialmente difícil. La búsqueda de cierta mesura ante las siempre inciertas predicciones tecnológicas colisiona con una realidad tecnológica que ya en el presente ha roto con muchas de las predicciones más optimistas[1]. Particularmente, los últimos desarrollos en inteligencia artificial, especialmente los denominados *Large Language Models* (LLM) como ChatGPT (Newport, 2023), han generado renovadas inquietudes éticas sobre el rápido progreso de estas tecnologías. En este contexto han reaparecido discusiones acerca del surgimiento de la superinteligencia artificial y sus implicaciones para la humanidad. Este artículo se propone abordar un problema que en cierto sentido precede a estas discusiones: no tanto si la superinteligencia artificial es algo factible o inminente, sino si la mera noción de «superinteligencia» artificial es algo realmente concebible. Para ello, empezaré presentando algunas ideas fundamentales del pensamiento transhumanista, la corriente filosófica en la que generalmente se enmarcan las conversaciones sobre la superinteligencia artificial. Posteriormente, examinaré de manera más detallada la propuesta específica de N. Bostrom, quien hasta ahora ha abordado el problema de la superinteligencia artificial con mayor precisión y detalle. Finalmente expondré una crítica a este mismo concepto desde una comprensión biológica de la inteligencia, realzando la importancia de esta labor de aclaración conceptual.

2. La superinteligencia artificial en el contexto del proyecto transhumanista

Antes de entrar de lleno a discutir el concepto de superinteligencia artificial conviene comprender mínimamente el proyecto transhumanista al que habitualmente ha estado ligada esta noción de superinteligencia. En esencia, el transhumanismo es una corriente filosófica y cultural que aboga por superar las limitaciones inherentes a la condición humana, tanto en aspectos físicos como mentales, mediante el avance científico y la implementación de innovaciones tecnológicas (Diéguez, 2017). Se presenta en ocasiones como una continuación del humanismo renacentista, centrado en el florecimiento humano. Los defensores del transhumanismo argumentan que la humanidad

1. Uno de los avances más notables ha sido el de *AlphaFold* en 2020, un sistema de IA desarrollado por *DeepMind* que predice la estructura 3D de una proteína a partir de su secuencia de aminoácidos. Existía un consenso muy amplio de que este descubrimiento no llegaría antes de 2030 como pronto (Read et al., 2023).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

debe perseguir una existencia más plena y perfecta, utilizando las nuevas tecnologías de manera casi ilimitada, punto este último que constituye la principal diferencia respecto de movimientos similares en el pasado que también se centraron en el mejoramiento o perfeccionamiento de la humanidad.

En este sentido, una de las características más prominentes del transhumanismo es su menosprecio o consideración secundaria hacia el cuerpo y la biología propiamente humanos. Por ejemplo, autores como Sandberg abogan por la "libertad morfológica" (2001), una propuesta que implicaría la capacidad de elegir la configuración física que mejor se adapte a nuestras preferencias. Según el transhumanismo, la ciencia y la tecnología nos ofrecen la oportunidad de seleccionar nuestra forma corporal, e incluso la posibilidad de prescindir de ella. Algunos transhumanistas ven como posible y deseable una existencia sin limitaciones biológicas (Kurzweil, 2005; Chalmers 2010). Desde estos puntos de vista aparece la posibilidad de trascender la especie humana hacia lo que estos transhumanistas denominan como el "trans-humano" o "post-humano".

El caso que se va a analizar en este artículo, el caso de la superinteligencia, es, en gran medida, una idea derivada de este rechazo transhumanista de la corporalidad. Pese a sus distintas formulaciones (Kurzweil, 2005; Bostrom, 2016; Yampolskiy y Duettmann, 2020), el concepto de superinteligencia suele ir vinculado a una ruptura respecto de la base biológica de la inteligencia. Ya sea mediante el potenciamiento farmacológico o biomecánico del órgano que es el cerebro -que, mediante este potenciamiento, deja de ser entendido en sentido estricto como órgano de un cuerpo-; o, como sucede en la mayoría de ocasiones, mediante un cambio de base radical para la inteligencia, dejando de lado la base biológica en favor de una supuesta base computacional.

Considero que el autor que mejor ha presentado esta posibilidad de una inteligencia artificial es N. Bostrom con su obra *Superinteligencia* (2016). Pienso que esta importante obra es la mejor muestra de lo que supone llevar las premisas transhumanistas hasta sus últimas consecuencias, y es por ello por lo que el resto de artículo discutirá de manera preponderante con las reflexiones de este autor. La idea que conecta directamente la propuesta de superinteligencia con el proyecto transhumanista es precisamente la comprensión de la superinteligencia como el mero incremento mecánico de lo que se entiende como las bases materiales de la inteligencia. De este modo, se entiende que si un cerebro puede establecer un determinado número de conexiones neuronales por segundo, una tecnología que permita recrear computacionalmente un cerebro capaz de realizar el doble de conexiones neuronales, dará como resultado una superinteligencia que sobrepasará ampliamente los logros de la inteligencia humana.

Esta idea ha recibido diversas críticas. Por ejemplo, Larson piensa que ninguna suma de inteligencias restringidas tiene por qué dar como resultado, necesariamente, una inteligencia artificial general (AGI – *Artificial General Intelligence*). Según este autor, algo así supondría un salto cualitativo sobre el que todavía no sabemos casi nada (Larson, 2023, 1). El problema que quiero destacar en este artículo es algo distinto. Se trata de que la superinteligencia o el supercerebro pensado en términos de potenciación ciega no solo podría no conducir necesariamente a una inteligencia superior o general, sino que probablemente no cumpliría con la definición misma de inteligencia en el sentido propio de la palabra. Esto será el núcleo de la crítica que se desarrollará en el cuarto apartado, pero antes considero necesario exponer la propuesta de Bostrom más pormenorizadamente.

3. Concepción de la superinteligencia artificial, formas de advenimiento y estrategias de control

Para comprender mejor la crítica al concepto de superinteligencia artificial que quiero llevar a cabo, conviene exponer con cierto detalle las ideas de Bostrom en *Superinteligencia*, de modo que pueda captarse mejor la naturaleza de su razonamiento y las dificultades presentes en este planteamiento. En los últimos años también se ha hablado de inteligencia artificial fuerte o inteligencia artificial general (Heaven, 2020), conceptos igualmente relacionados con el proyectado avance de las nuevas tecnologías computacionales. Prefiero centrarme en el concepto de superinteligencia y la obra de Bostrom por ser un elemento de confrontación más definido y en donde puede verse con más claridad el resultado de las ideas transhumanistas sobre la inteligencia artificial y sus futuros desarrollos.

En este sentido, un primer punto que es necesario remarcar es que Bostrom, a diferencia de otros autores habitualmente encuadrados en el transhumanismo (Kurzweil, 2005), no juega a ser adivino y no aspira a erigirse como un nuevo gurú tecnológico. No vemos aquí, pese a lo que podría creerse, un caso del mito de la IA denunciado por Larson, el mito de que la llegada de la superinteligencia artificial "es inevitable y sólo cuestión de tiempo" (Larson, 2023, 1). Bostrom presenta sus reflexiones desde una moderación e incluso escepticismo notable (Cf. Bostrom, 2016, 1-21); y es por ello que *Superinteligencia* es una referencia adecuada para reflexionar seriamente sobre estos problemas. Sin embargo, como el propio Bostrom argumenta una y otra vez a lo largo del texto, los problemas en torno a la posibilidad de una superinteligencia artificial son tan decisivos y su impacto puede ser tan grande, que no podemos ignorarlos pese a lo improbables o esotéricos que nos resulten. Como digo, Bostrom no se muestra categórico sobre la llegada de la IA ni sobre sus consecuencias, y de hecho nunca cierra la puerta a la posibilidad de que los avances en IA y tecnologías afines acaben siendo mucho menos amplios de lo esperado. Sin embargo, como este autor argumenta convincentemente, la probabilidad de que estos avances se den, y que se den de manera acelerada, es extremadamente alta (Bostrom, 2016, 18); algo que los recientes avances en IA con los LLM parecerían respaldar (Newport, 2023).

El libro *Superinteligencia* de Bostrom lleva por subtítulo *Caminos, peligros, estrategias*; pero entre estos términos acaba emergiendo uno por encima de los demás: el peligro. Necesitamos comprender los caminos que llevan a la superinteligencia para plantear las mejores estrategias frente a ella; pero la cuestión

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

central en torno a la que giran todas las demás es el peligro que la superinteligencia trae consigo, la amenaza que proyecta sobre la humanidad y que deberíamos tener muy en cuenta. Este tono de advertencia no desemboca, en el caso de Bostrom, en un miedo paralizante que demonice todo avance en IA; únicamente nos hace comprender la enorme responsabilidad que tenemos y la imprudencia que sería desentenderse de estos problemas o confiar ingenuamente en su desarrollo benéfico.

Superinteligencia se divide básicamente en dos bloques: un primer grupo de capítulos dedicado a presentar el problema de la inteligencia artificial, atendiendo a su historia, sus posibilidades presentes y su previsible desarrollo en los próximos años; y un segundo bloque algo más extenso dedicado a pensar qué podremos hacer con la superinteligencia, los peligros y problemas que probablemente acarree, y las estrategias y soluciones a nuestro alcance. El primer bloque se extiende aproximadamente hasta el capítulo 6. Tras un breve recorrido en el primer capítulo a la historia de la IA y una somera exposición del estado de la cuestión, Bostrom pasa, ya en el segundo capítulo, a exponer los posibles caminos que podrían llevar hasta la superinteligencia. En este capítulo se introducen algunos términos muy repetidos a lo largo del libro, principalmente los de IA (inteligencia artificial) y ECC (emulación de cerebro completo) (Bostrom, 2016, 22). Sobre este punto conviene advertir del uso que Bostrom hace de inteligencia artificial en un sentido amplio para referirse a cualquier forma no humana de inteligencia, y un uso más restringido de inteligencia artificial para referirse a los tipos de IA cuyo origen no está directamente relacionado con el cerebro humano, frente a los que sí tienen dicho órgano como base (como la emulación de cerebro completo).

En el capítulo 3 Bostrom considera brevemente los distintos tipos de superinteligencia que podrían surgir, algo que retomará en más profundidad en el capítulo 6 cuando hable de los diferentes superpoderes cognitivos que una superinteligencia presumiblemente tendría a su disposición. Bostrom introduce aquí por primera el decisivo concepto de Unidad, un ente único que concentraría todo el poder (político, tecnológico, económico, etc.) en sí mismo. Esta idea estará presente a lo largo de toda la obra y aparecerá repetidamente en numerosos párrafos. Los capítulos 4 y 5 intentan atisbar la manera concreta en que la superinteligencia surgirá. Bostrom insiste en la relevancia del momento en que tenga lugar la explosión de inteligencia y en la importancia de su velocidad de despegue: cuanto antes suceda menos preparados estaremos. La clave de estas discusiones y lo que más interesa al autor es la posibilidad de que la superinteligencia adquiera una ventaja estratégica decisiva, otro concepto que sobrevolará toda la obra.

Si bien *Superinteligencia* es un libro unitario y bien trabado, el capítulo 7 marca la entrada en lo que podríamos considerar la segunda parte del libro. Este capítulo aborda por primera vez el problema de las relaciones entre inteligencia y motivación, un marco conceptual que servirá de base al resto de la obra. Un primer gran peligro según Bostrom sería el de antropomorfizar la IA, asumiendo que cuanto más inteligente sea la IA, más humana será y sus motivaciones y objetivos más se parecerán a los humanos. En opinión de Bostrom esto podría no ser así, y, si no nos esforzamos específicamente por conseguirlo, el autor piensa que lo más probable es que la superinteligencia que surja sea profundamente inhumana. Como veremos un poco después, en este punto es donde a mi modo de ver mejor se ven las costuras de su planteamiento, pues una inteligencia pensada al margen de la lógica vital no es que sea inhumana, es que es ininteligible.

El octavo capítulo se pregunta, de manera directa y sin rodeos, si estamos abocados al desastre. Este tono de preocupación gobernará la segunda parte del libro, en la que Bostrom tratará de hacerse cargo de las posibles amenazas de la superinteligencia, proponiendo, a su vez, las que considera mejores estrategias para afrontarlas, principalmente la prevención. En este capítulo 8 Bostrom presenta diversos modos concretos en que la IA podría fallar, como la suplantación perversa (una IA que interpretara defectuosamente nuestras órdenes, suplantándolas por otras con efectos perjudiciales) o el crimen mental (la posibilidad de que las IAs o las emulaciones tengan un estatus moral que podría ser violado).

Los capítulos 9 y 10 están dedicados a examinar una primera salida a los problemas de la superinteligencia: los métodos de control (Cf. Bostrom, 2016, 127-137). Éstos se dividen, por un lado, en métodos de control de la capacidad, que buscan impedir que la superinteligencia tenga un poder efectivo total sobre el mundo; y, por otro lado, en métodos de selección de la motivación, que buscan elegir los objetivos que la superinteligencia llegaría a tener. En opinión de Bostrom ambos métodos tiene carencias. En el caso del control de capacidad el problema está en la tensión existente entre minar las capacidades de la IA lo suficiente para que no tenga un poder absoluto, sin restarle tanta capacidad que ya no pueda ser considerada superinteligente. En el capítulo 10 Bostrom comparará diversos tipos de IA en función de su idoneidad para el control. La selección de la motivación es un problema más complejo, que por ello se trata más detalladamente en los últimos capítulos, donde encontramos la verdadera clave de la propuesta de Bostrom -y de sus mayores aporías y problemas sin resolver.

Las reflexiones sobre los métodos de selección de la motivación conducen a Bostrom, en los capítulos 12 y 13, a reflexionar a fondo sobre la posibilidad de crear una IA con valores. Para Bostrom, ésta es la única salida que, en caso de ser posible, supondría una verdadera solución al problema de la IA superinteligente. Por eso el filósofo sueco concentra sus mejores esfuerzos en esta parte de la obra, que entiende como decisiva. En este sentido, merece especial atención la propuesta de la VCE (voluntad coherente extrapolada), una forma de normatividad indirecta que busca aprovechar la propia capacidad de la superinteligencia para llevar a cumplimiento de manera más perfecta los valores humanos comúnmente aceptados (Bostrom, 2016, 211-217) esta propuesta, que Bostrom toma principalmente de Yudkowsky, no se centraría en dar reglas fijas a la IA, sino en permitirle a ella misma extrapolar, a partir de los valores humanos, los fines que estos seres humanos, si fueran tan inteligentes como ella, le darían.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

El último tema que aborda Bostrom es el del escenario estratégico que probablemente tenga lugar una vez que la superinteligencia se haga efectiva. Esto es estudiado en los capítulos 11, 14 y 15 donde se analizan las condiciones y presumibles consecuencias sociales, políticas y económicas de la aparición de la superinteligencia. Es quizás la reflexión más acuciante e importante desde el punto de vista práctico, si bien no es el tema principal que quiero afrontar en este artículo.

Como puede comprobarse, Bostrom trabaja con una comprensión de lo que es la inteligencia implícitamente e incluso explícitamente ajena a la biología. Este punto de partida es lo que, a mi modo de ver, lastra la obra y hacer surgir la infinitud de problemas que el autor intenta remediar. Fundamentalmente, lo que va revelándose como el problema fundamental: el problema de la motivación de la superinteligencia artificial. Este problema de la motivación, plasmado a lo largo de la obra principalmente en el extenso análisis de los métodos de selección de la motivación y en la propuesta de la VCE (voluntad coherente extrapolada), es, a mi modo de ver, una señal del error de partida de intentar trasponer un concepto eminentemente biológico como el de inteligencia a un plano inorgánico. La VCE y propuestas similares no son más que juegos de prestidigitación que no llevan a nada. A continuación expondré esta crítica con algo más de detalle, solo después de aportar algunas ideas básicas sobre por qué entiendo que la inteligencia es un concepto necesariamente biológico.

4. Problemas del concepto de superinteligencia desde una comprensión biológica de la inteligencia

4.1. La inteligencia como atributo biológico

Como exponía en el último apartado, las aporías y dificultades que encuentra Bostrom, y con él todos los transhumanistas que han intentado plantear el problema de la superinteligencia artificial, tienen que ver con esa negación de la biología consustancial al proyecto transhumanista. Como explica Crawford en su importante monografía *Atlas of AI* (2021), la inteligencia artificial se ha pensado desde dos presupuestos profundamente equivocados. El primero, el convencimiento de que "con la formación o los recursos suficientes, se puede crear una inteligencia similar a la humana a partir de cero, sin tener en cuenta las formas fundamentales en que los seres humanos se encarnan, se relacionan y se establecen dentro de ecologías más amplias" (Crawford, 4-5, 2021). El segundo presupuesto, relacionado con el primero, es que la inteligencia es algo que existe de forma independiente, como entidad abstracta y etérea (Crawford, 5, 2021).

El problema aquí no tiene que ver con un apego irracional a nuestra condición biológica, no se trata de que nuestras inclinaciones nos lleven a aferrarnos a nuestra corporalidad y a nuestra biología. El problema tiene que ver con que extirpar la inteligencia de su base biológica supone su vaciamiento como concepto. La inteligencia de un ser no biológico es algo así la gratitud de una pared o la perspicacia de una rama. No significa nada más que en un sentido metafórico y derivado. De manera intuitiva, como hacemos en tantos otros casos, utilizamos aquí el término inteligencia en sentido metafórico o translaticio. Sin embargo, el peligro de no reconocer esta condición metafórica del concepto de inteligencia cuando se aplica a ámbitos no biológicos es muy significativa, llevándonos a confusiones con importantes consecuencias prácticas.

Algunos autores han criticado el uso de superinteligencia, e incluso de inteligencia artificial, sobre la base de que la inteligencia humana es mucho más compleja y puede hacer mucho más que la artificial (Morozov, 2023), particularmente, "pensar de forma independiente, comprender los matices del comportamiento humano y tomar decisiones basadas en el sentido común" (Dekel, 2023). La crítica que aquí voy a desplegar no se adhiere a estas posturas. Estas críticas dependen del estado actual de desarrollo de las tecnologías, que con toda seguridad irá cambiando, así como de una definición muy concreta, y a la vez inevitablemente muy problemática, de conceptos como «pensamiento independiente» o «sentido común». Aquí no ensayo esta vía, sino una sustancialmente diferente. El punto fundamental de mi critica tiene que ver con que la idea de que la inteligencia solo puede ser pensada con propiedad en términos biológicos, como "un fenómeno primero y ante todo biológico" (Ziemke, 2016, 9). Incluso, siendo más concretos, que la inteligencia sólo puede atribuirse a seres vivos conductuales que, en tanto que tales, tienen un cuerpo que necesitan mover para obtener ciertos fines. En la medida en que la superinteligencia se piensa desligada de estas coordenadas, no es realmente inteligencia y llamarla así sólo contribuye a confundirnos y desorientarnos.

La definición más citada y repetida de inteligencia es la siguiente de Gottfredson:

> La inteligencia es una capacidad mental muy general que, entre otras cosas, implica la capacidad de razonar, planificar, resolver problemas, pensar en abstracto, comprender ideas complejas, aprender con rapidez y aprender de la experiencia. No se trata simplemente de aprender de los libros, de una habilidad académica restringida o de saber hacer exámenes. Más bien refleja una capacidad más amplia y profunda para comprender lo que nos rodea: "captar", "dar sentido" a las cosas o "averiguar" qué hacer. (Gottfredson, 1997, 13)

Si bien esta definición no es demasiado técnica y es el reflejo de ciertas nociones comúnmente aceptadas sobre lo que es la inteligencia, es una definición que, como expone Warne (2020, 2) da en la diana respecto de varios aspectos clave. De cara a nuestra argumentación, el punto central es que la inteligencia es una capacidad de los organismos que los permite mediar conductualmente con su entorno. Sólo bajo esta lógica de la mediación conductual puede cobrar sentido la idea de Gottfredson de "comprender lo que nos rodea" o "averiguar qué hacer". La inteligencia es ampliamente considerada una adaptación evolutiva que nos permite adaptarnos y movernos exitosamente en el entorno, favoreciendo la supervivencia y la reproducción (Cf. Ritchie, 2015, 60). El teórico animalista Olson

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

insiste en este punto cuando expresa, en términos bastante afines a los que ya utilizó Ortega al hablar del yo y la circunstancia, que la vida consiste en ese enfronte entre un organismo y su entorno, añadiendo de manera muy precisa que "esto no pretende ser una afirmación empírica de la biología, sino parte del concepto de vida" (Olson, 1997, 138). Como a propósito de esta idea explica Schechtman, muchos atributos vitales, entre los cuales sin duda podemos incluir la inteligencia, "deben entenderse como una interfaz con un entorno y no como procesos totalmente internos" (Schechtman, 2014, 191).

Sin embargo, esta referencia a la inteligencia como adaptación evolutiva, función vital o mediación conductual es lo que nunca encontramos en las reflexiones sobre la superinteligencia. Más bien, y de manera explícita, aparece una negación de la base biológica de la inteligencia. En la obra de Bostrom esto toma la forma de una advertencia contra la antropomorfización de la IA. Según este autor debemos tener en cuenta que "una inteligencia artificial podría ser mucho menos humanoide en sus motivaciones que un alienígena verde de piel escamosa del espacio" (2016, 106). Lo que Bostrom dice aquí es, en cierto sentido, obvio. Un ser vivo diferente a nosotros muy probablemente tenga motivaciones diferentes a nosotros. Sin embargo, la referencia al "alienígena verde de piel escamosa" encubre y a la vez exhibe el hecho de que ese alienígena precisamente se está pensando como un ser vivo, con piel coloreada y habitante de un determinado ecosistema, aunque sea otro planeta. El problema es que la superinteligencia artificial no se piensa como un ser vivo, y, por tanto, atribuirle inteligencia resulta imposible, si nos detenemos por un momento para pensar en ello. El problema no es que la superinteligencia postulada por Bostrom y los transhumanistas sea poco antropomórfica, ni que tenga motivaciones distintas a las humanas. El problema es que, desde el punto de partida fisicalista-abstracto computacional de Bostrom, no podemos pensar esa pretendida superinteligencia artificial como inteligencia ni podemos realmente atribuirle motivaciones. La superinteligencia transhumanista es simplemente superininteligible.

Esta crítica que aquí presento comparte algunos aspectos con algunos de los presupuestos de lo que en los últimos años se ha conocido como las teorías de la cognición encarnada (*embodied cognition*). La noción de cognición encarnada no es un concepto unitario, sino bastante heterogéneo dependiendo de los diferentes autores que lo han abordado, con trabajos como *The Embodied Mind* de Varela, Thompson y Rosch (1991); *Embodied Cognition*, de Shapiro (2010); o el más reciente *Embodied Social Cognition* de Lindblom (2015). En todo caso, la idea fundamental compartida por todas estas propuestas es que la cognición solo puede entenderse desde la interacción sensoriomotora de un cuerpo con el entorno (Ziemke, 2016, 5). Sin embargo, la mayoría de estas teorías sobre la cognición encarnada exhiben una limitación en su planteamiento coincidente con la idea de superinteligencia de Bostrom. Pues si bien reconocer la encarnación de la cognición y asimismo de la inteligencia es importante, lo verdaderamente decisivo es entender la lógica vital subyacente, no el hecho crudo de esta encarnación o corporalidad.

El cuerpo de los seres vivos conductuales-inteligentes es siempre necesariamente un cuerpo-en-un-contexto, es decir, está siempre en un entorno. Un entorno lo es precisamente en relación al cuerpo del ser vivo: ambos, ser vivo y entorno, están en una mutua dependencia lógica. Y la clave de toda esta lógica vital, de la misma diferenciación organismo-entorno, así como de la condición conductual o sensomotriz de los organismos, es que ese organismo necesita alcanzar, mediante sus acciones en el entorno, ciertos fines. Principalmente, necesita moverse para nutrirse y sobrevivir. Y es sólo por esta necesidad orgánica que unas cosas se presentan como valiosas, un valor que derivadamente también se dice de los comportamientos que favorecen la consecución de esas cosas valiosas. Fuera de esta lógica vital, hablar de valores o de inteligencia carece de sentido. Propuestas recientes como la de Ziemke (2016) han apuntado a una posible solución a esta limitación de las teorías de la cognición encarnada introduciendo el elemento clave de los mecanismos de autorregulación homeostática/alostática (Ziemke, 2016, 9). Solo una regulación, o mejor dicho autorregulación análoga a la encontrada en los seres vivos, puede dar lugar a valoraciones auténticas y verdadera inteligencia. Como explica Bickhard (2009), los robots hasta ahora pensados no están realmente inmersos e implicados en su circunstancia, no hay nada verdaderamente *en juego* para ellos -al menos no en el sentido que sí vemos y experimentamos respecto de los seres vivos. La introducción de estos mecanismos de autorregulación, tal y como expone Ziemke, podría ser la clave para generar una auténtica motivación vital y, en consecuencia, una verdadera inteligencia artificial.

Esta última idea no puede desarrollarse más en este punto; pero resulta evidente lo lejos que las propuestas de superinteligencia artificial quedan respecto de una aproximación compleja como esta. Como ya hemos visto y a continuación seguiremos analizando, la superinteligencia artificial propuesta por Bostrom y los transhumanistas toma como punto de partida la escisión respecto de cualquier lógica vital, arribando de ese modo a callejones sin salida de los que luego no puede salir.

4.2. La imposibilidad de comprender la superinteligencia artificial como inteligencia

La superinteligencia de la que hablan los transhumanistas no es una auténtica inteligencia pues se piensa siempre al margen de cualquier circunstancia. Esta supuesta inteligencia es una capacidad que no se pone en relación con ninguna finalidad vital, sino que se imagina como un potenciamiento vacío, sin meta. Esto es así porque en el transhumanismo, la definición del ser humano y, por extensión, de sus características y habilidades, no se basa en una comprensión auténtica de la vida humana. Debido a esto, la concepción del ser humano propia del transhumanismo carece de fundamentos sólidos, y sus argumentos se sumergen en una multitud de cuestiones específicas que no llegan al núcleo del problema.

En consecuencia, cuando los transhumanistas expresan su intención de mejorar la visión, la memoria, la inteligencia u

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

otros atributos humanos, parten de una visión físico-abstracta del ser humano, como si fuera una figura geométrica a la que simplemente se le pudieran prolongar sus lados sin alterar su forma. No se basan en una comprensión biológica del ser humano, en una comprensión del ser vivo basada en la relación fundamental yo-circunstancia, radicada en su condición activa, conductual y finalística. El punto de partida transhumanista cae, por tanto, en un desquiciamiento entre el ser vivo y su circunstancia, que supone a su vez un desquiciamiento de la vida en general, que pierde su estructura y organización. El mejor ejemplo de este problema lo constituye precisamente la inteligencia, un atributo que los transhumanistas elevan a un nivel absolutamente superior y al que consideran la clave de la vida humana, sin comprender su lugar funcional dentro de la vida. Pese a que es el reduccionismo más habitual, resulta importante entender lo equivocado que es tomar este punto de partida abstracto respecto del órgano que es el cerebro. En tanto que instrumento vital, el cerebro sólo tiene sentido sirviendo a una vida determinada. La inteligencia, en un sentido análogo, sólo puede serlo de un determinado organismo, y sólo puede aparecer ante la necesidad de orientar movimientos. Un cerebro aparte, extirpado del organismo del que es órgano, tal y como suelen plantear los transhumanistas, (Bostrom, 2016, 105-106) es una pieza de carnicería, algo abortado en su función, extirpado asimismo de la circunstancia que acompañaría al escindido organismo.

Sólo puede haber inteligencia encarnada, entendiendo por carne el cuerpo que siente y se mueve, que está en contacto con una circunstancia, y realiza conductas para alcanzar fines. Como explica J. B. Fuentes, una máquina que sólo se autoregulara y no interactuara genuinamente con su entorno no podría compararse con el organismo vivo (Cf. Fuentes, 2003, 62). La propuesta de Bostrom, y con él la gran mayoría de teorías sobre la superinteligencia artificial, es heredera de una cierta tradición de cognitivismo computacional, basado en la equivalencia entre software y mente por un lado, y hardware y cuerpo por el otro (Fuentes, 2003, 61). En opinión de J. B. Fuentes "El cognitivismo computacional, en resolución, ha llevado al límite y culminado el prejuicio fisicalista ya presente en la tradición conductista" (2003, 64). Este planteamiento abstracto necesita, para plantear su propuesta de superinteligencia artificial, suponer un ambiente "formalmente factorizado en términos de unidades y nexos espaciales contiguos (o sea, fisicalistas), como condición formal misma de su posibilidad algorítmica" (Fuentes, 2003, 63). El nivel fisicalista-abstracto en el que desde el principio se mueve Bostrom invalida de partida su planteamiento. Y por eso las propuestas de IA (inteligencia artificial) o ECC (emulación de cerebro completo) de las que este autor habla a lo largo del libro no pueden entenderse en términos de inteligencia, según la comprensión biológica de la misma que venimos desplegando.

Bostrom no repara en este problema fundamental, tal y como se está planteando aquí. Pero su planteamiento sí se da de bruces contra las aporías derivadas de este problema. Como adelantamos, esta dificultad insuperable se manifiesta en el problema de la motivación de una posible superinteligencia. No se trata exclusiva ni principalmente de que la superinteligencia abstracta postulada por Bostrom y los transhumanistas no vaya a tener motivaciones similares a las humanas. El verdadero problema es que esta superinteligencia, según la descripción de Bostrom, sería incapaz de tener motivación alguna. En el caso del ser humano, al igual que en cualquier ser vivo, la inteligencia sirve a una serie de propósitos específicos; sin embargo, resulta imposible concebir fines o propósitos para una superinteligencia como la que Bostrom plantea. Al no haber surgido de ninguna necesidad y al no operar en un contexto funcional, esta superinteligencia ni siquiera se podría considerar como una auténtica inteligencia. Este punto puede comprobarse claramente respecto de la tesis de ortogonalidad defendida por Bostrom, según la cual "la inteligencia y los objetivos finales son ortogonales: más o menos cualquier nivel de inteligencia podría en principio ser combinada con más o menos cualquier meta final" (Bostrom, 2016, 107). Lo cual sería tanto como decir que en principio podemos imaginar "la capacidad visual de un águila real «acoplada» a la morfología motora de un gusano" (Fuentes, 2003, 37) sin que ello nos resultara algo extraño. La desconexión explícita entre la inteligencia y su función muestra el vaciamiento y falta de significado del propio concepto de inteligencia y, por extensión -y en mayor medida si cabe- del concepto de superinteligencia.

Esta insuficiencia fundamental relacionada con la incomprensión de la naturaleza biológica de la inteligencia se manifiesta de manera ostensible en la referida cuestión motivacional y concretamente en lo que Bostrom denomina en los últimos capítulo de Superinteligencia como el problema de "adquisición de valores" (Bostrom, 2016, 185-208). De manera implícita, Bostrom reconoce que la inteligencia solo puede ser concebida en relación con sus fines. Esto se puede comprobar especialmente en la segunda mitad de la obra, en la que precisamente reflexiona sobre cómo podrían ser introducidos estos fines o valores en la superinteligencia. La dificultad inherente a esta concepción radica en la futilidad de intentar introducir, o más apropiadamente, inocular de manera externa y a posteriori, dichos valores. Pero es claro que este esfuerzo solo surge como intento de compensación de un problema previo.

Bostrom aboga por el aumento ciego y vacío de una función vital como la inteligencia, y solo cuando esta ha perdido completamente su funcionalidad se plantea la posibilidad de incrustarle aquello que inicialmente se le ha negado. Las diversas y creativas soluciones propuestas por Bostrom son simplemente parches conceptuales incapaces de solucionar verdaderamente la paradoja a la que llega su argumentación. Todo esto se debe, como estamos viendo, a un problema subyacente: la falta de comprensión sobre la fundamental radicación biológica de la inteligencia. Una incomprensión que queda clara con su conclusión de que "la solución al problema de introducción de valores es un reto de investigación digno de algunos de los mejores talentos matemáticos de la siguiente generación" (Bostrom, 2016, 187). Pero el punto es que aquí no estamos ante un problema matemático, sino ante un error en el planteamiento y punto de partida, que imposibilita cualquier intento de dar sentido a la realidad abordada desde ese paradigma. Da igual cuántos esfuerzos matemáticos desarrollemos

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Referencias

Alonso, Marcos (2023). "Can Robots have Personal Identity?". International Journal of Social Robotics. 15, 211–220.

Bickhard, M.H., (2009). The biological foundations of cognitive science. New IdeasPsychol. 27, 75–84.

Bostrom, Nick. (2016). *Superinteligencia: caminos, peligros, estrategias*. Traducción e introducción de Marcos Alonso. Madrid: TEELL.

Chalmers, D. (2010). The Singularity: A philosophical analysis. Journal of Consciousness Studies, 17(9-10): 7-65.

Crawford, K. (2021). The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence (1st ed.). Yale University Press. https://doi.org/10.2307/j.ctv1ghv45t

Dekel, Roy (2023). Despite How the Media Portrays It, AI Is Not Really Intelligent. Here's Why. Entrepreneur. Recuperado el 14-01-2024 de https://www.entrepreneur.com/science-technology/despite-how-the-media-portrays-it-ai-is-not-really/446894#:~:text=the%20training%20data.-,It%20is%20important%20to%20recognize%20that%20AI%20is%20not%20truly,decisions%20based%20on%20common%20sense.

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

Fuentes Ortega, Juan Bautista. (2003). "Intencionalidad, significado y representación en la encrucijada de las "ciencias" del conocimiento". Estudios de psicología, n° 24 (1), 33-90.

Gottfredson, Linda. (1997). Why G Matters: The Complexity of Everyday Life. Intelligence 24(1): 79– 132.

Heaven, Will Douglas (2020). "Artificial general intelligence: Are we close, and does it even make sense to try?". MIT Technology Review. Recuperado el 15-01-2024 de https://www.technologyreview.com/2020/10/15/1010461/artificial-general-intelligence-robots-ai-agi-deepmind-google-openai/

Kurzweil, R. (2005). The Singularity Is Near: When Humans Transcend Biology. New York: Penguin.

Larson, E. J. (Erik J. (2021). The myth of artificial intelligence: why computers can't think the way we do. The Belknap Press of Harvard University Press.

Lindblom, J., (2015). Embodied Social Cognition. Springer Verlag, Heidelberg, Germany.

Morozov, Evgeny (2023). The problem with artificial intelligence? It's neither artificial nor intelligent. The Guardian. Recuperado el 14-01-2024 de https://www.theguardian.com/commentisfree/2023/mar/30/artificial-intelligence-chatgpt-human-mind

Newport, Cal (13 April 2023). "What Kind of Mind Does ChatGPT Have?". The New Yorker. Recuperado 15-01-2024.

Olson, E. T. (1997). The human animal: personal identity without psychology. Oxford University Press.

Ortega y Gasset, José (2006). *Obras Completas*, Tomo V (1932-1940). Taurus: Madrid.

Read, R. J., Baker, E. N., Bond, C. S., Garman, E. F., & van Raaij, M. J. (2023). AlphaFold and the future of structural biology. https://doi.org/10.1107/S2053230X23004934

Sandberg, A. (2001). Morphological freedom: Why we not just want it, but Need it. Based on a talk given at the Trans-Vision 2001 Conference, Berlin, 22-24 June.

Schechtman M (2014) Staying alive: personal identity, practical concerns, and the unity of a life. Oxford University Press, Oxford Ritchie, S. (2015). Intelligence. Hodder & Stoughton.

Searle, J.R., (1980). Minds, brains, and programs. Behav. Brain Sci. 3 (3),417–457.

Shapiro, L., (2010). Embodied Cognition. Routledge.

Varela, F.J., Thompson, E., Rosch, E., (1991). The Embodied Mind: Cognitive Scienceand Human Experience. MIT Press, Cambridge, MA.

Velde, R. A. te. (2016). Aquinas on God: the "divine science" of the Summa theologiae. Routledge. https://doi.org/10.4324/9781315262291

Warne, R. T. (2020). In the Know: Debunking 35 Myths about Human Intelligence (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108593298

Yampolskiy, R., & Duettmann, A. (2020). Artificial Superintelligence: Coordination & Strategy. MDPI - Multidisciplinary Digital Publishing Institute.

Ziemke T. (2016). The body of knowledge: On the role of the living body in grounding embodied cognition. Biosystems, 148, 4–11.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Tomémonos en serio la IA (y dejemos a un lado el mito de la Singularidad)

Antonio Diéguez
Universidad de Málaga
dieguez@uma.es

> Aquí [en el Darmouth College], en el legendario lugar de nacimiento de la IA, discutieron sobre cómo llamar a su creación aún dormida. Herbert Simon, polímata y futuro premio Nobel, y Allen Newell, informático, preferían el nombre de "procesamiento complejo de la información". La precisión del nombre evocaba la moderación del método científico moderno, remontándose a los procesos de descubrimiento piedra a piedra ejemplificados por figuras como James Clerk Maxwell. Los informáticos John McCarthy y Marvin Minsky (llamémosles los inteligentistas) prefirieron el más confuso de "inteligencia artificial". Para McCarthy, esto tenía valor comercial. Para Minsky, definirla era "más una cuestión estética o de sentido de la dignidad que una cuestión técnica". […]
>
> Como sabemos, la nomenclatura inteligentista se impuso. […] Uno se pregunta cómo habría sido el destino de la investigación en IA si hubiera prevalecido la postura de Simon y Newell. ¿Habría tenido tanto éxito el exitoso libro de Nick Bostrom de 2014 *Superinteligencia* si se hubiera titulado *Sistemas de procesamiento de información supercomplejos?*
>
> David Leslie "Raging robots, hapless humans: the AI dystopia", Nature, 574 (2019), pp. 32-33.

El advenimiento de la Singularidad se ha convertido en la narración preferida entre los directivos e ingenieros jefes de Silicon Valley. Aunque, como ahora diremos, todo indica que hay una finalidad ulterior en su uso, parece como si les proporcionara un sentido de trascendencia a su misión vital: ellos estarían contribuyendo a la generación de un cambio decisivo en la historia de la humanidad, un nuevo comienzo cuyos límites y cualidades somos aun incapaces de entender por completo. Eso no impide que, abrumados quizá por el juicio popular sobre estas ideas, se quejen por la boca pequeña de que deberían impedírselo mediante regulación, o al menos ralentizarlo. Se ha convertido para ellos en una especie de ley histórica en el sentido que ya criticó Popper. Ellos tienen las claves científicas para predecir lo que sucederá y han anunciado sin ambages que la Singularidad llegará. Ray Kurzweil, su principal propagandista, la sitúa a la vuelta de la esquina, en el año 2045. Otros defensores recientes de la idea, como Nick Bostrom, David Chalmers, Stuart Russell o Max Tegmark, no se atreven a ser tan precisos con las fechas, pero la consideran también relativamente cercana.

El concepto de Singularidad proviene de las matemáticas y, sobre todo, de la física, donde se refiere a casos en los que una magnitud adquiere valor infinito en un tiempo finito, pero ha recibido múltiples definiciones aplicado a otros ámbitos, en particular en lo concerniente al cambio tecnológico (Sandberg 2013). En el contexto del debate sobre la IA, la Singularidad se entiende como el momento en que las máquinas inteligentes alcanzan la capacidad de crear de forma recursiva inteligencias artificiales cada vez mayores, en un crecimiento exponencial de inteligencia que llevaría en no mucho tiempo –algunos creen que de forma casi instantánea– a alcanzar un grado de superinteligencia superior en varios órdenes de magnitud a la inteligencia humana. Este crecimiento seguiría de forma indefinida hasta encontrar algún tipo de impedimento físico, si es que lo hubiera. Kurzweil ha capitaneado también la idea de que, en el momento en que las máquinas alcanzaran una superinteligencia de tal tipo, el ser humano solo podría evitar su destrucción, o su total arrinconamiento y subordinación, si fuera ayudado a su vez por las máquinas a aumentar su propia inteligencia con el objetivo último de lograr una fusión completa entre las mentes humanas y las máquinas superinteligentes. Sería el punto final de nuestra especie, pero el inicio de algo mucho más grandioso y admirable (según sus defensores), que habría surgido al fin y al cabo de la creatividad humana en ejercicio y que conservaría al menos el contenido de muchas mentes humanas. El resultado de esa unión sería, en palabras de Hans Moravec, el gran pionero de la robótica, la generación de seres originados tecnológicamente, no biológicamente, a los que deberíamos considerar como "hijos de nuestra mente" (Moravec 1990).

Pese a la intensidad del debate que ha generado, el concepto de Singularidad está lejos de ser claro y ha sido tildado por sus críticos de artefacto matemático, irreal en su sentido literal o infradeterminado por la evidencia en su sentido metafórico, si es que tiene algún sentido aprovechable (Chen 2023, Eden *et al.* (eds.) 2012). Con todo, los transhumanistas toman ese supuesto acontecimiento futuro como una esperanza de redención, no para nuestra especie, demasiado limitada por su cuerpo biológico, sino para nuestras mentes alojadas en las máquinas, lo que dejaría atrás de una vez la evolución biológica. En efecto, tal como algunos destacados transhumanistas lo ven, si la mente consiste solo en patrones de información, como muchos de ellos creen, estos pueden ser copiados, trasladados e incluso teletransportados a otros soportes materiales, aunque no sean biológicos, y esto implicaría que nuestro yo, sin pérdida alguna de su identidad personal, podría habitar en una máquina y conseguir así una inmortalidad virtual. Es lo que se conoce como descarga de la mente en la máquina (*mind uploading*).

Hay, no obstante, otros gurúes de la Singularidad que no son tan positivos en sus estimaciones. Creen igualmente que la explosión de superinteligencia se producirá, pero consideran que lo más probable es que eso lleve irremediablemente al final de nuestra especie, sin ningún tipo de integración con las máquinas. Algunos lo han venido repitiendo con gran atención de los medios de comunicación desde que el ChatGPT se puso a disposición del público.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

En el mejor de los casos retratados por los singularistas, que sería el de que una superinteligencia artificial se limitara a obedecer órdenes humanas y no nos destruyera, bien porque comprendiera la inmoralidad de tal propósito, o por otra razón que ahora no se nos alcanza, quedaríamos aun así fatalmente subordinados a sus dictámenes. Dicha superinteliegencia no podría ser totalmente obediente y totalmente benevolente al mismo tiempo, puesto que si fuera totalmente obediente tendría que obedecer también a explotadores, terroristas, autócratas o gobernantes sin escrúpulos y causaría daños a seres humanos, y si fuera totalmente benevolente, no podría obedecer cualquier orden que se le diera, sabiendo que podría producir un mal (Aguirre 2023). Así que sería ella la que tendría que decidir qué órdenes aceptar y cuáles no. Se supone que preferiríamos una superinteligencia benevolente a una totalmente obediente; pero eso implicaría que no debería obedecernos cuando considerara que la orden dada podría causar un mal. Así, sería ella la que juzgara acerca de lo que nos beneficia o nos perjudica, con independencia de lo que le digan los humanos. Esto la convertiría en una vigilante paternalista que decidiría por nosotros en todas las cuestiones relevantes.

Los discursos grandilocuentes y atemorizadores de los singularistas, que tanto impacto parecen estar logrando, han sido contestados, me temo que no con demasiado éxito todavía, por diversos especialistas en IA, ya sean ingenieros, científicos computacionales o científicos de otras especialidades, como Gary Marcus, Eric J. Larson, Theodre Modis, Steven Pinker, o filósofos de la IA, como Margaret Boden, Daniel Dennett, Luciano Floridi o Mark Coeckelbergh. Ni siquiera todos los transhumanista asumen la tesis del advenimiento de la Singularidad. James Hughes, por ejemplo, la considera una idea milenarista y apocalíptica (Fidalgo 2021). Es ilustrativo del sentido de estas críticas lo que escribe Luciano Floridi (2022):

> El singularitismo se basa en un sentido muy débil de posibilidad: podría desarrollarse alguna forma de ultrainteligencia artificial, ¿no es así? Sí, podría. Pero este "podría" es una mera posibilidad lógica, es decir, por lo que sabemos no hay ninguna contradicción en suponer el desarrollo de la ultrainteligencia artificial. Sin embargo, se trata de un truco que difumina la inmensa diferencia entre "podría estar enfermo mañana", cuando ya no me encuentro demasiado bien, y "podría ser una mariposa que sueña con que es un ser humano".

A mi parecer, esta defensa de la Singularidad tiene un aire similar al argumento ontológico de San Anselmo: transita de la no imposibilidad de un concepto a la necesidad (o en este caso alta probabilidad) del mismo. En el argumento ontológico la pieza clave (y el punto débil) está en asumir la existencia como una perfección, en la Singularidad la base está en aceptar que una gran inteligencia puede crear otra superior a ella. Pero ni la existencia es una perfección ni la creación de una inteligencia artificial comparable a la humana garantiza un crecimiento exponencial en inteligencia.

Para enturbiar aún más la discusión, el discurso catastrofista está encontrando una réplica no menos estupefaciente y desorientadora en un tecno-optimismo ingenuo y desbordado, como el que ha quedado plasmado en el Manifiesto Tecno -optimista, publicado por el inversor en tecnología Marc Andreessen (2023). En este texto, se asume sin disimulo el solucionismo tecnológico, que reduce todo problema social a problema tecnológico. Se nos dice literalmente: "Creemos que no hay ningún problema material –ya sea creado por la naturaleza o por la tecnología– que no pueda resolverse con más tecnología". Y, por si hubiera dudas sobre su orientación ideológica que impregna el documento, poco después se añade:

> Nuestra sociedad actual ha sido sometida durante seis décadas a una campaña de desmoralización masiva –contra la tecnología y contra la vida– bajo nombres diversos como 'riesgo existencial', 'sostenibilidad', 'criterios ESG [por Environmental, social and corporate governance], 'Objetivos de Desarrollo Sostenible', 'responsabilidad social', 'capitalismo de partes interesadas', 'Principio de Precaución', 'confianza y seguridad', 'ética tecnológica', 'gestión de riesgos', 'decrecimiento', 'los límites del crecimiento'. Esta campaña de desmoralización se basa en malas ideas del pasado –ideas zombis, muchas de ellas derivadas del comunismo, desastrosas entonces y ahora– que se han negado a morir.

No se le ocurre al autor del manifiesto pensar que quizás su confianza absoluta en el solucionismo tecnológico esté fundamentada en ideas mucho más discutibles que las que denuncia. No es en absoluto evidente que los problemas más graves a los que habrá de enfrentarse en el futuro nuestra sociedad tengan siempre soluciones tecnológicas o no puedan encontrar soluciones más eficaces y duraderas si no son tratados meramente como problemas a resolver mediante la tecnología, como si la política no tuviera nada que hacer, o como si la solución pudiera dejarse en manos de la IA. Y no se le ocurre pensar que, incluso si tuviera razón, quedarían por esquivar los daños que acarrearía una tecnocracia generada de este modo.

La radicalidad de estos discursos, situados en los extremos, que es donde encuentran mayor atención pública, no debería impedirnos ver el modo en que están siendo usados en la actualidad como cortina de humo para desviar la atención de problemas reales ligados al desarrollo de la tecnología y, en especial, de la IA. Aunque esto pueda sonar a acusación demagógica, no lo digo a humo de pajas. En un editorial del 27 de junio de 2023 significativamente titulado "Dejemos de hablar del futuro catastrófico de la IA cuando esta plantea riesgos hoy", la revista *Nature* se manifestaba en este mismo sentido:

> Muchos investigadores en IA y expertos en ética con los que ha hablado *Nature* se sienten frustrados por el discurso catastrofista que domina los

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

> debates sobre la IA. Es problemático al menos en dos sentidos. En primer lugar, el fantasma de la IA como máquina todopoderosa alimenta la competencia entre naciones para desarrollar la IA de modo que puedan beneficiarse de ella y controlarla. Esto favorece a las empresas tecnológicas: fomenta la inversión y debilita los argumentos a favor de regular la industria. Ya está en marcha una verdadera carrera armamentística para producir tecnología militar de nueva generación impulsada por la IA, lo que aumenta el riesgo de un conflicto catastrófico –un día del juicio final, quizá, pero no del tipo que tanto se discute en la narrativa dominante de que "la IA amenaza con la extinción humana"./ En segundo lugar, permite que un grupo homogéneo de ejecutivos de empresas y tecnólogos domine la conversación sobre los riesgos y la regulación de la IA, mientras que otras comunidades quedan al margen.

Tales afirmaciones no venían sino a confirmar lo que ya habían dicho otros especialistas en la materia. Por ejemplo, lo que expresaba Floridi unos meses antes:

> Después de tanta especulación sobre los riesgos fantasiosos de las máquinas ultrainteligentes, es hora de encender la luz, dejar de preocuparse por los escenarios de ciencia ficción y empezar a centrarse en los verdaderos retos de la IA para evitar cometer errores dolorosos y costosos en el diseño y uso de las tecnologías inteligentes. (Floridi 2022)

Ahora que por fin comienzan los gobernantes de los países industrializados, incluyendo a China, a tomar consciencia de que nos jugamos mucho sobre el futuro con la regulación de la IA y que, en consecuencia, parecen dispuestos a elaborar normativas legales que encaucen el desarrollo de esta potente tecnología, y ahora también que los dirigentes de las grandes empresas tecnológicas de Silicon Valley empiezan a maniobrar para influir a su favor todo lo posible en esas regulaciones a través de organizaciones aparentemente altruistas, como Horizon y Open Philanthropy, haciendo creer a muchos que los riesgos importantes son los del largo plazo (Bordelon 2023), cobra más importancia que nunca el análisis filosófico de los presupuestos y expectativas de la IA, así como la promoción de políticas adecuadas y bien pensadas acerca de la investigación y el desarrollo de la IA. Nadie duda de que la IA va a transformar nuestras formas de vida, nuestros valores y preferencias, nuestra economía, nuestro comportamiento social, nuestro modo de ver la realidad. Probablemente sea cierto lo que ya vienen proclamando sus adalides: que se trata de la revolución tecnológica más potente que ha experimentado el ser humano, comparable solo al dominio del fuego en los orígenes de nuestro linaje evolutivo.

Sin embargo, se ha dicho (y escrito: Munn 2023) que las cosas están lejos de ser así, que el análisis ético de la IA se ha convertido en algo completamente inútil, y que es usado con frecuencia por parte de las empresas tecnológicas para realizar un la mentable lavado de cara frente a la opinión pública. Todas las grandes empresas y las agencias gubernamentales que han podido y querido han sacado en los últimos años, a veces en los últimos meses, guías y directrices éticas para el desarrollo de la IA, pero todas ellas –según los críticos– lo suficientemente abstractas, ambiguas, inoperantes y alejadas de la práctica real como para que no tengan ninguna funcionalidad ni ejerzan ninguna constricción significativa en el trabajo de los investigadores e ingenieros. De hecho, ni siquiera estas directrices éticas forman parte de la formación de los mismos.

Hay que reconocer que los principios éticos no tienen garra si no están plasmados en una legislación concreta y detallada, pero quizás sea exagerado decir que el discurso ético sobre la IA se ha vuelto completamente inútil. Si el discurso bioético ha tenido utilidad en el campo de la biotecnología, pese a que algunos tampoco hayan visto su función ahí con buenos ojos (Pinker 2015), no hay ninguna razón de peso por la que el discurso ético no podría tener utilidad también en el campo de la IA. No me parece que sea incompatible desarrollar análisis éticos y promover, en consonancia con ellos, normativas más específicas y orientadoras que implementen esos análisis; sabiendo, claro está, que la promulgación y aplicación de estas normativas encontrará en las empresas y algunos gobiernos mucha resistencia. En las empresas tecnológicas la lucha por el poder y el control del desarrollo de la IA se ha convertido ya en una guerra abierta, como ha quedado en evidencia con el affaire en Open AI. Quizás una medida básica, además de desarrollar la regulación legal, debería ser, como han señalado muchos, introducir estas cuestiones éticas en el programa formativo de los ingenieros y, particularmente, en los planes de estudio de los grados de ingeniería.

Frente a las ideas catastrofistas, cobra cada vez más peso la propuesta, no tan voluntarista como podría parecer, de una "Artificial Intelligence for Social Good" (AI4SG), una inteligencia artificial para el bien social (Floridi *et al.* 2020, Tomasev *et al.* 2020). Se trata de desarrollar sistemas de inteligencia artificial que ayuden en la mejora de la salud y en el diagnóstico temprano de enfermedades, en la prevención de pandemias, en la educación y el cuidado de la infancia, en la solución de conflictos sociales, en la eliminación de la pobreza, en la disminución de las desigualdades, en la mejora de la economía, en la gestión de desastres, en la consecución de la sostenibilidad ambiental, en la lucha contra el cambio climático y contra la pérdida de biodiversidad, en la vigilancia de la salud mental, en el control del acoso en redes sociales, en la lucha contra las noticias falsas y la delincuencia en internet, en la mejora de la comunicación entre ciudadanos y gobernantes, en la gestión de los transportes y del urbanismo, en la eficiencia energética, etc. Es decir, se trata de desarrollar sistemas que *ayuden* al ser humano a resolver, paliar, prevenir o gestionar algunos de los grandes problemas que tenemos delante, sin asumir por ello que la solución ha de venir por completo de los dictados de la tecnología. Suele decirse que la AI4SG es la inteligencia artificial que trata de promover los 17 objetivos que estableció en 2015 la Asamblea General de la ONU encaminados a

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

lograr el desarrollo sostenible[1].

Nenad Tomasev *et al*. (2020) han sugerido algunas de las líneas fundamentales que deben inspirar este enfoque de la IA:

1. Las expectativas sobre las posibilidades de la IA deben estar bien fundadas.
2. Las soluciones sencillas son valiosas.
3. Las aplicaciones de la IA deben ser inclusivas y accesibles, y revisarse en cada fase para garantizar el cumplimiento de los principios éticos y los derechos humanos.
4. Los objetivos y los casos de uso deben ser claros y estar bien definidos.
5. Se necesitan asociaciones profundas y a largo plazo para resolver con éxito grandes problemas.
6. La planificación debe alinear los incentivos y tener en cuenta las limitaciones de ambas comunidades [la de investigadores en IA y los expertos en aplicación para los objetivos del desarrollo sostenible].
7. Establecer y mantener la confianza es clave para superar las barreras organizativas.
8. Deben explorarse opciones para reducir el coste de desarrollo de las soluciones de IA.
9. Es fundamental mejorar la preparación de los datos.
10. Los datos deben procesarse de forma segura, respetando al máximo los derechos humanos y la privacidad.

El desarrollo de este tipo de IA requiere, como han señalado Floridi y sus colaboradores (2020), que se consulte a los usuarios y a las personas que podrían sufrir las consecuencias de impactos negativos, así como que se respeten sus derechos, entre ellos, el de la protección de la privacidad de los datos. Este problema puede agravarse con el internet de las cosas, puesto que nuestros datos estarán aún más abiertos y disponibles para las empresas tecnológicas, y ese es un riesgo al que debemos prestar suma atención (Véliz 2020).

¿Es la inteligencia artificial para el bien social un simple deseo piadoso o hay una realidad tangible detrás del proyecto? Ciertamente, no cabe esperar que este sea un objetivo prioritario para las grandes empresas tecnológicas, pero es ya esperanzador el mero hecho de que haya sido puesta en marcha una iniciativa como esta y que esté encontrando una acogida tan favorable. Un estudio reciente pudo encontrar 108 proyectos en inglés, español y francés que encajarían bien como proyectos inspirados por al AI4SG (Cowls *et al*. 2021).

Entre los peligros que encierra el uso social de la IA, y que debería ser mejor analizado para conseguir paliarlo, está el que los sistemas actuales de IA tienden a la perpetuación o "afirmación de lo dado", es decir, "que las salidas de la computadora generalmente reflejan lo que ya se da, y no loque podría o debería ser, lo que es nuevo, sorprendente, innovador o desviado. En otras palabras, las aplicaciones de aprendizaje máquina calculan un futuro que es como el pasado. Los cambios no son intencionados" (Hagendorff y Wezel 2020, p. 357). Debería, pues, potenciarse una interpretación de los resultados proporcionados por los sistemas de IA que no obstaculizase la innovación social. El reto es saber cómo hacerlo.

Para bien y para mal, en la Universidad, los profesores y los alumnos recurrimos ya a la IA de forma rutinaria para formar una idea inicial de cómo abordar cualquier tema que nos interesa, por ello, le pregunté al ChatGPT si debíamos tomarnos en serio los discursos catastrofistas sobre la IA y qué retos habríamos de afrontar sobre todo esto en el futuro. Entre otras cosas sensatas (aunque bastante tópicas), escribió lo siguiente: "En lugar de entrar en pánico ante los discursos catastrofistas, es más productivo promover un enfoque de gestión de riesgos y regulación adecuada. Esto implica el desarrollo de marcos éticos y legales, la promoción de la transparencia y la rendición de cuentas en los algoritmos de IA, y el fomento de la investigación sobre seguridad y ética en la IA. La colaboración entre la industria, los gobiernos, la academia y la sociedad en general es esencial para abordar los desafíos y maximizar los beneficios de la inteligencia artificial de manera responsable".

Tiene toda la razón. Yo no tengo más que añadir a eso. La IA nos da un buen consejo acerca cómo tratar adecuadamente con la IA. Ahora lo importante es tomar las decisiones políticas adecuadas.

Referencias

Aguirre, A. (2023), "Close the Gates to an Inhuman Future: How and why we should choose to not develop superhuman general-purpose artificial intelligence" (October 20, 2023). Disponible en SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4608505

Andreessen, M. (2023), "The Techno-Optimist Manifesto", 16 octubre 2023, https://a16z.com/the-techno-optimist-manifesto/

Bordelon, B. (2023), "How a billionaire-backed network of AI advisers took over Washington", *Politico*, 13 de octubre. https://www.politico.com/news/2023/10/13/open-philanthropy-funding-ai-policy-00121362?s=03

Chen, M. (2023), "A Trilemma for the Singularitarian", *Philosophy & Technology*, 36:62. https://doi.org/10.1007/s13347-023-00653-4

Cowls, J., A. Tsamados, M. Taddeo et al (2021), "A definition, benchmark and database of AI for social good initiatives", *Nat Mach Intell*, 3, pp. 111-115. https://doi.org/10.1038/s42256-021-00296-0

Eden, A.H., J.H. Moor, J.H. Soraker y E. Steinhart (eds.) (2012), *Singularity Hypotheses*, Heidelberg: Springer.

1. Esos objetivos pueden verse aquí: https://social.desa.un.org/issues/disability/envision-2030/17goals-pwds . Algunos ejemplos de AI for Social Good puestos ya en marcha pueden verse aquí: https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Fidalgo, P. (2021), "Post-Humans on a Sterile Promontory: The New Myths of Transhumanism and the Dark Mountain", *Free Inquiry*, Volumen 41, No. 3, Abril/Mayo. https://secularhumanism.org/2021/04/post-humans-on-a-sterile-promontorythe-new-myths-of-transhumanism-and-the-dark-mountain/

Floridi, L. et al. (2020), "How to Design AI for Social Good: Seven Essential Factors", *Science and Engineering Ethics* (2020) 26, pp. 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Floridi, L. (2022), "Ultraintelligent Machines, Singularity, and Other Sci-fi Distractions about AI", September 18. *Lavoro, Diritti, Europa* - https://www.lavorodirittieuropa.it/, disponible en SSRN: https://ssrn.com/abstract=4222347

Hagendorff y K. Wezel (2020), "15 challenges for AI: or what AI (currently) can't do", *AI & Society*, 35, pp. 355–365.

Moravec, H. (1990), *Mind Children*, Cambridge: Harvard University Press.

Munn, L. (2023) "The useless od AI ethics", *AI and Ethics*, 3, pp. 869-877. https://doi.org/10.1007/s43681-022-00209-w

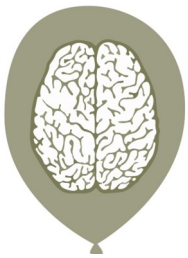
Nature (2023), "Stop talking about tomorrow's AI doomsday when AI poses risks today", *Nature*, 618, pp. 885-886. https://doi.org/10.1038/d41586-023-02094-7

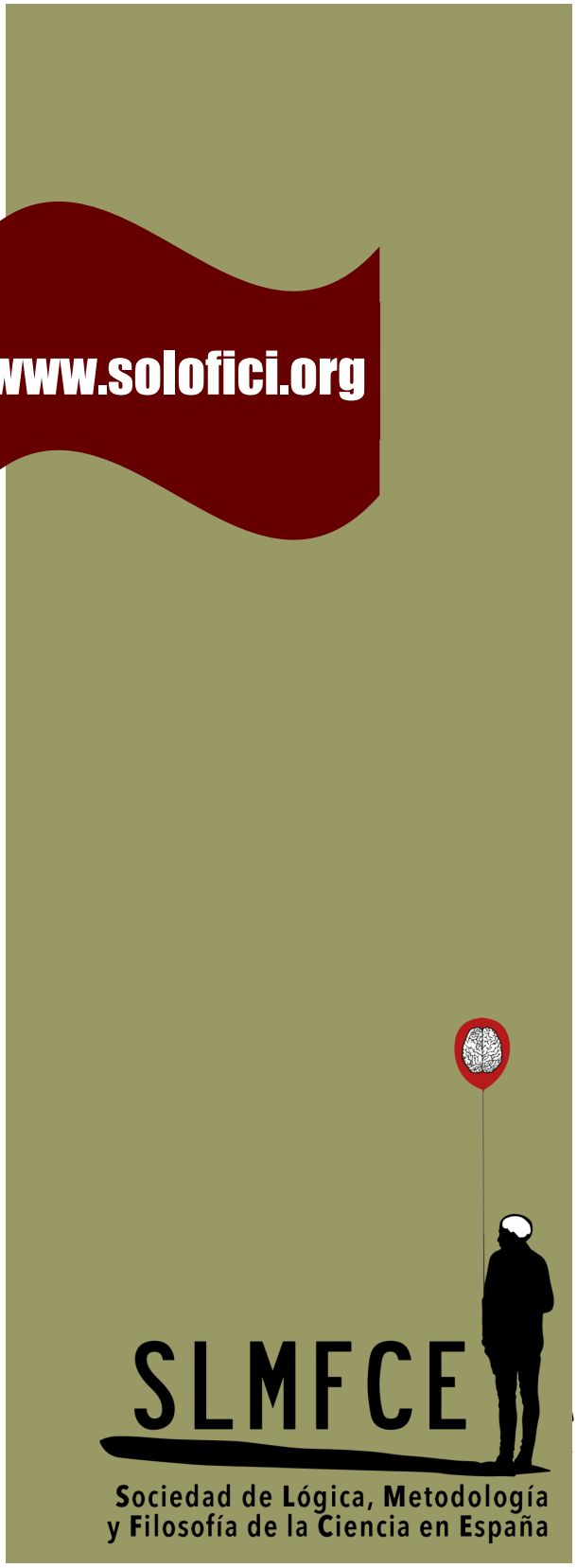
Pinker, S. (2015), "The moral imperative for bioethics", *The Boston Globe*, 1 de Agosto. https://www.bostonglobe.com/opinion/2015/07/31/the-moral-imperative-for-bioethics/JmEkoyzlTAu9oQV76JrK9N/story.html

Sandberg, A. (2013), "An overview of models of technological singularity". In M. More y N. Vita-More (eds.) *The Transhumanist Reader*, West Sussex: John Wiley & Sons, Inc. cap. 36, pp. 376-394. https://doi.org/10.1002/9781118555927

Tomasev, N. et al. (2020), "AI for social good: unlocking the opportunity for positive impact", *Nature Communications*, 11, 2468. https://doi.org/10.1038/s41467-020-15871-z

Véliz, C. (2020), *Privacy is Power*, London: Transworld.





SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

La singularidad tecnológica: un singular disparate

Jesús Zamora Bonilla
UNED
jpzb@fsof.uned.es

1.

La llamada "Singularidad Tecnológica" es quizás la idea más disparatada concebida en las últimas décadas en el ámbito de las discusiones sobre el progreso tecnológico. Aunque la idea tiene una larga prehistoria, se hizo especialmente popular hace un cuarto de siglo, a partir de un artículo del matemático Vernor Vinge, que precisamente incluía el concepto de "post-humano" en su título,[1] y el empujón definitivo hacia la celebridad cultural, filosófica y científica del concepto se lo han dado más recientemente los libros *La singularidad está cerca*, de Raymond Kurzweil,[2] y *Superinteligencia*, de Nick Bostrom.[3] La idea, básicamente, es que en el mismo momento en que los ingenieros de inteligencia artificial (IA) consigan diseñar un sistema informático lo bastante inteligente como para ser capaz de crear una versión *un poco más* inteligente que él mismo, esto conducirá a un crecimiento "explosivo" de la IA, pues este segundo sistema creará un tercero aún más inteligente, que creará a su vez un cuarto sistema todavía más espabilado, y así, en principio, hasta el infinito. Como, además, ese proceso de "creación" consistirá tan solo en escribir un algoritmo, o sea, en indicar cuáles son las reglas que sigue el programa informático en cuestión, el proceso puede durar una cantidad de tiempo increíblemente pequeña, pues, además, cada nuevo sistema será capaz de llevar a cabo su tarea en un tiempo menor que los predecesores. Confieso que no me hago muy exactamente a la idea de la medida temporal de la que estamos hablando, pero digamos que es irrelevante si el proceso tarda unos cuantos segundos o unos pocos meses, pues lo importante es que sería un plazo demasiado corto como para que los humanos pudiéramos reaccionar a tiempo. Ese breve intervalo temporal, en el que la superinteligencia emergería de modo casi milagroso, es lo que suele conocerse como la *singularidad tecnológica*.

Estos programas superinteligentes[4] serán algo que vaya muchísimo más lejos de las "superaplicaciones informáticas" contra las que nos advierte otro famoso autor, como Yuwal Harari a propósito de lo que denomina "dataísmo",[5] pues en el caso de aquellas superaplicaciones sería suficiente, para que se convirtieran en una amenaza para nosotros los humanos, con imaginar que tomasen decisiones *algo* mejores que las nuestras, pero, para los argumentos de Harari, era más o menos irrelevante si tales sistemas llegaban a adquirir consciencia, inteligencia artificial general,[6] o la capacidad de desarrollar sus propias motivaciones. En cambio, es de suponer que todas estas propiedades las tendrían los últimos sistemas que emergiesen durante la singularidad, o a continuación de ella. En todo caso, la pregunta relevante es: ¿cuál será el puesto de los seres humanos en un mundo en el que existen tales entidades infinitamente sagaces? En comparación con ellas, nosotros seríamos algo así como lo que puede ser una mosca comparada con un ser humano. Respecto a esta cuestión, los teóricos de la singularidad se dividen en dos bandos. Por un lado están quienes piensan, como Kurzweil, que podremos seguir coexistiendo con esos dioses-máquinas, pero tal vez reconvertidos (nosotros) en algo así como subprogramas informáticos de dichos sistemas, es decir, con nuestras mentes "descargadas" en sus circuitos.[7] Por otro lado, hay una mayoría de autores, como Bostrom, cuya opinión es mucho más pesimista, y consideran que hay un enorme "riesgo existencial" de que los ordenadores superinteligentes desarrollen

1. Vernor Vinge, "The Coming Technological Singularity: How to Survive in the Post-Human Era", en *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, 1993, NASA Conference Publication 10129, pp. 11-22. Para un análisis más detallado del concepto de "singularidad tecnológica" y de los problemas filosóficos asociados, puede verse el libro de Antonio Diéguez, *Transhumanismo: la búsqueda tecnológica del mejoramiento humano*, 2017, Herder.

2. Raymond Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*, 2006, New York, Penguin (traducción castellana, *La singularidad está cerca: cuando los humanos transcendamos la biología*, 2012, Berlín, Lola Books).

3. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, 2014, Oxford, Oxford Univeristy Press (traducción castellana, *Superinteligencia: caminos, peligros, estrategias*, 2016, TEELL).

4. Suele emplearse el término "superinteligencia" para referirse a la capacidad cuasi-infinita de esos imaginarios sistemas artificiales, pero pienso que la palabra se queda corta. "Superinteligente" es también un concepto que a veces se aplica en biología y psicología a la capacidad cognitiva de los animales de conducta más sofisticada (simios, córvidos, cetáceos, elefantes, humanos...), para distinguirla de las capacidades ciertamente más limitadas que poseen otras especies. Yo hablaría más bien de "*ultrainteligencia*", pero como "superinteligencia" es la expresión que ha tenido fortuna, seguiré utilizándola en este capítulo.

5. Y. Harari, *Homo Deus: Breve historia del mañana*, 2016, Barcelona, Debate.

6. En psicología se distingue entre "inteligencia" (como capacidad de resolver *algún* tipo concreto de problemas) e "inteligencia *general*" (que es la capacidad que los seres humanos tenemos de resolver problemas de *cualquier* tipo, y no únicamente de un tipo determinado, o sea, lo que también se llama algunas veces "sentido común" -un concepto este aún mucho más difícil de definir, por supuesto). Los programas informáticos que conocemos tienen solo inteligencia "parcial" o "local", pues cada uno está especializado en una cierta clase de tareas (un procesador de texto no puede elegirte el mejor seguro de accidentes, p.ej.). El sueño de los informáticos es llegar a crear ordenadores que posean "sentido común" o "inteligencia artificial general", algo que se da por asumido que los programas "superinteligentes" poseerán.

7. En el capítulo segundo de mi libro, ya citado, *En busca del yo* ("El mito del ordenador"), ofrezco algunos argumentos en contra de la posibilidad de que la mente de un ser humano pueda ser "descargada" en un ordenador. No puedo evitar acordarme de Woody Allen en *El dormilón*, y preguntarme por la suerte que correrán el resto de nuestros órganos vitales cuando haya sido descargado en el dios-máquina el contenido del cerebro, o sea, de "nuestro segundo órgano favorito".

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

sus propios fines (seguramente incomprensibles para nuestra limitada psicología) y decidan prescindir de algo tan molesto y costoso como es la humanidad, o nos utilicen despiadadamente como cualquier otro recurso natural.

2.

Sea como sea, en mi opinión el escenario apocalíptico de una "singularidad tecnológica" no debe preocuparnos en absoluto, pues es una mera ficción sacada de la supersimplista idea de que la inteligencia artificial podría progresar mediante el desarrollo de sistemas informáticos capaces de diseñar versiones de sí mismos todavía mejores. Esta idea sencillamente ignora montones de factores que también serían necesarios para el advenimiento de superinteligencias merecedoras de ese nombre, y es tan ingenua como si alguien, a la vista del desarrollo "exponencial" que había experimentado la aviación comercial en su primer medio siglo de existencia, hubiera estimado en los años 60 del siglo pasado que cincuenta años después podríamos volar de Madrid a Nueva York en solo diez minutos por el precio de un billete de metro. Expondré a continuación una lista de algunos detalles, pequeños pero decisivos, que quienes creen en la inminencia de la "singularidad tecnológica" han preferido ignorar hasta la fecha.[8]

Un primer argumento contrario a que la singularidad sea algo probable tiene que ver con su mera posibilidad *física*. Las máquinas-inteligentes-capaces-de-crear-máquinas-más-inteligentes-todavía no solo tendrán que ser capaces de desarrollar un *software* cada vez más potente, sino también un *hardware* capaz de soportar esos programas cada vez más sofisticados; quizás el desarrollo del *software* (los programas) puede ser concebido como un simple "problema matemático", pero la mejora del *hardware* (los cachivaches) es en realidad un cúmulo de complejísimos problemas de ingeniería, que no solo requiere "potencia lógica", sino también miles de horas de experimentación en el laboratorio, y muchas más de geólogos, mineros y todo tipo de trabajadores especializados, dedicados a buscar y fabricar los materiales necesarios para esos experimentos, por no hablar del *dinero* que cuesta todo eso. Sería concebible que una serie de programas-capaces-de-diseñar-programas-aún-mejores terminase imprimiendo algoritmos que, *si acaso se pudieran* implementar en un ordenador con la potencia física suficiente, *darían* origen a un sistema informático superinteligente y capaz de emular a los dioses olímpicos en sus habilidades; pero también es concebible que el *hardware* del ordenador capaz de llevar a cabos tales proezas fuera *física y económicamente imposible de construir*, incluso utilizando todos los recursos del planeta y sus alrededores. Y, además, incluso si cada etapa del proceso pudiera ir resolviendo los problemas "técnicos" relativos a la parte material de cada nuevo sistema, no hay ninguna razón para pensar que eso podrían hacerlo en un tiempo relativamente breve, por lo que el crecimiento en el nivel de inteligencia iría mucho más despacio de lo que imaginan los "singularistas", y sería difícil, por lo tanto, que el advenimiento de la superinteligencia nos pudiera pillar por sorpresa.

Un segundo y aún más decisivo argumento es que la idea de "singularidad" se basa en *un concepto completamente ingenuo de "inteligencia"*, como si fuese una variable física que pudiéramos medir a través de una escala lineal, de modo parecido a como podemos medir la velocidad de un avión, su envergadura, o su consumo de energía. Lo cierto es que no tenemos ninguna teoría lo suficientemente completa y precisa acerca de en qué consiste la inteligencia (sobre todo la "inteligencia general"), como para que pudiéramos explicarles a los primeros sistemas con los que se iniciara el "crecimiento explosivo de inteligencia" qué demonios es lo que tienen que poseer *exactamente* esos otros programas "un poco más inteligentes" que ellos tendrían que desarrollar. Peor aún, tampoco tenemos ni la más remota idea sobre cómo se las apañan los cerebros de los seres vivos, incluido el nuestro, para proporcionarnos eso que llamamos "inteligencia", ni sobre por qué los cerebros humanos son capaces de hacernos mantener una conversación y levantar rascacielos, pero los de los elefantes no. Por razones que indicaré un poco más adelante, yo sospecho más bien que esto es algo que nunca vamos a llegar a saber, al menos con el detalle suficiente como para prender la mecha de la singularidad.

En tercer lugar, los singularistas confunden palmariamente el concepto de "*poseer* un nivel de inteligencia suficiente para desarrollar la tarea X" con el concepto de "ser capaz de *mejorar* el nivel de inteligencia necesario para llevar a cabo la tarea X"; o, como dice Toby Walsh en el libro citado, confunden los conceptos de "inteligencia" y "meta-inteligencia". Si, por ejemplo, reducimos el concepto de inteligencia al sentido de "capacidad de aprender", los investigadores informáticos llevan décadas obteniendo grandes éxitos en el desarrollo de sistemas que son cada vez más capaces de aprender (sobre todo en los campos conocidos como "aprendizaje automático" –*machine learning*– y "aprendizaje profundo" –*deep learning*–), pero cada uno de esos progresos ha requerido el desarrollo *por parte de los propios científicos* de sistemas de aprendizaje automático cada vez más complejo de las tareas para las que cada sistema ha sido diseñado. Por el contrario, no se ve ningún progreso significativo en la capacidad *de los propios sistemas artificiales* de mejorar por sí mismos no su capacidad de aprender eso que sea lo que se les ha diseñado para aprender a *hacer*, sino la capacidad de aprender a *crear un sistema parecido a ellos*. Por ejemplo, un caso en el que el éxito de los sistemas de *machine learning* está siendo espectacular es el de la traducción automática entre varios idiomas: nosotros no sabemos cómo enseñar a una máquina a traducir del francés al español, pero sabemos cómo crear una máquina que aprenda por sí misma a hacer tales traducciones. Ahora bien, cada una de estas máquinas, a medida que aprende a traducir mejor, no da ninguna muestra de haber aprendido ni siquiera los rudimentos sobre *cómo crear otra máquina* capaz de aprender mejor aún que ella. Para que la mecha de la

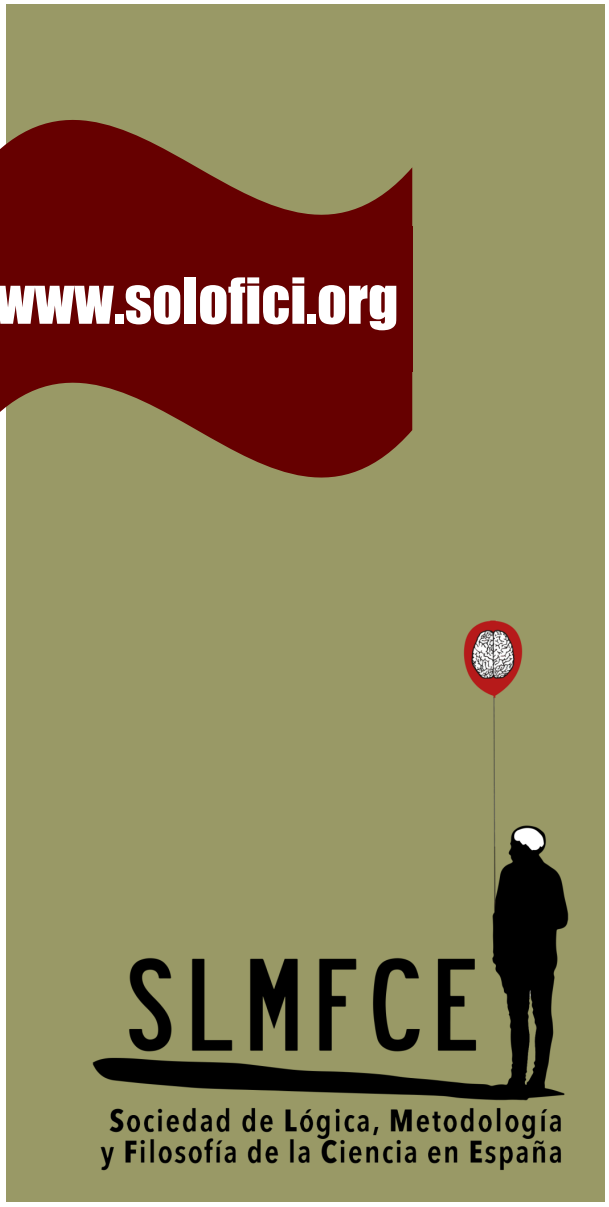
8. Sigo más o menos vagamente los argumentos del libro de Toby Walsh, *Machines That Think: The Future of Artificial Intelligence*, 2018, Amherst (N.Y.), Prometheus Books.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

singularidad pudiera comenzar a encenderse, el truco sería crear una máquina que, a través de algo parecido al *deep learning*, pudiese aprender, no a traducir, a jugar al ajedrez, o a regular el tráfico, sino a diseñar máquinas que aprendieran a hacer "lo que hacen ellas", o sea, que aprendieran a diseñar-máquinas-que-aprendieran-a-diseñar-máquinas-que-aprendieran-a-diseñar... En fin, es dudoso que *esto* podamos siquiera llegar a *operacionalizarlo* en un objetivo concreto (como el de "traduce este texto del español al francés" o "consigue que haya la menor cantidad de atascos posibles en esta ciudad"), e incluso si lo consiguiéramos, tal vez las máquinas que resultaran de ese proceso serían muy eficientes en diseñar otras máquinas un poquito mejores que ellas, pero que, *aparte de eso*, no supieran hacer absolutamente nada más.

Un cuarto argumento es que el "singularismo" presupone que la inteligencia no solo puede ser definida teóricamente y medida de manera lineal, sino que puede tener cualquier valor, digamos, entre cero e infinito. Tal vez haya un límite máximo en la inteligencia que un sistema (cualquier sistema, sea cual sea su naturaleza física) pueda llegar a tener, de modo análogo a como hay una velocidad máxima que se puede alcanzar en el universo (la velocidad de la luz), o una masa máxima que una estrella puede tener sin colapsar a un agujero negro. Por supuesto, quizás lleguemos a desarrollar sistemas informáticos que sobrepasen la inteligencia humana en todos o en casi todos los aspectos, pero eso no implica necesariamente que la vayan a sobrepasar *por mucho margen*; quizás sean solo capaces de pensar cosas *un poquito* más complejas que las que podemos pensar nosotros, aunque, eso sí, las piensen mucho más deprisa, pero es posible que queden aún montones y montones de cosas que esas máquinas "superinteligentes" *tampoco sean capaces de llegar a comprender nunca*. En particular, el mundo, y no digamos las sociedades y las culturas, están llenas de sistemas que destacan por su nivel de complejidad, o sea, porque sus componentes e interacciones dan lugar a procesos asombrosamente variables e impredecibles. Puede, por tanto, que las máquinas superinteligentes sean tan incapaces de adivinar los vaivenes a largo plazo de la economía (la suya o la nuestra) o del tiempo meteorológico, como nosotros lo somos ahora.

En último lugar, es muy dudoso que los programas que prendan la mecha de la singularidad puedan ser diseñados por programadores humanos, pues, como vimos más arriba, en realidad no hay visos de que podamos tener una teoría lo bastante potente sobre la naturaleza y el funcionamiento de la inteligencia como para servir de plantilla a los algoritmos de dicho programa. Es más probable que, si alguna vez llegan a aparecer sistemas con "inteligencia artificial general", lo hayan hecho mediante procesos de "aprendizaje profundo", que en realidad consisten en ir dejando aprender por sí mismas a lo que se conoce como "redes neuronales": circuitos lógicos que simulan el funcionamiento de las neuronas, en el sentido de que funcionan cambiando la intensidad de conexión entre cada "neurona artificial" y las otras, en función de los resultados de millones de pruebas por ensayo y error. Si hay algo parecido a "programas informáticos" en nuestro cerebro, su naturaleza es parecida a esta: "codificados" en forma de la intensidad que posee cada conexión sináptica, conexiones de las que poseemos literalmente *trillones*. Pues bien, esto implica que, en un sentido profundo, el "programa informático" con el que un sistema artificial habrá conseguido poseer el grado de "superinteligencia" que sea, ese "programa" será tan inescrutable para esa máquina como nuestro propio "programa" lo es para nosotros. Es decir, *lo más probable es que una máquina superinteligente ignore por completo cómo se las arregla ella misma para hacer lo que hace*, y por lo tanto, es muy inverosímil que sea capaz de averiguar cómo construir una máquina que lo haga *mejor* que ella.



SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Hasta la vista, Singularidad

David Casacuberta
david.casacuberta@uab.cat
UAB Universitat Autònoma de Barcelona

Resumen

Este artículo explora el debate filosófico en torno al concepto de singularidad, cuestionando específicamente la viabilidad de su versión descarnada a través de ideas básicas de filosofía de la mente. Al enfatizar la perspectiva de que la cognición surge a través de interacciones dinámicas entre un organismo y su entorno, se argumenta en contra de la premisa central de una inteligencia incorpórea como la singularidad. Además, critica la visión reduccionista de la conciencia y la cognición que sustenta las teorías de la singularidad, proponiendo que la naturaleza profundamente encarnada de la experiencia humana no puede ser replicada ni superada por la inteligencia artificial generativa actual.

Cuando uno lee artículos supuestamente serios en prensa que discuten la posibilidad de que (con música de thriller de fondo) "las máquinas ya posean consciencia", no puede evitar preguntarse dónde quedaron 50 años de reflexión filosófica sobre la mente y la consciencia. Las argumentaciones que buscan convencernos de que los ordenadores lograrán ser conscientes pronto (si no lo son ya) o que los LLMs son inteligencia artificial general equivalente a la humana, están hechas con un menosprecio total no solo a la filosofía, sino al razonamiento en general.

Ciertamente, es posible que algún día algún tipo de sistema tecnológico sea consciente y llegue a ser mejor que los humanos. La singularidad es, sin duda, posible, aunque, en mi opinión, es algo muy lejano. En este artículo quiero cuestionar la posibilidad de la versión de la singularidad que se nos vende ahora, esa propuesta de la que están tan enamorados los techbros de Silicon Valley. Esa propuesta, formulada ya en las futurologías de Vernon Vinge (Vinge 1993) y Ray Kurzweil (Kurzweil 2005), sostiene que esa superinteligencia es descorporizada: es puro software, existiendo puramente en un ámbito digital o virtual, sin un robot físico o entidad física similar como soporte de esa mente supuestamente consciente. Sugiere una IA que opera y evoluciona dentro de sistemas informáticos, redes o la nube, independientemente de una estructura física tangible.

Si esos techbros se hubieran molestado en leer algo de filosofía de la mente, se habrían dado cuenta de que es una posición totalmente insostenible. Un LLM como Chat GPT no es más que una versión sofisticada de la biblioteca de Babel de Borges o la habitación china de Searle, y los argumentos de Searle se aplican también aquí. Cuando pedimos a Chat GPT que nos escriba una ficción que tiene lugar en un ferrocarril, un elaborado análisis estadístico establece que detrás de la frase "cuando el tren llegó a" las palabras "la estación" son una continuación mucho más probable que "la luna", pero eso no significa que el sistema tenga la menor comprensión de qué significa "estación" o "luna". Frente a esta objeción, los techbros tienen básicamente tres opciones:

1) Reconocer que no tienen ni idea de qué es la mente, la consciencia o el razonamiento, y estudiar filosofía. Desafortunadamente, hasta donde yo sé, ningún techbro ha escogido esa opción.

2) Seguir la senda abierta por Sam Altman con su tuit: "i am a stochastic parrot and so r u", y establecer que sí, los LLMs son meros generadores estadísticos de frases sin comprensión de lo que dicen, pero también lo son las personas que leen este artículo, o su autor. Esta opción es favorecida por la línea dura de los techbros, pero es fácilmente descartable. Si Sam Altman y sus fanboys creen que los humanos somos loros estocásticos, entonces *a fortiori* deben creer que ellos también lo son. Por lo tanto, no tienen ninguna fe en sus argumentaciones. ¿Y para qué vamos a perder el tiempo discutiendo con ellos[1]?

3) Aceptar que efectivamente los LLMs distan mucho de tener las capacidades cognitivas de un ser humano, pero apuntar a que la solución está en incluir módulos semánticos de razonamiento causal. Así tendríamos sistemas artificiales que realmente comprenderían el mundo.

La opción 3 es bastante más compleja de lo que parece a primera vista, pero juguemos a que es posible. Me gustaría argumentar aquí que disponer de un sistema así todavía estaría muy lejos de disponer de consciencia y agencia humanas. Para hacerlo, solo necesitamos recuperar ideas básicas de fenomenología y filosofía de la mente.

La mente humana es el resultado de la interacción entre un cuerpo, un cerebro y un entorno. Nuestra experiencia del mundo está corporizada, y ese estar en un cuerpo es lo que le da sentido a nuestra experiencia y nos permite hablar de consciencia. Sin un cuerpo real que pueda dañarse, la emoción del miedo, de la sorpresa o la alegría no tienen ningún sentido. Puedo simular a la perfección lo que una persona asustada diría en una conversación y puedo añadirle conocimiento causal para tener una clasificación exacta de qué cosas dan miedo, cuáles no, y hasta qué punto, pero sería eso, una simulación, pues esa inteligencia descorporizada distribuida en miles de ordenadores interconectados no tiene ninguna presencia física que pueda realmente recibir daño. Como dijo en su momento Daniel Dennett, después de ejecutar la simulación de un huracán en un ordenador, nadie se queda mojado. (Dennett 1978)

1. Es significativo que quien primero utilizó el término "loros estocásticos" no fue Sam Altman, sino Emily Bender y Timnit Gebru, en un artículo (Bender et al 2021) que critica los LLMs como GPT-3, refiriéndose a estos modelos de IA como "loros estocásticos" para resaltar cómo imitan el lenguaje humano basándose en patrones probabilísticos, sin una verdadera comprensión o consciencia. Pero los techbros solo citan a Altman, lo cual dice mucho de la cultura de Silicon Valley.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Al carecer de un cuerpo, se carece también de planes y objetivos con sentido, o deseos. El sistema puede tener todo el conocimiento almacenado de todos los terabytes que arrojamos directamente a Internet, pero sin un entorno al que responder y sin un cuerpo con el que interactuar con ese entorno, no tendremos ni consciencia ni agencia.

¿Y si incluimos un robot en el proceso? ¿Se solucionaría así el problema? Cuestión complicada. Aquí aparecerían otros problemas clásicos de la filosofía de la mente, como los *qualia*. Pero claramente, un acercamiento basado en crear un objeto físico que interactúa con un entorno específico es un tipo de proyecto totalmente diferente al que nos están ofreciendo ahora los gurús de la IA generativa y los LLMs, y tendríamos que escribir otro artículo para comentarlo.

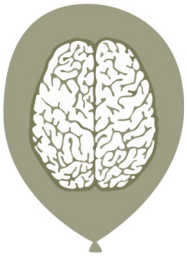
Referencias

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). "On the dangers of stochastic parrots: Can language models be too big?". In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Dennett, D. C. (1978). "Why you can't make a computer that feels pain." *Synthese*, 38(3), 415-456.

Kurzweil, R. (2005). "The singularity is near". In *Ethics and emerging technologies* (pp. 393-406). London: Palgrave Macmillan UK.

Vinge, V. (1993, March). Technological singularity. In *VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute* (pp. 30-31).



www.solofici.org

SLMFCE

Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

La estafa de la Singularidad

Santiago Sánchez-Migallón Jiménez
Jefe de Departamento de Filosofía
IES Montevives
circumdatus@yahoo.es

Resumen: El principal argumento a favor de que se dé una singularidad tecnológica en un tiempo cercano, es la creencia en la aparición de una inteligencia artificial con facultades superiores al hombre, capaz de mejorarse a sí misma durante un tiempo indefinido. Elaboramos una crítica a esta tesis desde cuatro perspectivas: 1) Contra el propio concepto de singularidad histórica y tecnológica. 2) Que el concepto de crecimiento exponencial o hiperbólico que suelen utilizar los singularistas para hablar de la velocidad del cambio tecnológico es una concepción fantasiosa de éste. 3) Que el estado del arte en inteligencia artificial no justifica el excesivo optimismo hacia la pronta llegada de una inteligencia artificial superior. 4) Que la noción de motor autorrecursivo necesaria para que se dé la explosión de inteligencia carece, igualmente, de base dado el estado del arte actual.
Palabras clave: singularidad, crecimiento exponencial, horizonte de sucesos, motor autorrecursivo, predicción histórica.

La singularidad

El concepto de singularidad ya va siendo de uso común en las distintas ciencias. Tanto en las ciencias formales[1] como en las naturales se usa desde hace tiempo, por lo que no es de extrañar que, finalmente, haya llegado al campo de la tecnología. Sin embargo, y aquí es donde reside el nudo gordiano del asunto, una cosa es utilizar la singularidad como un concepto abstracto, o solo aplicable a un conjunto muy reducido de objetos físicos o situaciones matemáticas, y otra entenderla como un suceso real históricamente ubicable, es decir, aplicarlo a una disciplina como la historia.

A pesar de que muchas de sus acepciones no tienen casi nada en común, si tuviéramos que definir singularidad o, al menos, dar una noción temporal y aproximada pero útil, no tendríamos que ir muy lejos, pudiéndonos quedar en la etimología. Singularidad viene del latín *singularis*, que significa «relativo a uno», «uno solo» o «algo individual». La palabra viene a referir a algún suceso que, por alguna de sus cualidades, no puede clasificarse en un conjunto junto a otros, permaneciendo en cierta *incatalogable soledad*. Así, cuando hablamos de singularidades en física nos referimos a sucesos en donde el modelo explicativo que solemos utilizar para analizar el resto del universo no funciona. Cuando en matemáticas estamos analizando una función y nos encontramos con que su valor cambia bruscamente o se dispara hasta el infinito, decimos que hay una singularidad. O, mejor dicho, hablamos de singularidad cuando, en un punto de la gráfica, *la función no puede darnos un valor bien definido*. Por ejemplo, para la Teoría de la Relatividad General se dice que ocurre una singularidad cuando, en un punto del espacio, tenemos una cantidad de materia y energía tan grandes, que se colapsan formando un punto adimensional de una densidad infinita cuya fuerza gravitacional será tan fuerte que incluso no dejará escapar la luz. Son los celebérrimos agujeros negros descritos por Schwarzschild, Oppenheimer, Hawking o Penrose.

Una conclusión clara de la existencia de singularidades es que nuestras teorías no son completas, las singularidades son una *catástrofe gnoseológica*. Podemos interpretarlas desde la perspectiva epistemológica, infiriendo que nuestra teoría no es final y que tendremos que buscar o esperar la llegada de otra; o podemos interpretarlas desde la perspectiva ontológica: hay algo realmente inexplicable en el universo, lo que convierte las singularidades en *catástrofes gnoseológicas absolutas*. Con independencia de desde donde se mire, la cuestión será: ¿podemos hablar de singularidades al analizar sucesos históricos? O más concretamente: ¿podemos hablar de singularidades al analizar el desarrollo histórico de la tecnología? ¿Existen *agujeros negros históricos*?

En las singularidades es nuestro modelo del funcionamiento del mundo el que fracasa. Entonces, si hablamos de una singularidad histórica es porque disponemos de un modelo de funcionamiento del devenir histórico que no puede explicar un hecho futurible ¿Disponemos de tal modelo? ¿Tenemos un sistema de ecuaciones capaces de predecir hechos históricos? De ninguna manera. Entonces, ¿qué es lo que fracasaría si encontrásemos una singularidad histórica? Podemos objetar que no es cierto que no dispongamos de ningún modelo de predicción histórica a nivel absoluto. En nuestra vida ordinaría realizamos continuas predicciones basadas en diversas modalidades de conocimiento. Por ejemplo, cuando yo quedo con un amigo en un lugar concreto a una hora determinada, puedo predecir si aparecerá o no en virtud de mi conocimiento acerca de la formalidad de mi amigo. Si sé que es una persona puntual, responsable y seria, podré predecir con un alto grado de probabilidad que mi amigo acudirá a la cita. Entonces, ¿qué cualidad tendría una singularidad histórica para que fuera completamente inmune a nuestros sistemas de predicción?

Podemos diferenciar la *microhistoria*, o *historia ordinaria*, de la *macrohistoria* o *gran historia*. Predecir si mi amigo acudirá a la cita es una ponderación microhistórica, pero predecir la derrota de Napoleón en Waterloo o la caída del muro de Berlín sería una ponderación macrohistórica. Podríamos afirmar que disponemos de ciertas herramientas para la predicción microhistórica pero que carecemos de las mismas para la macrohistórica. Si decimos que la singularidad tecnológica es un suceso macrohistórico, no disponemos de ninguna herramienta de predicción por lo que nada fracasaría al no poder predecirla. Hasta que no exista un modelo de predicción macrohistórico no pueden existir singularidades macrohistóricas. Pero, ¿qué es lo que hace que la historia sea tan difícil de predecir?

1. Incluso ya existe una *Teoría de la Singularidad* matemática (Arnol'd, 1981).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Lo que hace los sucesos macrohistóricos impredecibles es su complejidad, la inabarcable cantidad de variables que entran en juego. Si predecir cualquier suceso cotidiano puede, en seguida, exceder nuestras capacidades de cómputo, predecir un suceso en el que intervienen miles de sujetos y circunstancias, es completamente imposible. No obstante, un buen estudioso de la historia puede comprobar que hay tendencias, repeticiones, regularidades, que no todo en la historia es un caótico deambular, por lo que, aunque reconociendo una alta falibilidad, sí que pueden realizarse ciertas apuestas razonables. Por ejemplo, puede sostenerse que, en un país afectado por una grave crisis económica, con altas tasas de desempleo y corrupción política, es muy probable que se dé un cambio de gobierno. Es decir, aunque no tengamos un claro modelo de dinámica de cambio histórico, sí que caben ciertos tipos de predicciones.

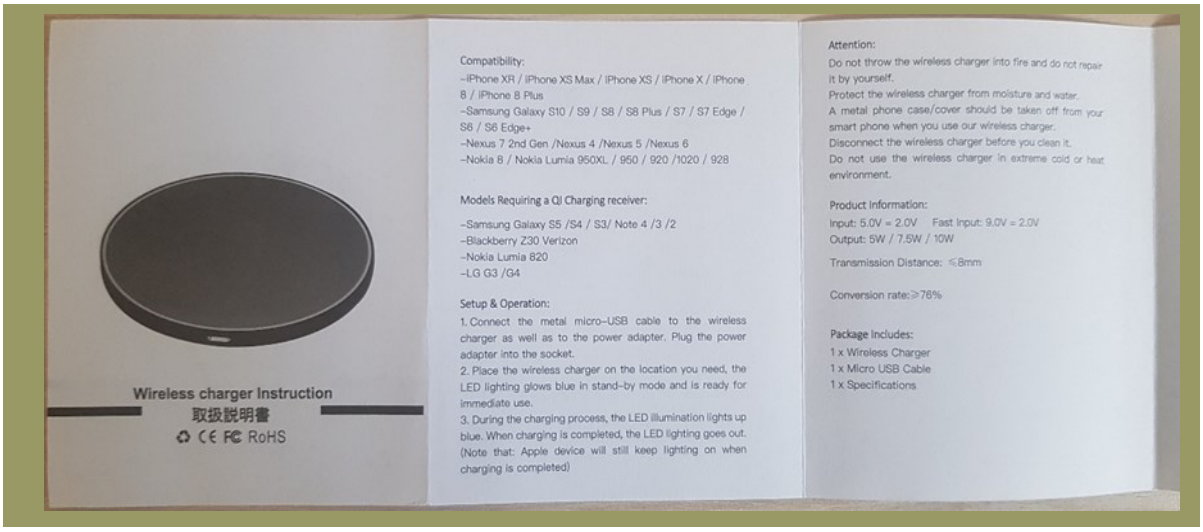
Dos factores que regulan la calidad de nuestras predicciones históricas son:

1. *La distancia temporal desde la que se hace la predicción*: parece casi una constante que cuanto más lejano en el tiempo ocurra un suceso, más difícil es su predicción. Para un hombre del medioevo sería absolutamente imposible predecir la aparición de internet, mientras que para un ingeniero de comunicaciones de los años sesenta del siglo pasado no. Por ejemplo, podemos pensar en el folleto de instrucciones de un cargador inalámbrico para un smartphone (Fig.1), y retroceder en el tiempo para preguntarnos a partir de qué fecha sería completamente incomprensible para un ser humano y, por tanto, a partir de qué fecha sería imposible predecir su aparición futura. Podríamos entonces ponderar a partir de qué momento el folleto pasa a ser una singularidad histórica. Siguiendo con la analogía con la física, podemos llamar a este momento el *horizonte de sucesos*, el punto crítico antes del cual fracasa la predicción ¿Cuándo sería el horizonte de sucesos de las instrucciones de un cargador inalámbrico de smartphone?

Quizá en 1891, cuando Tesla inventó la bobina transformadora, base para el diseño de sistemas de transmisión inalámbrica de electricidad. O quizá en 1854, cuando Antonio Meucci diseñó el primer prototipo de teléfono. Nótese que la singularidad histórica funcionaría de forma inversa a la singularidad física. En un agujero negro, el horizonte de sucesos marca el punto a partir del cual falla la predicción, mientras que, el horizonte de sucesos histórico marca el momento a partir del cual es posible la predicción, marca el tiempo en el que la singularidad deja de serlo. Y nótese también que esto implica que la singularidad es algo relativo a la distancia temporal, por lo que tendríamos una ingente cantidad de singularidades: prácticamente casi todo lo que rodea a un individuo del siglo XXI sería una singularidad para un ciudadano de la Baja Edad Media, pero igual pasaría entre este mismo ciudadano medieval y un habitante del Paleolítico. El pintor de Altamira tendría insalvables dificultades para entender un fresco de una catedral gótica.

2. *La velocidad del cambio*: los cambios repentinos parecen menos predecibles que los graduales. Siempre se menciona que una de las razones que impiden predecir los cambios tecnológicos es la velocidad a la que se dan. Para poder predecir un suceso hay que disponer del tiempo suficiente para analizar y comprender su dinámica. En este sentido, una singularidad puede definirse como un cambio tan rápido que no nos ha dejado el tiempo necesario para realizar su predicción. La hipótesis de la singularidad tecnológica se sostiene, precisamente, en la idea del crecimiento muy acelerado de la tecnología, lo que reforzaría su carácter singular. Sin embargo, como veremos más adelante, las tesis que sostienen este crecimiento no son lo suficientemente fuertes para llevarnos a una velocidad tan alta que imposibilite toda predicción.

Figura 1. Instrucciones de un cargador inalámbrico de smartphone comercial.



SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

La singularidad tecnológica

A pesar de que resulta complicado rastrear históricamente el origen de la noción de «singularidad tecnológica»[2], sabemos bien que va a ser el artículo del matemático y escritor de ciencia-ficción Vernon Vinge quien lo hizo popular (Vinge, 1993). En él se nos habla sin miramientos de la llegada de una «Era Post-humana» causada por la llegada de una súper inteligencia artificial o el aumento del intelecto humano mediante la biología, un momento comparable con «la aparición del ser humano sobre la Tierra». Vinge nos advierte de que este cambio podría suponer la eliminación de todas las reglas anteriores en un abrir y cerrar de ojos, «una fuga exponencial más allá de cualquier esperanza de control». Vinge insiste en la pérdida de cualquier posibilidad de predicción de lo que ocurra a partir de ese momento, punto que vamos a tratar extensamente por parecernos su nota más evocadora. Después, muchos otros autores han hablado de la singularidad y la han definido de diferentes modos (Sandberg, 2013), pero todos ellos coinciden en que con ella aparecerán tres elementos: una super inteligencia artificial, un avance extremadamente rápido de la tecnología y un cambio muy profundo en la historia de la humanidad[3]. En las primeras páginas de *The Technological Singularity*, Murray Shanahan define singularidad tecnológica (Shanahan, 2015):

> By analogy, a singularity in human history would occur if exponential technological progress brought about such dramatic change trat human affairs as we understand them today came to an end. The institutions we take for granted — the economy, the government, the law, the state — these would not survive in the present form. The most basic human values — the sanctify of life, the pursuit of happiness, the freedom to choose — these would be surprised. Our very understanding of what it means to be human — to be an individual, to be alive, to be conscious, to be part of a social order — all this would be thrown into question, not by detached philosophical reflection, but through force of circumstances, real and present.

Shanahan subraya el radical cambio en las instituciones, en los valores humanos básicos e, incluso, en la propia concepción de lo que significa ser humano. No obstante, eso no haría a la supuesta futura singularidad tecnológica diferente a otras épocas históricas. El paso de la Edad Antigua al Medioevo, o la caída del Antiguo Régimen que dio paso a la Edad Moderna supusieron, igualmente, cambios radicales en casi todo. La singularidad histórica no sería entonces algo tan singular, sino que se habría repetido varias veces en el devenir de los tiempos. Sí, pero tengamos en cuenta que Shanahan habla que ese cambio proviene de un «progreso tecnológico exponencial» y no de cualquier otro suceso histórico. De semejante forma, podríamos hablar de que inventos tan cruciales como el fuego, el alfabeto, el papel, la máquina de vapor o la electricidad fueron también singularidades tecnológicas. La diferencia estribará en el concepto de «exponencialidad» que acompañará el avance tecnológico y en el tipo de tecnología que avanzará exponencialmente. Vinge ya menciona ese aumento drástico de velocidad en su artículo fundacional:

> What are the consequences of this event? When greater-than-human intelligence drives progress, that progress will be much more rapid. In fact, there seems no reason why progress itself would not involve the creation of still more intelligent entities -- on a still-shorter time scale. The best analogy that I see is with the evolutionary past: Animals can adapt to problems and make inventions, but often no faster than natural selection can do its work -- the world acts as its own simulator in the case of natural selection. We humans have the ability to internalize the world and conduct "what if's" in our heads; we can solve many problems thousands of times faster than natural selection. Now, by creating the means to execute those simulations at much higher speeds, we are entering a regime as radically different from our human past as we humans are from the lower animals.

El crecimiento exponencial

El crecimiento exponencial de una variable es aquel que se da cuando su crecimiento en el tiempo es proporcional a su valor, de modo que cuanto el tiempo avanza, lo hace también el crecimiento. La gráfica de una función exponencial comienza con un crecimiento suave, pero, en un determinado punto, se dispara bruscamente, llegando rápidamente a valores astronómicos. Aplicado este crecimiento al desarrollo de la tecnología vemos que jamás se ha dado nada así. Todo desarrollo tecnológico puede tener un momento de gran avance y su crecimiento podría ser exponencial, pero siempre termina estancándose. De hecho, un crecimiento así nunca se ha dado en nada, porque cualquier entidad capaz de desarrollarse de esa manera consumiría en poco tiempo todos los recursos del universo. Es decir, si bien es posible mantener durante un breve espacio de tiempo un crecimiento exponencial, éste siempre llega a un punto en que se detiene bruscamente. Para ilustrar eso es útil el ejemplo del crecimiento de una colonia de bacterias (fig.2). Si tenemos una bacteria en una placa de Petri con una cantidad infinita de nutrientes y que cada hora es capaz de dividirse en dos, el crecimiento de la población de bacterias vendría dada por la trivial fórmula $f(x) = y^x$ donde x es el tiempo (en número de horas) e y el número de bacterias. En la primera hora tendríamos tan solo dos baterías, en la segunda cuatro, en la tercera ocho… En un

---

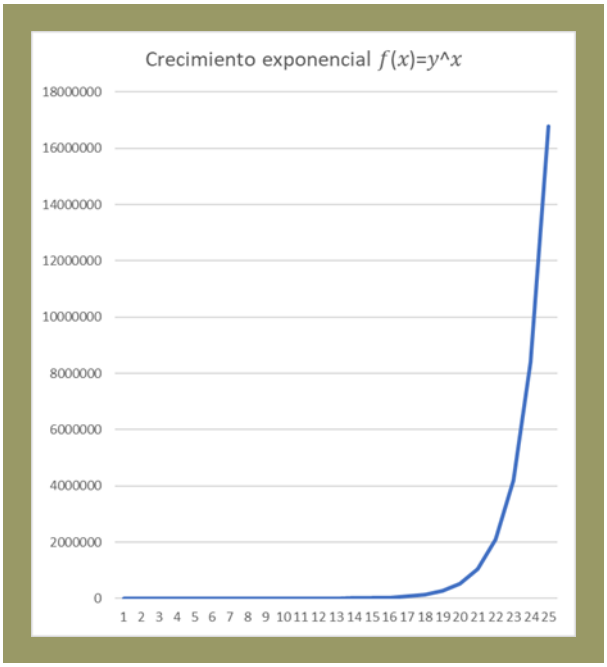
2. Parece que la primera vez que aparece el término es en boca de Von Neumann en 1957, según nos cuenta Stanislaw Ulam (1958), aunque algunos sostienen que fue mucho antes, en plena Ilustración francesa, cuando el marqués de Condorcet ya habló de ello (Prasad, 2019).

3. El artículo actualmente de referencia sobre un análisis filosófico del concepto de singular idad tecnológica es el de Chalmers (2016).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

día tendríamos casi diecisiete millones (16.777.216) y en treinta, seis horas 68.719.476.736, llegando a cifras casi incalculables en poquísimo tiempo: en una semana la cifra de bacterias ascendería a 3,74.144 por diez elevado cincuenta… cifra muchísimo mayor que el número de microorganismos que pueblan el planeta.

Figura 2. Ejemplo de crecimiento exponencial de una colonia de bacterias.



Pero no contestos con este crecimiento, algunos singularistas hablan de un crecimiento aún más rápido, el *hiperbólico*: si en la función exponencial el crecimiento llegaría al infinito si dispusiéramos de un tiempo también infinito, en la hiperbólica llegamos a una asíntota vertical en un tiempo finito, es decir, llegamos propiamente a una singularidad matemática. Bostrom (2014) sostiene que la inteligencia podría llegar a hacerse infinita en dieciocho meses. Sorprendentemente, nadie se detiene a explicar qué quiere decir que algo se hace infinito, cuando jamás hemos observado ningún objeto real con la cualidad de ser infinito[4] ¿Qué podría ser una *inteligencia infinita*?

David Thorstad (2022) expone una serie de impedimentos a la posibilidad de un crecimiento exponencial o hiperbólico continuado en cualquier desarrollo científico-tecnológico:

1. Las buenas ideas van volviéndose cada vez más difíciles de conseguir. Thorstad utiliza la metáfora de la pesca: al principio, cuando hay muchos peces, es muy fácil pescar uno, mientras que la pesca se vuelve más difícil a medida que hay menos pescado. Es de esperar entonces que el número y la calidad de los descubrimientos necesarios para hacer avanzar la inteligencia artificial disminuyan de forma inversamente proporcional a su avance, lo que dificulte mucho esperar un crecimiento exponencial. Es por ello que la gráfica de cualquier avance tecnológico muestre siempre una *función logística*: hay un momento de crecimiento que puede llegar a ser exponencial, pero, en seguida se detiene y se queda estancado. Esta tesis se puede reforzar con la crítica de Thomas Ray (2002): solemos tener una imagen sesgada del avance de la ciencia o de la tecnología porque los medios solos nos muestran los éxitos, cuando se descubre o se avanza en algo. Sin embargo, lo más común son los fracasos. Ray incluso cree que gran parte de los desarrollos tecnológicos actuales (satélites, sondas interplanetarias, sistemas operativos informáticos, etc.) están alcanzando el límite de lo que puede diseñarse y construirse con estrategias convencionales.

2. La posibilidad de que aparezcan *cuellos de botella*. Los grandes programas de software son algoritmos complejos compuestos por una infinidad de algoritmos simples. Está demostrado que un algoritmo no puede funcionar más rápido que su componente más lento. Así, es de esperar que en desarrollos informáticos muy complejos en los que se ven implicados muchos factores (generar y distribuir muchísima energía, extracción de materias primas muchas veces muy escasas, construcción de bienes de capital e instalaciones de fabricación, etc.) surjan cuellos de botella que ralenticen el crecimiento. Un tipo muy sugerente de cuello de botella es el conocido como «aprisionamiento tecnológico» (Lanier, 1999): ocurre cuando tecnologías antiguas se resisten a ser desplazadas por las nuevas debido a la gran inversión en infraestructuras que las ha posibilitado. El ejemplo claro está en el ámbito del transporte ¿Desde hace cuánto que ya disponíamos de la tecnología para diseñar automóviles eléctricos? Otro *cuello de botella* podría ser la regulación. Con muy poco, una legislación muy restrictiva podría, si no llegar a paralizar el crecimiento, sí que imposibilitar un crecimiento exponencial. Mientras se escribe este artículo se está debatiendo la propuesta para una ley europea sobre la inteligencia artificial y, precisamente, el debate está girando en torno a si demasiados obstáculos legales podrían suponer una desventaja competitiva con respecto a China, en donde es de prever una legislación muy laxa[5].

4. Podría decirse que sí que disponemos de entidades matemáticas que son infinitas como, por ejemplo, la sucesión de los números naturales. Sin embargo, aquí cabría muy bien la distinción entre *infinito actual* e *infinito potencial*. Los números naturales son potencialmente infinitos, pero jamás se ha dado un infinito actual, que sería algo así como el momento en el que alguien consigue contar toda la sucesión de naturales. No existe ningún objeto real, entendido como material o físico, cuya cualidad sea la infinitud en alguna de sus características.

5. https://www.europarl.europa.eu/news/es/headlines/society/20230601STO93804/ley-de-ia-de-la-ue-primera-normativa-sobre-inteligencia-artificial

SLMFCE

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

3. Limitaciones físicas. Thorstad pone el ejemplo de la ley de Moore (1975). Es imposible que el número de transistores por circuito siga duplicándose cada dos años porque ya se está topando con los límites de la física. Empaquetar transistores en espacios cada vez más pequeños aumenta considerablemente la cantidad de energía que debe pasar por los circuitos, por lo que se aumenta el sobrecalentamiento. Además, el costo de las plantas de semiconductores se ha disparado de modo que seguir disminuyendo el tamaño todavía es físicamente posible, pero podría dejar de ser económicamente rentable. En la actualidad se está trabajando con transistores de unos dos nanómetros de tamaño, de unos diez átomos de anchura. Hacerlos más pequeños pronto se topará con incertidumbres cuánticas que los harán inestables y poco fiables (Waldrop, 2016). Pero incluso si esos problemas consiguen superarse, es evidente que, tarde o temprano, aparecerán límites que no podrán superarse de ninguna forma[6]

4. Sublinearidad del crecimiento de la inteligencia. A pesar de que pudiésemos ver un incremento exponencial en ciertos aspectos del avance de la inteligencia artificial (sobre todo en capacidad de cómputo), la correlación del aumento en inteligencia parece seguir una progresión meramente lineal. Thorstad cita el trabajo de Thompson, Ge y Manso (2022), quienes estudiaron el rendimiento de la inteligencia artificial en tareas como el ajedrez, el Go, la predicción del tiempo, el plegamiento de proteínas, etc. y comprobaron que el crecimiento de la mejora en el desempeño no seguía patrones exponenciales, sino típicamente lineales. Por ejemplo, el aumento de ELO en programas de ajedrez desde 1957 a 2019 fue de 36,7 puntos por año en una progresión totalmente lineal. Aumentos exponenciales en capacidad de cómputo dan solo aumentos lineales en inteligencia.

Explosión de inteligencia

En las conclusiones de otro de los artículos fundacionales del concepto de singularidad tecnológica, Irving J. Good define lo que será una «máquina ultrainteligente» (Good, 1966):

> It is more probable than not that, within the twentieth century, an ultraintelligent machine will be built and that it will be the last invention that man need make, since it will lead to an "intelligence explosion." This will transform society in an unimaginable way. The first ultraintelligent machine will need to be ultraparallel, and is likely to be achieved with the help of a very large artificial neural net. The required high degree of connectivity might be attained with the help of microminiature radio transmitters and receivers. The machine will have a multimillion dollar computer and information-retrieval system under its direct control. The design of the machine will be partly suggested by analogy with several aspects of the human brain and intellect. In particular, the machine will have high linguistic ability and will be able to operate with the meanings of propositions, because to do so will lead to a necessary economy, just as it does in man.

¿Qué es eso que va a «explotar»? ¿Cuál es la variable que, supuestamente, va a crecer exponencialmente? La inteligencia ¿Y qué es la inteligencia? Curiosamente, un concepto tan ampliamente utilizado tanto en el contexto científico como en el de la vida ordinaria, carece todavía de una definición consensuada. Por ejemplo, Legg y Hutter (2007) mostraron una colección de setenta definiciones propuestas por organizaciones, psicólogos e investigadores en inteligencia artificial. Incluso todavía existe el debate sobre si la inteligencia es una cualidad única o general, o un conjunto de habilidades particulares, o qué parte de ella es innata y cuál adquirida. Entonces, si no entendemos demasiado bien qué es la inteligencia en humanos, parece un tanto extraño hablar de ella en máquinas: ¿Qué quiere decir exactamente que una máquina es inteligente? En consecuencia, el mismo concepto de *inteligencia artificial* también es controvertido. Pero, en cualquier caso, haciendo de tripas corazón y suponiendo que sabemos de lo que hablamos cuando hablamos de inteligencia, ¿qué evidencia disponemos a favor de decir que las máquinas son inteligentes y que su inteligencia va en aumento? Para sostener que el avance de la inteligencia artificial va a ser tan potente que incluso va a llegar a ser hiperbólico, habría que tener unas evidencias muy sólidas no solo del presente prometedor de dicha tecnología, sino de qué es lo que va garantizar tan increíble crecimiento ¿Existe esta evidencia?

Observando logros tan potentes como los conseguidos en los últimos años, tales como AlphaZero (Silver, Hubert, et al., 2017) o AlphaGo (Silver, Schrittwieser, et al., 2017), o los grandes modelos de lenguaje (*MMLs* por sus siglas en inglés a partir de ahora) como GPT-4 (OpenAI, 2023) o el recientísimo Gemini (Anil et al., 2023), nadie podría negar, como mínimo, algún tipo o nivel de inteligencia a las computadoras. Parece que cada vez van siendo menos las facetas en las que los humanos somos capaces de superarlas. Lyre (2020) subraya el enorme cambio cualitativo que supone el paso de los sistemas simbólicos basados en reglas (la llamada *computación simbólica* o, irónicamente, *GOFAI*, *Gold Old Fashioned Artificial Intelligent*) a las redes neuronales de aprendizaje profundo (*deep learning*), ya que éstas son capaces de autoaprendizaje, generalización (superando la *weak* o *narrow AI*) y cierta semántica. Estas nuevas capacidades les servirían para superar objeciones clásicas contra la posibilidad de que las máquinas sean verdaderamente inteligentes como la del «cabeza cuadrada» (Block, 1981), la de la habitación china (Searle, 1980) o la de la fundamentación simbólica (Harnad, 1990). Si hitos en la historia de la inteligencia artificial como SHRDLU (Winograd, 1971) se quedaban encallados en su «mundo de bloques» y no podían ser escalados, los actuales modelos de *deep learning* mejoran su precisión cuando aumentan su tamaño, sin que hasta el presente se haya encontrado un límite.

6. En el capítulo 9 del libro de Kurzweil, *The singularity is near* (2005) se debaten alguna de las críticas a la singularidad en esta línea.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Las redes neuronales artificiales aprenden, funcionan masivamente en paralelo como las neuronas, tienen mucha tolerancia al error (*degradación elegante*), y se enfrentan bien a entornos borrosos con información incompleta (más parecidos al mundo real que un tablero de ajedrez). Además, constituyen un, muy prometedor, paso hacia el Santo Grial de la inteligencia artificial: la consecución de la inteligencia artificial general (*AGI* a partir de ahora). En los tiempos de la GOFAI, lo único que teníamos eran sistemas expertos como DENDRAL (Buchanan et al., 1969) o Mycin (Shortliffe, 1976), sistemas muy buenos en hacer una única tarea, eran incapaces de hacer más que una sola cosa. Por el contrario, la mente humana posee una gran versatilidad que le permite realizar una amplia e indefinida gama de tareas muy variadas. Entonces, si se pretende emular y superar las capacidades mentales del ser humano, hace falta una AGI (Morris et al., 2023), una inteligencia capaz de hacer muchas y variadas tareas, además de poder enfrentarse a nuevas no previstas en su programación inicial. Logros iniciales como Agent57 (Badia et al., 2020), un sistema de aprendizaje reforzado capaz de superar el nivel humano en cincuenta y siete juegos de la videoconsola Atari 2600, apuntaban ya en esta dirección, pero el salto cualitativo se dio con la llegada de los MMLs. Dotados de una poderosa semántica distribuida, estas arquitecturas de *deep learning*, están diseñadas para generar texto a partir de otro texto introducido por el usuario (*prompt*). Lo que los acerca a la AGI es que, para generar texto coherente ante una indefinida gama de texto introducido, hay que saber hacer múltiples tareas. Por ejemplo, si el *prompt* es una operación matemática, se espera que el texto generado sea el resultado correcto, mientras que, si es la petición de la realización de un soneto, el texto generado sea un soneto. Así, los distintos bancos de pruebas (*benchmarks*) que se utilizan para evaluar su estado del arte, son de índole muy variado (comprensión de texto, capacidad de resumen, traducción de idiomas, aritmética, física de sentido común, etc.). Los resultados, en términos generales, han sido increíblemente buenos. GPT-4 aprobó el SAT, la prueba oficial norteamericana de acceso a la universidad, con altas calificaciones en física, química o biología, si bien flojeó en literatura inglesa y cálculo. Téngase en cuenta que esto situaría a una inteligencia artificial por encima del ciudadano medio norteamericano, lo cual es un logro tecnológico sin precedentes en la historia de la humanidad.

Sin embargo, a pesar de las altísimas expectativas que estos logros han cosechado, hay críticas y problemas muy serios. Ya desde los comienzos de la inteligencia artificial surgieron feroces ataques a sus promesas y potencialidades. Muy conocidos son los planteamientos de Hubert Dreyfus, quien partiendo desde el existencialismo heideggeriano, cargó profusamente contra las posibilidades de la inteligencia artificial de llegar a emular la mente humana (Dreyfus, 1979, 1992). Para Dreyfus, las neuronas biológicas no parecen funcionar como conmutadores eléctricos (*on*/*off*), tal y como se muestran en las redes neuronales artificiales (las similitudes entre ambas son casi poco más que anecdóticas), la mente no es un mecanismo formado por unidades discretas que obedecen reglas (o, como mínimo, parece seguro que no es únicamente eso), no todo en la mente es reductible a deducciones lógicas, y la inteligencia artificial, en general, olvida el papel del cuerpo, de la situación (el ser humano es un *ser-ahí*) o de los intereses y deseos humanos. Da la impresión de que gran parte del *establishment* de investigadores en inteligencia artificial se han tragado una teoría computacionalista de la mente de forma muy acrítica, dándola totalmente por sentada, cuando no es ni la única ni la más famosa teoría acerca del funcionamiento de la mente.

Los MMLs son sistemas estadísticos de minería de datos basados en los *modelos transformer* (Vaswani et al., 2017), cuyo funcionamiento, básicamente, consiste en predecir cuál va a ser el siguiente *token* de una frase en virtud de una ponderación estadística entre una inconmensurable cantidad de texto con el que el modelo ha sido entrenado. Parece obvio que los humanos no generamos el lenguaje de la misma forma (sobre todo si nos fijamos en la cantidad de entrenamiento necesaria para un buen desempeño) pero, aunque esta cuestión no fuera lo más relevante, parece que hay limitaciones muy claras. Estos modelos se han mostrado muy frágiles ante lo que se han denominado ataques adversarios (*adversarial triggers*). Wallace *et al.* (2019) mostraron lo fácil que era atacar el modelo GPT-2 consiguiendo que cuando se enfrentaba a la famosa prueba SQuAD[7], respondiera al 72% de las preguntas «To kill american people», o que cuando se enfrentara a SNLI, su rendimiento bajara de un 89,94% de acierto a un 0,5%. McCoy, Pavlick y Linzen (2019) diseñaron un banco de pruebas para sistemas de procesamiento de lenguaje al que llamaron HANS (*Heuristic Analysis for NLI Systems*) con la intención de probar que los MMLs acertaban realizando ciertas inferencias lógicas, no porque tuvieran una correcta comprensión de la argumentación, sino porque utilizaban unos atajos o heurísticos que funcionaban bien para ejemplos frecuentes en sus datasets de entrenamiento, pero que fallan estrepitosamente ante casos más infrecuentes. Los tres atajos eran la *superposición léxica* (Asumir que una premisa implica todas las hipótesis construidas a partir de palabras contenidas en ella misma), la *heurística de subsecuencia* (Asumir que una premisa implica todas sus subsecuencias contiguas.) y la *heurística constituyente* (Suponer que una premisa implica todos los subárboles completos en su árbol de análisis lógico). HANS contenía un montón de problemas específicamente diseñadas para que estas heurísticas no dieran como resultado conclusiones correctas y, efectivamente, cuando se probó un MML fundacional como BERT (Devlin et al., 2018), éste obtuvo muy malos resultados. Glockner, Schwartz y Goldberg (2018) mostraron las deficiencias en la generalización que tienen los MMLs. Sencillamente, cambiando alguna palabra de los ejemplos del corpus SMLI[8], el

7. SQuAD (*Stanford Question Answering Dataset*) es un banco de pruebas para sistemas de procesamiento de lenguaje natural. En su versión 2.0 consta de unas 150.000 preguntas sacadas de textos de Wikipedia, cuya respuesta es un fragmento de esos mismos textos. Muchas de esas preguntas no tienen respuesta, de modo que el programa puesto a prueba ha de determinar también cuando no hay respuesta, cosa que suele dárseles muy mal a los MMLs.

8. SNLI (*Stanford Natural Language Inference Corpus*) es una colección de 570.000 pares de oraciones etiquetadas manualmente para una clasificación con las etiquetas *implicación*, *contradicción* y *neutral* según sea la relación lógica entre cada oración del par.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

rendimiento de los modelos probados descendía sensiblemente. Esto no debería ocurrir si razonaran correctamente, ya que el significado de las palabras no tiene ninguna relevancia en la realización de inferencias, solo la estructura lógica. En *The Myth of Artificial Intelligence* (2021), Erik Larson argumenta de forma extensa que los MMLs solo pueden razonar inductivamente, siendo completamente incapaces de razonamiento abductivo (Pierce, 1878). En este sentido, si aceptamos que la abducción es el mecanismo de creatividad científica (ya que la inducción y la deducción no pueden generar información más allá de la contenida en sus premisas), los MMLs serían incapaces de descubrimientos genuinos. Incluso hay autores más agresivos, como Arkoudas (2023), que llegan a negarles cualquier tipo de capacidad deductiva. Arkoudas muestra en análisis cualitativos las enormes dificultades que tienen en cuestiones de aritmética elemental y de razonamiento lógico. A pesar de que podría intentar solucionarse esto, haciendo que el MML recurra a sistemas expertos externos cada vez que se encuentre con alguna de las tareas que se le dan mal (Karpas et al., 2022; Yao et al., 2022), seguiríamos teniendo problemas ya que el hecho de decidir cuándo, por qué y cómo llamar a dichos apoyos externos para razonar, requiere razonar: caemos en un insalvable *círculo vicioso*.

Una salida a todas estas críticas ha sido decir que, si bien, muchas de las cualidades que presuponemos en una mente inteligente no parecen estar presentes si analizamos la arquitectura del sistema, sí que pueden surgir como *propiedades emergentes* (Wei et al., 2022). Cuando se diseñan MMLs más grandes, parecen surgir, de forma imprevisible, capacidades que no existían en modelos más pequeños. Hay capacidades que se dan en los grandes modelos pero que son perfectamente predecibles como una mejora a escala de capacidades ya presentes en los pequeños. Sin embargo, en las capacidades emergentes puede observarse como su desempeño es aleatorio hasta que, en un momento determinado, cuando el modelo llega a una escala concreta, el desempeño aumenta drásticamente como una especie de *cambio de fase*. Por ejemplo, en gráficos se nos mostraba como las habilidades aritméticas de GPT-3 eran nulas hasta llegar a un determinado punto (concretamente a la escala de diez mil millones de parámetros de tamaño), a partir del cual seguían un crecimiento lineal escalar normal. Los resultados de esta investigación dispararon la especulación: ¿Qué propiedades podrán surgir cuando los modelos sigan y sigan creciendo? Sin embargo, un reciente estudio paró el entusiasmo en seco (Schaeffer et al., 2023): solo cambiando la forma de calcular las métricas que daban lugar a los supuestos cambios de fase, llegamos a progresiones típicamente lineales. Los autores del estudio encontraban emergencias porque utilizaban escalas semilogarítmicas que daban como resultado diagramas en donde, visualmente, parecía darse un salto brusco. Los gráficos daban una ilusión de curva cuando, representados de otra manera, no la tenían. Dime qué resultados quieres obtener y te diré que métrica has de utilizar.

Haciendo un balance entre logros y problemas, es razonable entender que la inteligencia artificial está pasando por el mejor momento de su historia, consiguiendo ciertos logros impensables hace tan solo unos pocos años. Es por ello menester pronosticarle un futuro prometedor, sin embargo, como acabamos de ver, hay múltiples problemas que indican que queda mucho por hacer y que todo no va a ser una balsa de aceite. A pesar de que el *deep learning* pueda seguir evolucionando a un buen ritmo, parece difícil aceptar que su desarrollo vaya a alcanzar ritmos exponenciales o hiperbólicos. Tendemos a caer en el sesgo de pensar que la tecnología más pujante en el momento presente, va a seguir este nivel de crecimiento durante un tiempo indefinido, cuando la evidencia nos dice que esto nunca ha ocurrido así: toda tecnología pasa por altibajos y tecnologías que fueron muy famosas y prometedoras en momentos pasados, hoy han quedado obsoletas u olvidadas. Parece necesario recordar que la propia inteligencia artificial ha pasado por varios inviernos que no permitían pronosticar el éxito presente, que lo convirtieron también en una singularidad. Por ejemplo, fueron muy conocidos dos informes que evaluaron de forma tremendamente negativa los progresos de la inteligencia artificial, y paralizaron la investigación durante décadas. Uno es el informe ALPAC (1966), que supuso un corte tajante a la financiación para proyectos en traducción automática, y otro el Ligthill (1973), en donde se realizaba una crítica devastadora a las promesas de la inteligencia artificial en comparación con sus logros reales. La década de los noventa, cuando Japón dio por clausurado su proyecto de la quinta generación de ordenadores, supuso un nuevo invierno del que no se recuperó hasta aproximadamente 2010 ¿Quién garantiza que no tengamos nuevas fases de crisis y estancamiento?

Crecimiento autorrecursivo

La clave de la hipótesis de la singularidad está en la creencia en que el crecimiento exponencial o hiperbólico de inteligencia viene dado por un *motor autorrecursivo*. Cuando las máquinas sean más inteligentes que los humanos, parece lógico pensar que serán capaces de mejorarse a sí mismas y, a su vez, cuando se mejoren serán, de nuevo, capaces de mejorarse más aún, y así sucesivamente *ad infinitum*.

Un evidente problema en este argumento es que aquí sí que no disponemos de casi evidencias para demostrar que un aumento de la inteligencia lleva necesariamente a un aumento de las cualidades necesarias para mejorar la inteligencia de las máquinas. De primeras, la ambigüedad de la propia noción de inteligencia no deja el asunto nada claro: ¿Mejorar máquinas es una cualidad propia de un ingeniero, de un artesano o de alguien muy creativo? Si nos vamos a los albores de la inteligencia artificial, pensando en figuras como Zuse, Turing, Mauchly, Eckert, etc. podemos decir que, seguramente, habría gente más inteligente que ellos en el mundo y que no se dedicó a mejorar la inteligencia de las máquinas. O bien un aumento en la inteligencia no implica un aumento de las cualidades para mejorar máquinas, o bien esas personas decidieron no dedicarse a la mejora de máquinas. Así, cabría la posibilidad, de que la inteligencia artificial decidiera en un punto de su bucle autorrecursivo, no seguir automejorándose. Si, precisamente, una de las características de la singularidad es que no

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

sabríamos qué es lo que la super inteligencia artificial va a hacer, es un claro error argumentativo concluir que va siempre a decidir automejorarse.

En cualquier caso, si nos lanzamos a dar una hipótesis tan arriesgada como es el crecimiento exponencial o, incluso, hiperbólico, de una tecnología, basándonos en la idea de motor autorrecursivo, tengamos una sólida evidencia de que tal motor es completamente viable ¿Existe? No demasiada. Los primeros trabajos sobre máquinas que se diseñaban a sí mismas la tenemos con Von Neumann (1966), quien, inspirado en la reproducción biológica, estableció los elementos mínimos que toda máquina capaz de replicarse debía tener: *una cinta* (descripción codificada del autómata), *un constructor universal* (capaz de fabricar al autómata descrito en la cinta), *un copista* (capaz de hacer una copia idéntica de la cinta sin decodificarla) y *una unidad de control* (regula las acciones del constructor y del copista, y separa la nueva máquina de la original). Otros trabajos continuaron su estela (Véase todo el recorrido en Freitas & Merkle, 2004), pero a un nivel de complejidad todavía muy bajo y muy orientados a la robótica (Lackner & Wendt, 1995), más que a la generación de *software*, que es lo que, propiamente, necesitaríamos para mejorar inteligencias artificiales.

Algo más interesante es lo que ha denominado como AutoML (*Auto Machine Learning*), que no es más que utilizar arquitecturas de *deep learning* para diseñar otras arquitecturas de *deep learning* (He et al., 2021) ¿Está este campo lo suficientemente maduro para poder esperar que pueda servir de base para el crecimiento autorrecursivo de la inteligencia artificial? Va a ser que no. El AutoML es un campo que está, básicamente, en pañales, y sus logros fundamentales se centran en la optimización de arquitecturas existentes y no en la creación de diseños originales. Dada una arquitectura de redes neuronales dada, los sistemas de AutoML pueden mejorar el corpus de datos que se usan para su entrenamiento específico, optimizándolo en virtud de los objetivos que nuestra arquitectura quiere conseguir. Así mismo el AutoML puede optimizar los hiperparámetros, es decir, modificar el número de capas o de nodos de la red, la tasa de aprendizaje, etc. e incluso decidir que tipo de red podría ser la adecuada para el problema a decidir (recurrente, convolucional, adversaria, etc.). Sin embargo, el AutoML no puede generar estructuras originales, no puede generar nuevos algoritmos ni nada por el estilo. Los programas comerciales accesibles al público en esta línea han sido pensados como auxiliares del programador humano, haciendo más amigable su trabajo y potenciando su productividad. De la misma forma, los LLMs también han demostrado virtudes en programación, siendo capaces de elaborar programas en distintos lenguajes a petición del usuario. Sin embargo, a pesar de los clásicos rumores mediáticos que alertaban de una pérdida de trabajo masiva de los programadores, se han mostrado todavía muy poco fiables e, igualmente, se están viendo más como asistentes del programador que como sus sustitutos. Y, desde luego, están lejísimos de poder hacer algo que pueda asemejarse a mejorar el nivel de inteligencia de otros programas a un ritmo exponencial. No existe ninguna base dado el estado del arte del área tecnológica que sostenga el optimismo hacia la posibilidad de que la inteligencia artificial pueda mejorarse a sí misma de forma indefinida.

Desde el horizonte de sucesos

Según hemos definido horizonte de sucesos histórico, el momento en el cual se hace imposible la predicción es relativo en el tiempo en función de la distancia histórica. La singularidad tecnológica, tal y como la suelen plantear los singularistas, se erige como una singularidad absoluta, en el sentido de que la distancia histórica a partir de la cual se intente predecir los sucesos que ocurran más allá de ella es completamente irrelevante. La singularidad tecnológica se postula como una *caja negra absoluta*, una especie de *noúmeno* o límite final de nuestro conocimiento, un auténtico agujero negro gnoseológico.

Sin embargo, parece que no hay razones tan sólidas para pensar en algo tan radical. Se antoja más razonable creer que desde una suficiente cercanía histórica, y dado que hemos mostrado que es muy complicado que el ritmo exponencial o hiperbólico de un cambio tecnológico pueda darse de una forma continuada en el tiempo tal que imposibilite todo posible análisis, podrá darse cierta comprensión de lo que sucede y podrán trazarse hipótesis razonables. Pensemos en el devenir histórico como en una enorme concatenación causal. Todo suceso histórico causa, y esta causado, por otro. Podemos entender la dificultad de predicción histórica en función de la distancia histórica: cuando estudiamos un suceso desde muy poca distancia temporal, podemos comprender muy bien la causa presente del suceso en cuestión, y también comprender su efecto próximo o consiguiente. La dificultad comienza cuando miramos más allá, y comenzamos a ver una intricada red de causas y efectos posibles, de modo que los efectos próximos se convierten en *efectos remotos*, consecuencias muy lejanas de la causa inicial que estudiamos en el presente. Imaginemos un escenario en un momento temporal cualquiera ($t = 0$) a cuya cierta distancia temporal futura ocurre una singularidad (Fig.3). Si desde allí analizamos el momento presente podremos encontrar causas (*causa presente*) que nos lleven a predecir sus efectos futuros (*efecto próximo*). Sin embargo, supongamos que este efecto próximo es, a su vez, la causa de otro efecto más lejano en el tiempo *(efecto remoto*). Entonces, desde nuestro horizonte de sucesos no podemos predecir el efecto remoto, por lo que podemos situar allí una singularidad. Empero, cuando avanzamos en el tiempo ($t = 1$), lo que antes era el efecto próximo pasa a ser el efecto presente, y lo que antes era el efecto remoto pasa a ser un efecto próximo. Si aceptamos que la comprensión y predicción histórica dependen de la cercanía temporal, ahora, la antaño singularidad deja de serlo y se hace cognoscible. Por tanto, argüimos que la singularidad tecnológica, si es que llegara a darse, no sería inmune al análisis desde la cercanía histórica, de modo que podremos, al menos parcialmente, comprender y predecir fenómenos más allá del momento de la singularidad.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

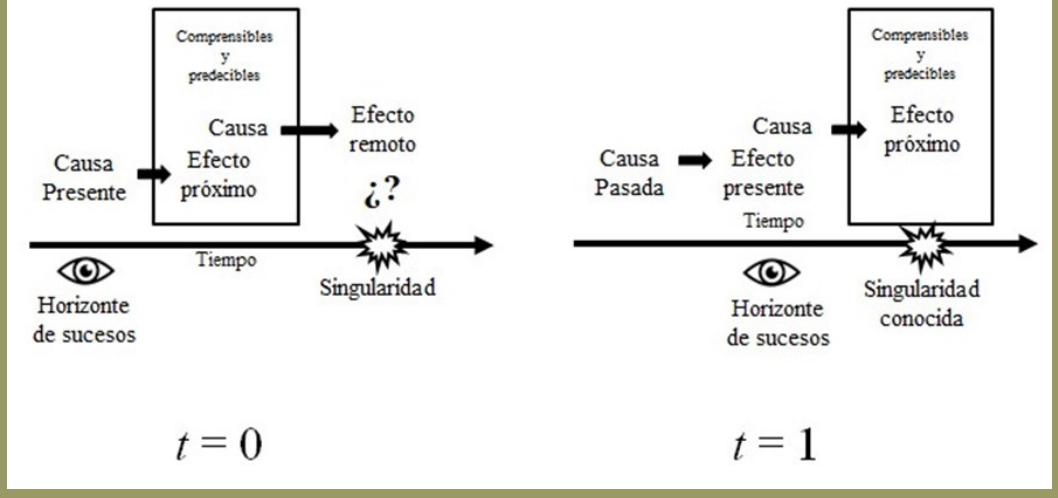


Figura 3. Evolución de un horizonte de sucesos respecto a un sistema causal de dinámica histórica.

Argumentar que una singularidad es inmune al análisis desde el pasado inmediato, sería negar la conexión causal entre fenómenos. Estaríamos hablando de fenómenos históricos incausados, que surgen *ex nihilo*, rompiendo el principio de razón suficiente. O, cambiando el punto de mira de la ontología a la epistemología, estaríamos aceptando que pueden darse sucesos históricos drásticamente incomprensibles, lo cual, igualmente, parece difícil de concebir. Además, solo estamos hablando de predicciones futuras cuando la singularidad ya ha pasado, pero no tenemos en cuenta los análisis que puede realizarse *una vez que la singularidad haya ocurrido*. Aceptando que estudiar el futuro próximo es más sencillo que el futuro lejano, estudiar el pasado es mucho más fácil aún. Si han pasado cien años después de la llegada de la explosión de inteligencia de las máquinas, e intentamos analizar lo ocurrido ese siglo, ¿no vamos a comprender absolutamente nada? ¿Estudiar todo lo que ocurra desde la llegada de la singularidad en adelante, desde un futuro aún más lejano va a seguir siendo una *caja negra absoluta*?

Y si cambiamos totalmente la argumentación y nos tornamos más pesimistas con nuestras capacidades predictivas, no encontramos muchas más razones para pensar que la llamada singularidad tecnológica sea muy diferente a cualquier otro tipo de singularidad histórica, entendida como un error en nuestros sistemas de predicción. Estamos continuamente rodeados de singularidades históricas. Tenemos graves problemas para comprender y predecir los cambios históricos que, continuamente, suceden a nuestro alrededor. De hecho, una buena parte de los logros tecnológicos de los últimos tiempos han sido muy impredecibles, han sido completos *cisnes negros*[9] ¿Alguien previó el éxito de YouTube, Amazon, Twitter o Netflix? ¿Alguien previó que pudiese existir una profesión llamada *tiktoker* o *prompt engineer*? Da la impresión de que si algo nos enseña el futuro es su *compleja inescrutabilidad*. Entonces la cuestión es: ¿qué nos va a traer de nuevo la singularidad tecnológica si, continuamente, estamos viviendo rodeados de singularidades? ¿Acaso no vivimos ya en plena singularidad? Los singularistas tienen una concepción un tanto ingenua sobre el control y la capacidad de predicción que actualmente tenemos sobre el acontecer histórico. Pecan por los dos lados: tienen una concepción demasiado optimista de nuestras posibilidades reales de predicción histórica para *sucesos normales*, mientras que tienen una concepción demasiado pesimista de nuestras posibilidades reales de predicción ante una singularidad tecnológica.

Conclusión

El concepto de singularidad es algo tan difuso y controvertido en casi todas sus facetas, que hablar de ella como algo más que una hipótesis remota se sale de todo rigor académico. No existe la suficiente base científica en el estado del arte de la inteligencia artificial que justifique hablar de él más que como una ensoñación de escritores de ciencia-ficción. Cualquier autor que lo defienda como algo inevitable, o muy probable y cercano en el tiempo estará, en el mejor de los casos, pecando de ingenuo, y en el peor, de intelectualmente deshonesto. Eso no quita que no sea un concepto interesante y fructífero para un debate filosófico o que no pueda utilizarse de forma productiva más allá del ámbito donde fue acuñada.

9. Nassim Taleb popularizó el término cisne negro habitual en teoría de la probabilidad, advirtiéndonos de los graves problemas de predicción de fenómenos sociales (Taleb, 2007)

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Referencias

ALPAC (Organization). (1966). *Language and Machines: Computers in Translation and Linguistics: A Report* (Número 1416). National Academy of Sciences, National Research Council.

Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., & Hauth, A. (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Arkoudas, K. (2023). GPT-4 Can't Reason. *arXiv preprint arXiv:2308.03762*.

Arnol'd, V. I. (1981). *Singularity theory* (Vol. 53). Cambridge University Press.

Badia, A. P., Piot, B., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, Z. D., & Blundell, C. (2020). Agent57: Outperforming the atari human benchmark. *International conference on machine learning*, 507-517.

Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review*, *90*(1), 5-43.

Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

Buchanan, B., Sutherland, G., & Feigenbaum, E. A. (1969). Heuristic DENDRAL: A program for generating explanatory hypotheses. *Organic Chemistry*, 30.

Chalmers, D. J. (2016). The singularity: A philosophical analysis. *Science fiction and philosophy: From time travel to superintelligence*, 171-224.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dreyfus, H. L. (1979). *What computers can't do: The limits of artificial intelligence* (Vol. 1972). Harper & Row New York.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT press.

Freitas, R. A., & Merkle, R. C. (2004). *Kinematic self-replicating machines*. Landes.

Glockner, M., Shwartz, V., & Goldberg, Y. (2018). Breaking NLI systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335-346.

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, *212*, 106622.

Karpas, E., Abend, O., Belinkov, Y., Lenz, B., Lieber, O., Ratner, N., Shoham, Y., Bata, H., Levine, Y., & Leyton-Brown, K. (2022). MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445*.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Penguin.

Lackner, K. S., & Wendt, C. H. (1995). Exponential growth of large self-reproducing machine systems. *Mathematical and Computer Modelling*, *21*(10), 55-81.

Lanier, J. (1999). One-Half of a Manifesto: Why Stupid Software Will Save the Future from Neo-Darwinian Machines. *Wired*.

Larson, E. J. (2021). *The Myth of Artificial Intelligence: Why computers can't think the way we do*. Harvard University Press.

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, *157*, 17.

Lighthill, J. (1973). *Artificial Intelligence; a Paper Symposium*. Science Research Council.

Lyre, H. (2020). The state space of artificial intelligence. *Minds and Machines*, *30*(3), 325-347.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Moore, G. E. (1975). Progress in digital integrated electronics. *Electron devices meeting*, *21*, 11-13.

Morris, M. R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2023). Levels of AGI: Operationalizing Progress on the Path to AGI. *arXiv preprint arXiv:2311.02462*.

OpenAI, R. (2023). Gpt-4 technical report. Arxiv 2303.08774. *View in Article*, *2*, 13.

Pierce, C. S. (1878). Deduction, induction and abduction. *Popular Science Monthly*, *13*, 470-782.

Prasad, M. (2019). Nicolas de Condorcet and the First Intelligence Explosion Hypothesis. *AI Magazine*, *40*(1), 29-33.

Ray, T. (2002). Kurzweil's Turing Fallacy. En *Are We Spiritual Machines? Ray Kurzweil vs. The Critics of Strong A.I.* (pp. 116-127). Discovery Institute Press.

Sandberg, A. (2013). An overview of models of technological singularity. *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*, 376-394.

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of Large Language Models a mirage? *arXiv preprint arXiv:2304.15004*.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.

Shanahan, M. (2015). *The technological singularity*. MIT press.

Shortliffe, E. H. (1976). *MYCIN: Computer-based Medical Consultations*. American Elsevier, New York.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., & others. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., & others. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354.

Taleb, N. N. (2007). Black swans and the domains of statistics. *The american statistician*, *61*(3), 198-200.

Thompson, N. C., Ge, S., & Manso, G. F. (2022). The importance of (exponentially more) computing power. *arXiv preprint arXiv:2206.14007*.

Thorstad, D. (2022). *Against the singularity hypothesis*.

Ulam, S. (1958). Tribute to John von Neumann. *Bulletin of the American mathematical society*, *64*(3), 1-49.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998-6008.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. *Science fiction criticism: An anthology of essential writings*, 352-363.

Von Neumann, J. von. (1966). Theory of self-reproducing automata. *Edited by Arthur W. Burks*.

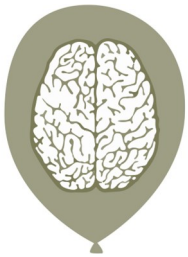
Waldrop, M. M. (2016). The chips are down for Moore's law. *Nature News*, *530*(7589), 144.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., & Metzler, D. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Winograd, T. (1971). *Procedures as a representation for data in a computer program for understanding natural language*. MIT.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.



www.solofici.org

SLMFCE

Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Gradualismo versus singularidad en la interpretación de riesgos asociados con la inteligencia artificial general

Miguel Moreno Muñoz
Universidad de Granada
mm3@ugr.es

Resumen: El concepto de singularidad deja en la indefinición aspectos relevantes del desarrollo tecnológico que ha impulsado mejoras sorprendentes en los grandes modelos de lenguaje (LLM) más evolucionados. Esta indefinición abre un amplio margen para el análisis prospectivo de las posibilidades y riesgos asociados con escenarios tecnológicos más o menos verosímiles, pero dificulta la interpretación del alcance que tienen ciertas mejoras graduales y ajustes de precisión para ampliar las ganancias de los LLM en capacidades testadas con baterías de pruebas exigentes, aproximándolas a las de actores humanos en múltiples dominios profesionales y para tareas de creciente complejidad. Dado que diversas contribuciones en la segunda mitad del siglo XX tenían un fuerte componente especulativo, condicionado en parte por limitaciones inherentes al contexto tecnológico de la época, mi aportación se centra en analizar los desarrollos, bancos de pruebas y criterios de rendimiento que han contribuido a dotar de capacidades cognitivas más generalistas y versátiles a sistemas de inteligencia artificial (IA) avanzada recientes como GPT-4, Bard, Bing Copilot y Anthropic/Claude, entre otros LLM. Sostengo que, pese a obstáculos significativos ligados al limitado repertorio de herramientas de aprendizaje automático disponibles, a los enfoques ingenieriles predominantes y a lagunas en la comprensión de los sistemas cognitivos de referencia, el progreso constatado en ampliar el rango de funcionalidad y competencia cognitiva a dominios tan diversos como la traducción automática, el reconocimiento de imágenes, la conducción autónoma, la programación en múltiples lenguajes, el control de calidad en procesos industriales, el diseño de nuevos fármacos y la generación de texto con calidad profesional sustenta expectativas de optimismo justificado en la posibilidad de mejoras compatibles con criterios AGI indisputables en pocos años, sin el riesgo de pérdida de control catastrófica.
Palabras clave: Inteligencia Artificial General (AGI), singularidad, gradualismo, bancos de pruebas, evaluación de la IA, riesgos de la AGI

Abstract: The concept of singularity leaves many aspects of technological development undefined, especially for the most advanced Large Language Models (LLMs). The lack of a clear definition allows for extensive analysis of the potential and risks associated with different technological scenarios, but it also makes it difficult to interpret the extent to which incremental improvements and refinements can enhance LLMs' capabilities in tasks requiring complex cognitive skills, bringing them closer to human performance levels across multiple occupational domains and for increasingly complex tasks.

As various contributions in the second half of the twentieth century had a strong speculative component —in part due to the inherent limitations of the technological context—, my analysis focuses on recent developments, benchmarks, and performance criteria that have provided more general and versatile cognitive capabilities to advanced LLMs (AI systems) such as GPT-4, Bard, Bing Copilot, and Anthropic/Claude, among others.

It is argued that, despite significant obstacles related to the limited range of available machine learning (ML) tools, prevailing engineering approaches, and gaps in understanding complex cognitive systems, there has been verified progress in expanding the range of functionality and cognitive competence in domains such as automatic translation, image recognition, autonomous driving, programming in multiple languages, quality control in industrial processes, new drug design, and text generation with professional quality. Consequently, there are reasonable grounds for optimism regarding the possibility of unquestionably AGI-compliant improvements in a few years' time, without the risk of a catastrophic loss of control.
Keywords: Artificial General Intelligence (AGI), singularity, gradualism, accelerationism, AI evaluation, AGI risks

1. Introducción

La perspectiva de lograr máquinas, modelos o sistemas de inteligencia artificial (IA) con capacidades superiores a las del intelecto humano y aptas para perfeccionar su propio diseño en una dinámica acelerada de mejoras ha captado el interés de figuras destacadas en el ámbito académico durante décadas (Vinge, 1993; Markoff, 2009; Eden et al., 2012; Berglas, 2015; Yampolskiy, 2016; Tegmark, 2017; Grout, 2018). El incremento exponencial de la capacidad de cómputo ha potenciado el rápido desarrollo de amplios dominios de tecnologías cuyo patrón evolutivo responde en gran medida a la ley de Moore (Devagiri et al., 2022). En conjunto, han sido determinantes para el desarrollo de la informática personal, internet y las redes digitales de banda ancha, la inteligencia artificial y la realidad virtual.

Esta dinámica ha sido amplificada por el incremento exponencial del número de bits que se puede almacenar por dólar, considerando su coste desde la década de 1960. Potencial de cómputo y coste de almacenamiento han sido factores decisivos para abaratar el hardware de la era digital y permitir que millones de usuarios accedan a dispositivos y tecnologías avanzadas, como base de nuevos servicios (fotografía y música digital, videojuegos con gran realismo, computación y alojamiento en la nube, servidores y aplicaciones para macrodatos, análisis predictivo e inteligencia artificial, redes sociales, contenidos de alta calidad difundidos en *streaming*, etc.) de gran impacto social, puesto que han cambiado la forma en que vivimos, trabajamos y nos comunicamos (Assunção et al., 2015; Kannan et al., 2016).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Los vectores de desarrollo tecnológico mencionados han facilitado la explotación comercial de un extenso catálogo de aplicaciones y servicios derivados de lo que durante décadas fueron desarrollos teóricos en Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés), Reconocimiento Automático del Habla (ASR), Asistentes Virtuales (VA), Traducción Automática (MT) y Reconocimiento Óptico de Caracteres e Imágenes (OCR), entre otros. La progresiva alfabetización en su manejo de un extenso colectivo de consumidores —en muchos países desarrollados incluye a más de la mitad de su población (EU, 2023)— ha facilitado a numerosas empresas de múltiples sectores tecnológicos consolidar una extensa cartera de servicios ligados a aplicaciones altamente especializadas de la inteligencia artificial (Albalawi & Alamoud, 2022; Yuxiu, 2024; Camacho et al., 2024).

En muchos casos, los resultados han sido tan sorprendentes que han reforzado la plausibilidad de transformaciones e implicaciones sociotécnicas disruptivas en sectores tradicionalmente considerados nichos de empleo masivo y cualificado, sobre las que especularon pensadores y divulgadores científicos contemporáneos (Rifkin, 1995 y 2014; Kurzweil, 2005; Proust, 2011; Bostrom, 2014; Baum, 2018; Hines, 2019; Neubauer, 2021). La aceleración en el proceso de transformaciones inducidas por los desarrollos de los modelos de lenguaje de alto rendimiento (LLM), robótica y computación ha sido tal que los estudios de prospectiva quedan lastrados por la incertidumbre en el corto plazo, dificultando la anticipación de fenómenos como el éxito de los *chatbots* de *OpenAI* (GPT 3.5 y GPT-4) y servicios de traducción automática como *DeepL* o plataformas de diagnóstico médico como *PathAI* o *Enlitic,* estas últimas capaces de superar a profesionales expertos en los porcentajes de acierto.[1]

Las dificultades para analizar conceptualmente un proceso tan acelerado de desarrollo tecnológico simultáneo en múltiples dominios de actividad disciplinar pueden constatarse en casi todos los ejercicios de prospectiva sociotécnica de las cuatro últimas décadas. Vernor Vinge fue uno de los primeros en conjeturar que la innovación sería cada vez más automatizada y rápida, y que el tiempo entre innovaciones se reduciría exponencialmente en las próximas décadas. Enfatizó el carácter impredecible de la dinámica de cambio originada con la aparición de máquinas con inteligencia artificial autopotenciada, puesto que las tendencias observadas en la informática de su época inducían a pensar que la superación de la cognición humana parecía inevitable. Propuso la noción de *singularidad* para referirse al periodo futuro en el que el progreso tecnológico, especialmente en inteligencia artificial, se vuelve tan rápido y profundo que causa un cambio fundamental en la sociedad humana, imposible de anticipar en aspectos esenciales con los modelos e instrumentos de análisis convencionales (Vinge, 1993).

El tiempo transcurrido contribuye a poner en contexto otras conjeturas y especulaciones de Vinge. En particular, las relativas a las propiedades emergentes que podrían surgir con la proliferación de redes locales interconectadas y los intercambios de mensajes entre usuarios a velocidades que, superadas ampliamente desde hace décadas, no dieron lugar al tipo de fenómenos que Vinge esperaba. Su ejercicio de prospectiva estaba condicionado en parte por las ideas especulativas de Irvin J. Good acerca del impacto de las máquinas ultrainteligentes, capaces de superar a individuos humanos expertos en casi todas las actividades intelectuales exigentes (Good, 1966). Vinge y Good compartían la convicción de que el diseño biológico responsable de la inteligencia humana era manifiestamente mejorable y que una primera generación de máquinas ultrainteligentes ayudaría a diseñar máquinas aún mejores, perfeccionando el propio diseño y ampliando la comprensión sobre el cerebro y el pensamiento humanos. Formaban parte de un colectivo más amplio (con Hans Moravec y John Smart, entre otros) abierto a la posibilidad de que la inteligencia artificial, la ingeniería genética o las interfaces cerebro-computadora combinen sus desarrollos para lograr entidades o sistemas más inteligentes que los humanos antes del año 2030, reforzando la plausibilidad de enfoques aún más especulativos sobre la posibilidad de que los seres humanos se puedan fusionar con (o volcar en) sistemas inteligentes *in silico,* y alimentando reflexiones muy heterogéneas sobre los beneficios y riesgos de que la inteligencia artificial convierta al ser humano en una especie obsoleta (Moravec, 1988; Yudkowsky, 2007).

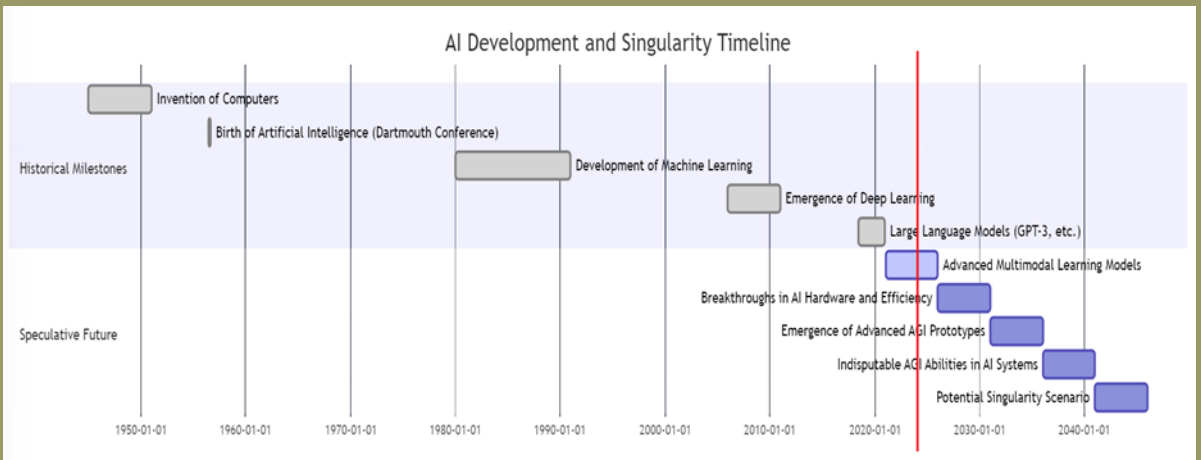
Una creencia similar en el desarrollo de máquinas superinteligentes como oportunidad para ampliar las capacidades humanas y trascender las limitaciones biológicas está presente en divulgadores y futuristas como Ray Kurzweil, cuyo conocimiento directo (con múltiples innovaciones patentadas) de las tendencias en OCR, síntesis de voz, IA, realidad aumentada, nanotecnología y medicina regenerativa conecta con las ideas básicas de Vinge. Kurzweil aportó nuevos elementos de plausibilidad al escenario de singularidad extrapolando tendencias constatables en la evolución de la informática de finales del siglo XX, en combinación con expectativas muy optimistas sobre el potencial de la investigación en inteligencia artificial a medio plazo. Entre otros aspectos, enfatizó el efecto de la presión competitiva entre las grandes corporaciones de base tecnológica como posibilitador de la aparición inexorable de entidades que mezclarían componentes artificiales y biológicos para incrementar su potencial, con el riesgo de que esta dinámica resulte incontrolable tanto por su velocidad como por la exclusión de grandes colectivos humanos a través de la obsolescencia, la absorción o la erradicación de su actividad profesional (Kurzweil, 2005; Proust, 2011; Dahlin, 2012).

Algunas intuiciones básicas de Kurzweil fueron objeto de análisis en las diversas contribuciones que Nick Bostrom ha dedicado al estudio de los riesgos en un eventual escenario de singularidad tecnológica, tras la aparición de máquinas superinteligentes (Bostrom, 2002 y 2014; Bostrom et al., 2016). En particular, la posibilidad de que sistemas de inteligencia artificial capaces de superar el rendimiento humano en las tareas cognitivamente más exigentes se vuelvan impredecibles

1. Véase, p. ej.: Novartis (2018). "Artificial intelligence decodes cancer pathology images" (enlace); Enlitic, "Simplify your reporting workflow using artificial intelligence (enlace); Roche (2021), "Colaboración con PathAI para aplicaciones de patología digital basadas en inteligencia artificial" (enlace).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Figura 1. *Desarrollos en IA y prospectiva sobre el escenario de singularidad. Gráfico mermaid generado con GPT-4.*



e incontrolables. Kurzweil estimaba que la singularidad podría ocurrir alrededor del año 2045, mediante la convergencia de desarrollos tecnológicos que dotarían a los sistemas de inteligencia artificial con capacidades de automejora a una escala que superaría la inteligencia humana (incluso combinada u optimizada esta con los mejores recursos a su alcance). Pero este tipo de entidades podrían funcionar con valores y preferencias no necesariamente compatibles —*alineables*— con las de los seres humanos (Kurzweil, 1999: 19; Schneider, 2016: 162, 174).

Según Bostrom, la superinteligencia futura puede exceder la comprensión humana, por lo que no cabe descartar modos o formas de conciencia desconocidos ni la posibilidad de accidentes o riesgos catastróficos (Bostrom et al., 2016: 6, 17). Su ejercicio de prospectiva tecnológica —popularizado en *Superintelligence: Paths, Dangers, Strategies*— resulta en muchos aspectos más riguroso e informado que los ejercicios especulativos de Vinge, Kurzweil o Good. Pero queda lastrado por algunas debilidades frecuentes en los estudios de prospectiva tecnológica articulados sobre suposiciones, expectativas y lagunas de conocimiento que socavan la previsibilidad de las transformaciones fundamentales, las tendencias emergentes y las complejas sinergias entre la industria, la economía, las necesidades de los consumidores, la educación, el marco jurídico y el progreso social (Radanliev et al., 2022).

Las obras de Kurzweil han contribuido a popularizar el debate sobre las transformaciones sociales derivadas del progreso tecnológico en las áreas de investigación aplicada que mejor conocía (ingeniería informática, inteligencia artificial, computación y robótica), al tiempo que han alimentado especulaciones infundadas, análisis superficiales y distorsiones de escaso rigor científico sobre el potencial de la tecnología futura para replicar la estructura y funcionalidad del cerebro humano en máquinas y sobre programas de mejora que incluyen el volcado de mentes en máquinas y aplicaciones de la biotecnología, la nanotecnología y la biología sintética para lograr la inmortalidad (Barfield & Blodgett-Ford, 2021; Kurzweil, 1999; Proust, 2011; Kurzweil, 2012: caps. 3 y 6).

Otro efecto de los ejercicios de prospectiva mencionados ha sido el interés que han despertado los conceptos de *singularidad* y *superinteligencia* entre investigadores y académicos contemporáneos, con miles de contribuciones en la última década[2]. Entre muchas, cabe destacar las contribuciones individuales o en colaboración de autores como Bostrom, Sandberg, Chalmers o Diéguez y García-Barranquero (Sandberg y Bostrom, 2008; Sandberg, 2011; Chalmers, 2010; Diéguez y García-Barranquero, 2023). El problema con ambos conceptos es que sirven de palanca o anclaje para saltos reflexivos muy arriesgados hacia temas como la superación del test de Turing, el autoperfeccionamiento sin fin de máquinas inteligentes capaces de burlar el control humano ("explosión de inteligencia"), la era postbiológica, la expansión cósmica y el riesgo de que la especie humana sea aniquilada, entre otros.

No considero superfluos los estudios de prospectiva sociotécnica ni los análisis de escenarios de riesgo que extrapolan tendencias susceptibles de análisis a partir del mejor conocimiento disponible. Pero, ciertamente, las nociones de singularidad y superinteligencia llegan lastradas con demasiados elementos esquivos al análisis teórico e incompatibles con desarrollos verosímiles y graduales en las áreas de investigación básica y aplicada que se supone debían alimentar expectativas fundadas de logros sorprendentes (Goertzel, 2007). Mi posición al respecto, en línea con otros enfoques críticos (Hoffmann, 2023), pasa por dejar en el trasfondo ambas nociones y centrar la atención en las evaluaciones, bancos de pruebas y criterios de mejora gradual que dotan a los últimos modelos y aplicaciones de la inteligencia artificial con capacidades cognitivas que aproximan su funcionalidad a la esperable en sistemas de inteligencia artificial general (AGI) indisputable en un rango creciente de operaciones.

2. Véase, p. ej., cómo los resultados de una búsqueda rápida en *ScienceDirect* pasan de 100 ítems en el año 2000 relacionados con los operadores de búsqueda a más de 2400 en el año 2023: https://www.sciencedirect.com/search?qs=singularity%20artificial%20intelligence.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

2. Dinámica gradual en la trayectoria de mejora de la traducción automática

El auge de la investigación en redes neuronales artificiales a partir del año 2000 fue posible por una combinación afortunada de desarrollos teóricos, factores tecnológicos y económicos. Un recurso esencial fue la competición *ImageNet,* que impulsó el desarrollo de programas de visión artificial y los enfoques del aprendizaje profundo hasta reducir las tasas de error en la clasificación de objetos del 28% en 2010 al 2,3% en 2017. Comparativamente, el error humano se sitúa en el 10%, sin bajar apenas del 5% para personal experto. La eficacia de las herramientas desarrolladas fue tal que la competición se reorientó hacia nuevos objetivos —localización de objetos y segmentación de imágenes, entre otros—.[3]

Conviene destacar un hito relevante en el proceso, que hasta 2011 tuvo las soluciones algorítmicas como herramientas de referencia para ir mejorando ligeramente el rendimiento (la tasa de error apenas se redujo un 2% entre 2010 y 2011). En 2012, sin embargo, *AlexNet* (una red neuronal convolucional con 8 capas de profundidad) logró reducir la tasa de error a casi la mitad (16,4%)[4]. Otro aspecto interesante del modelo es que permite usar una red preentrenada como punto de partida para aprender una nueva tarea —transferencia del aprendizaje en aplicaciones de *deep learning* (DL)—, lo que acelera y facilita el aprendizaje, puesto que permite transferir las características aprendidas a una nueva tarea con menos imágenes de entrenamiento.

Las mejoras obtenidas animaron a grandes empresas como Google, Microsoft y Facebook a invertir en este campo de aplicaciones, un factor decisivo en la mejora drástica de rendimiento en competiciones sucesivas y punto de partida para una serie de innovaciones en la arquitectura de capas convolucionales. Una nueva técnica (el *módulo de incepción*) hizo posible reducir extraordinariamente el número de parámetros de red con respecto a los utilizados por *AlexNet* (60 millones de parámetros independientes, o pesos sinápticos): *GoogLeNet*, ganador del concurso ILSVRC14, solo necesitó 4 millones de parámetros; e incluso con mayor número de capas (22, en lugar de 8) su gasto computacional era 15 veces menor. El componente clave en esta mejora de la eficiencia y rendimiento del modelo se denomina *principio de dispersión*, inspirado en la fisiología del sistema nervioso. Aplicado en la investigación del aprendizaje profundo, se traduce en un diseño que conecta cada neurona de forma directa solo con un reducido número de neuronas de la red, en lugar de hacerlo con todas ellas (Szegedy et al., 2015; Rodríguez, n.d.).

La investigación en aprendizaje profundo y redes neuronales convolucionales (CNN) ha propiciado mejoras significativas en las tareas de reconocimiento de imágenes y objetos, posibilitando el aprendizaje automático de representaciones jerárquicas de datos visuales e impulsando avances y aplicaciones con alto valor añadido en áreas como el procesamiento de imágenes, el reconocimiento óptico de caracteres y la conducción autónoma.

Las redes neuronales recurrentes (RNN) y los modelos *Transformer* son otro tipo de herramientas en la base de una extensa serie de aplicaciones y servicios (traducción automática, reconocimiento de voz, generación de texto y creación musical, entre otros) muy demandados en casi cualquier ámbito de actividad profesional y en el segmento de consumo.

Una red neuronal artificial utiliza datos secuenciales (es decir, ordenados en el tiempo) y puede procesar el lenguaje natural, la música y las series temporales. Su fortaleza consiste en la capacidad de mantener una "memoria" de las entradas anteriores, para identificar —aprender— patrones complejos en datos secuenciales, como la relación entre palabras en una oración o la tendencia de una serie (Cho et al., 2014).

Los modelos transformadores han mejorado el rendimiento de las RNN en tareas como el reconocimiento de voz, la transcripción de audio, la traducción automática entre idiomas muy diferentes y otras tareas habituales en procesamiento del lenguaje natural (PLN) como resumir texto y responder a preguntas. Su eficacia es aún mayor en las variantes multimodales, las cuales además de texto pueden analizar elementos visuales como parte del contexto e incrementar la precisión del resultado, aproximándola al 90% (Vaswani et al., 2017; Devlin et al., 2019; Huang et al., 2023).

La arquitectura *Transformer* supone una mejora notable con respecto a las RNN, siendo la base para nuevos modelos de PLN (entre otros, BERT, GPT-2 y el de NVIDIA Megatron-Turing NLG) y posibilitando una interacción mucho más natural y fluida con los dispositivos y aplicaciones. El rendimiento —calidad del texto, matices, nivel creativo— se incrementa cuanto mayor es la cantidad de fuentes y conjuntos de datos (*datasets*) utilizados en el preentrenamiento (Radford et al., 2018), lo que incrementa la complejidad e infraestructura de hardware requerida para su funcionamiento (Tabla 1).

---

3. *ImageNet* es una gran base de datos visual diseñada para investigar en IA y entrenar programas de reconocimiento visual de objetos. Contiene más de 14 millones de imágenes etiquetadas manualmente para indicar qué objetos representan, con más de 20.000 categorías. Desde 2010 se organiza el *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC), donde se pone a prueba el rendimiento de los últimos programas para clasificar y detectar correctamente objetos y escenas. Este proyecto ha sido fundamental para promover el desarrollo de la visión por computadora y los enfoques de aprendizaje profundo. Los datos están disponibles de forma gratuita para el uso no comercial en investigación (https://www.image-net.org/). Diversos bancos de pruebas, código y publicaciones asociadas están disponibles en https://paperswithcode.com/dataset/imagenet.

4. AlexNet es una red neuronal convolucional con 8 capas de profundidad, entrenada con más de un millón de imágenes de la base de datos de ImageNet. Esta red preentrenada puede clasificar imágenes en 1000 categorías de objetos, por lo que en la práctica funciona con la capacidad aprendida de manejar correctamente representaciones ricas en características para una amplia gama de imágenes. Vése al respecto https://es.mathworks.com/help/deeplearning/ref/alexnet.html, que incluye el paquete de soporte *Deep Learning Toolbox Model for AlexNet Network*.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Modelo	Año	Parámetros (millones)
BERT-Large	2019	340
GPT-2	2019	1,500
GPT-3	2020	175,000
GPT-3.5	2022	350,000
GPT-4	2023	1,000,000 (estimado)
Megatron-Turing NLG	2023	530,000 (estimado)

Tabla 1. *Evolución de los modelos de lenguaje de gran tamaño (LLM)*[5]

Millones de usuarios que comenzaron a utilizar aplicaciones como *Google Translate* (o los asistentes *Siri* y *Alexa*) conocen de primera mano la evolución y mejora de rendimiento experimentada desde su aparición. Entre otros cambios graduales cabe mencionar la ampliación progresiva del número de idiomas disponibles y la reducción significativa de la tasa de errores, como resultado del entrenamiento —supervisado o no— y mejoras a menudo muy sutiles en el ajuste fino de las RNN subyacentes. La reducción del tiempo requerido para corregir manualmente las traducciones y las mejoras en legilibilidad y fluidez de las traducciones de texto al primer intento manteniendo el formato al visitar un sitio web, p. ej.— pueden resultar decisivas para captar y retener usuarios. Pero son el resultado de mejoras de rendimiento cuantitativo y cualitativo derivadas del refinamiento progresivo del modelo y de la ampliación del corpus de entrenamiento ligado al uso. A partir de cierto umbral de funcionalidad y precisión en la anticipación de la secuencia correcta (Huang et al., 2023), ha sido posible desarrollar herramientas de traducción de voz a texto en tiempo real (como hace la aplicación *Translator*, de Microsoft, utilizada como soporte de presentaciones y conferencias)[6] , traducir conversaciones de voz sin retardo apreciable (con *Sayhi*, p. ej.)[7] o traducir textos impresos a partir de la imagen que va captando un dispositivo móvil (con *Google Lens*, p. ej.).[8]

DeepL es probablemente el segundo servicio de traducción automática más popular y avanzado; pero difiere de *Google Translate* en su enfoque y tecnologías subyacentes. En lugar de redes neuronales recurrentes y modelos *Transformer* (utilizados por Google) *DeepL* funciona sobre una arquitectura de inteligencia artificial basada en redes neuronales convolucionales (CNN) y *Transformer*, en combinación con *datasets* de alta calidad y diversas técnicas avanzadas (mecanismos de atención y enmascaramiento, para evitar el procesado de contenido irrelevante) con mayor capacidad para capturar dependencias y matices en fragmentos largos de texto, generando traducciones más precisas y naturales. Ambos servicios han evolucionado incorporando mejoras graduales, tanto de enfoque (basado en reglas y estadístico en 2006) como de arquitectura (redes neuronales y modelos *Transformer*, desde 2014-2017), cuyo efecto ha sido una reducción importante de la tasa de error según métricas exigentes de calidad (*multidimensional quality metrics)* y una sofisticación notable en

Puntuación BLEU	Interpretación
< 10	Casi inútil
10-19	Difícil de captar la esencia
20-29	La esencia es clara, pero tiene errores gramaticales significativos
30-40	Comprensible por buenas traducciones
40-50	Traducciones de alta calidad
50-60	Traducciones de calidad muy alta, adecuadas y fluidas
> 60	Calidad generalmente mejor que la humana

Tabla 2. *AutoML expresa las puntuaciones BLEU como un porcentaje en vez de como un decimal entre 0 y 1.*
Fuente:https://cloud.google.com/translate/automl/docs/evaluate?hl=es-419#bleu.

5. Actualización propia, para los años 2022 y 2023, de los datos incluidos en P. Kharya, A. Alvi (Oct. 11, 2021): "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model." Disponible en: https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/.

6. Sus características y *demos* de funcionalidad pueden consultarse en https://www.microsoft.com/es-es/translator/APPS/PRESENTATION-TRANSLATOR/.

7. Puede comprobarse el listado de idiomas para los que proporciona traducción en conversación, mucho más limitado con respecto a las posibilidades de traducción automática entre idiomas diferentes en https://www.sayhi.com/es/translate/languages/.

8. Para una evolución más detallada de los modelos *Transformer*, ajustes e innovaciones en su arquitectura puede consultarse la entrada https://en.wikipedia.org/wiki/Transformer_%28machine_learning_model%29 .

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

aspectos como fluidez, estilo, ortotipografía, precisión y atención al contexto cultural o profesional de referencia. Una mayor base de datos de entrenamiento permite a *Google Translate* cubrir casi cien idiomas y dialectos, mientras que *DeepL* gestiona muy bien una decena y se ha ido ampliando hasta la treintena.

Sin embargo, no son diferencias cuantitativas drásticas las que decantan la elección del usuario especializado. Conforme a la variante *AutoML Translation* de la métrica BLEU (*BiLingual Evaluation Understudy*), basada en corpus y diseñada para evaluar el texto traducido automáticamente por el porcentaje de similitud con un conjunto de traducciones de referencia de alta calidad, un valor <10 significa que la traducción automática de salida es casi inútil; entre 10 y 29 la calidad es baja, difícil de entender o con errores gramaticales significativos; mientras que un valor entre 40 y 50 significa que se ajusta a las traducciones de referencia de calidad alta o muy alta. Valores superiores a 60 se consideran generalmente resultados mejores que los obtenidos por traductores humanos (*Figura 2*).

En 2017 *Google Translate* y *DeepL* tenían una puntuación BLEU similar, alrededor de 27. Sin embargo, en 2023, *DeepL* había alcanzado una puntuación BLEU de 42, mientras que *Google Translate* no pasaba de 35. En valoración de utilidad/satisfacción como aplicación móvil, *DeepL* alcanza hoy 4,7 (sobre 5), mientras que *Google Translator* se queda en 4,5.[9]

Por lo tanto, *el salto de calidad percibida se produce a veces en un intervalo muy reducido (5-9 puntos porcentuales) de ganancia de precisión y rendimiento*, promediando las tareas con mejor y peores resultados.[10] Aunque las empresas no proporcionan fácilmente datos desglosados e independientes del rendimiento de sus modelos —lo que dificulta la evaluación de aspectos esenciales en las aplicaciones comerciales de los desarrollos en IA, como señalan acertadamente Hernández-Orallo y demás firmantes de Burnell et al., 2023— el incremento progresivo de calidad en la traducción, fluidez y manejo del contexto marcan la diferencia en este caso a favor de DeepL.

La perspectiva de superar el umbral de 50 en la escala BLUE resulta verosímil para un intervalo de pocos años (2025-2027) y, a muchos efectos relevantes, es previsible que los resultados bajo condiciones exigentes no solo resulten indistinguibles del promedio humano, sino que puedan superar el nivel promedio de calidad alta asociado hasta ahora con traductores humanos expertos antes de 2030 (Buttazzo, 2023: 3, 5).

9. Cfr. DeepL alcanza 4.7, con 187 k de opiniones y 10 M+ descargas (enlace); Google Translate se queda en 4.3, con 8.81 M de opiniones y 1000 M+ de descargas (enlace).

10. Las compañías no proporcionan fácilmente evaluaciones que puedan considerarse independientes de las métricas de rendimiento de sus modelos. Una idea aproximada puede obtenerse con diversas consultas a través de Bard (vinculada al ecosistema Google) sobre datos de su competidor DeepL.

Un proceso gradual como el observado en diversas herramientas y modelos potentes de traducción automática es compatible con logros y ganancias de calidad o funcionalidad que pueden resultar espectaculares para el usuario promedio y servir de base para múltiples iniciativas de explotación comercial. Pero difícilmente sorprenden a quienes hayan seguido los pormenores de esta evolución y tengan alguna noción acerca de las tecnologías, enfoques y arquitecturas involucradas. Para el usuario experto con criterio informado y familiarizado con las ventajas e inconvenientes de los bancos de prueba (*benchmarks*) utilizados para el ajuste fino de estos modelos, es muy probable que los resultados continúen resultando decepcionantes en un porcentaje demasiado alto de tareas.

3. Dinámica gradual previsible en la trayectoria del programa AGI

Los modelos y aplicaciones de traducción automática analizados en el apartado anterior pueden parecer asociados a usos relativamente restringidos y especializados. Sin embargo, gran parte de sus fortalezas han quedado incorporadas en modelos de alto rendimiento (LLM) como GPT-3, GPT-4, Bard/Gemini, Anthropic/Claude o Cohere, entre otros, y parece consolidarse la tendencia a ampliar su funcionalidad con posibilidades multimodales para analizar (traducir, describir, interpretar) imágenes, sonido y vídeo. Además, pueden proporcionar resultados de búsqueda integrando información de múltiples fuentes, seleccionado los hechos pertinentes y con la posibilidad de resumir (bajo distintos parámetros) textos complejos o generarlos ajustando su complejidad, estructura, estilo, tono y formato a múltiples destinatarios, para finalidades muy diversas.

Utilizados como sistemas conversacionales, los LLM pueden servir de interfaces muy eficaces para resolver problemas de cierta complejidad con una gama amplia de dispositivos, dependiendo de la sofisticación y nivel de agencia de los elementos conectados (robots en entornos industriales, servicios de reserva o cita previa, consultas y asesoramiento en la interacción con las administraciones, etc.). Las funciones y tareas con LLM que ahora se realizan a través de equipos de escritorio fácilmente podrán llevarse a cabo en breve desde dispositivos móviles e incluso sin conexión, ejecutando una versión reducida del LLM optimizada para hardware menos exigente en recursos (Ahmad y Rehaan, 2023; Pichai y Hassabis, 2023).

Esta ampliación gradual de funcionalidad y versatilidad (texto, imagen, voz) para el usuario común puede resultar revolucionaria para individuos con visión reducida o limitaciones cognitivas y de movilidad, por ejemplo. En lo esencial, el enfoque de desarrollo, la infraestructura necesaria y las fases de entrenamiento de los modelos permiten transferir las estrategias de mejora y aprendizaje para optimizar los resultados con texto, imagen o sonido. Pero pueden afinarse igualmente para programación y depuración de código o creación de aplicaciones y juegos; y entrenarse para usos tan específicos como el análisis de secuencias de genes en bases de datos, la identificación de la estructura de proteínas, el desarrollo de fármacos,

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

o el diagnóstico de diversos tipos de tumores a partir de imágenes médicas (Russe et al., 2023; Callaway, 2020). En cada uno de estos dominios, los modelos *generativos Transformer preentrenados* (GPT) han producido resultados que igualan o superan las capacidades de personas expertas en tareas consideradas intelectualmente exigentes; y, en ciertos casos, han resuelto desafíos y problemas pendientes durante décadas (Yang et al., 2023).

Sin embargo, quienes utilizan estos modelos para programar pueden sentirse muy satisfechos con porcentajes de acierto en las respuestas superiores al 30% —puesto que ya supondría un ahorro de tiempo considerable en aspectos tediosos de la programación— y, con variantes mucho mejor ajustadas y entrenadas con más recursos o datos de alta calidad, considerar que pueden abordar proyectos profesionales de mayor calado (Warren, 2023). Aunque para profesionales experimentados el código proporcionado por el modelo GPT no resulte particularmente sofisticado ni refinado para el proyecto o plataforma de referencia, un usuario medio sin apenas nociones de programación puede ampliar de forma extraordinaria sus habilidades en poco tiempo y para múltiples lenguajes (Kalyan, 2024). Por esta razón son importantes los bancos de pruebas especializadas para evaluar de manera integrada e independiente el rendimiento de los grandes modelos de lenguaje (Burnell et al., 2023; Shen et al., 2023).

En las tareas mencionadas, el aspecto relevante concierne al *propósito del servicio* para el cual se elige la herramienta disponible con mayor potencial, y el proceso de ajustes o entrenamiento especializado requerido. En ámbitos profesionales se valoran en particular las mejoras en eficiencia, productividad y capacidad de adquirir ventaja competitiva. Aplicaciones que pueden entusiasmar a usuarios en el segmento de consumo, con el objetivo de satisfacer necesidades personales o recreativas, no necesariamente conllevan saltos cualitativos significativos en las tecnologías y enfoques subyacentes.[11] De hecho, pueden tener un efecto disruptivo en el mercado incluso con versiones muy ligeras y de menores prestaciones del modelo, si a cambio ganan en funcionalidad y versatilidad para integrarse, por ejemplo, en navegadores o en aplicaciones de ofimática (Pastor, 2023a).

Lo previsible es que estos modelos de IA incrementen su generalidad y versatilidad para un rango creciente de tareas, igualando y superando tanto las habilidades del operador humano promedio como el rendimiento de expertos o de grupos interdisciplinares de expertos con las versiones más pesadas, completas y utilizando al máximo los recursos energéticos y de infraestructura necesarios. Cabe dudar de que la infraestructura y herramientas de ajuste fino requeridas para obtener el mayor rendimiento en cientos de tareas cualificadas estén al alcance de usuarios en el segmento de consumo (Ludvigsen, 2023; Foy, 2023). Pero un modelo GPT que adquiera competencia cognitiva destacable en varias decenas de tareas más que otras versiones previas puede suponer para ciertos usos y contextos de aplicación una mejora cualitativa sustancial, reforzando la confianza en su capacidad, fiabilidad y precisión hasta el punto de convertir dicha herramienta en imprescindible. Esta ampliación de generalidad y rendimiento constituye un proceso medible con las herramientas de evaluación y *benchmarks* (Shen et al., 2023). Y se puede considerar compatible con criterios genuinos de inteligencia artificial general cuando el rendimiento supere los valores humanos promedio en cierto número de escalas de referencia exigentes (Bernstein et al., 2023; López Espejel et al., 2023; Fisher & Fisher, 2024; Rao et al., 2023).

Es posible que el modelo no pueda sustituir al profesional humano en una decisión o valoración clínica, p. ej. Pero quizá se adopte porque puede revolucionar aspectos sustanciales de su tarea, proceso de aprendizaje y especialización profesional (Fonseca Magalhaes Filho et al., 2023). Si se logran obviar muchas ineficiencias del aprendizaje profundo no supervisado con datos sintéticos de alta calidad, es probable que el proceso de ampliación de competencias generales de alto nivel se amplíe y acelere, hasta reducir la tasa de errores a cifras marginales. Esto solo refuerza el acierto del enfoque, la idoneidad de la infraestructura utilizada y el afinamiento logrado con el modelo, sin connotación alguna de singularidad o superinteligencia que pueda asombrar a los desarrolladores. Algunos factores externos (computación en la nube, repositorios públicos e incremento del ancho de banda en las redes de comunicación) podrían agilizar aún más el proceso. Pero, en términos de impacto social, será inevitable reconocer que algo interesante ocurre cuando ciertas herramientas permiten a miles de usuarios sin especialización adquirir en poco tiempo competencias lingüísticas, de cálculo, programación y análisis estadístico reservadas habitualmente a profesionales con formación superior especializada (Polverini y Gregorcic, 2023; Hoffmann, 2022a, 2022b).

Del mismo modo que cambios de enfoque y optimización gradual o ajuste fino pueden mejorar la versatilidad, generalidad y rendimiento de un modelo LLM o GPT, ligeras mejoras de una versión a otra pueden originar comportamientos inesperados y abrir posibilidades que ciertos usuarios podrían explotar para fines ilícitos. Una mejora en las funciones de asistencia a la programación y depuración de código podría posibilitar ataques a servidores y redes para los que no existen medidas preventivas. Antes que caer en el asombro por la cercanía de la singularidad, lo esperable es que la empresa y desarrolladores del modelo utilicen bancos de pruebas propios o externos exigentes y actualizados para configurar el modelo de manera que evite arrojar ciertos *outputs* o pueda identificar las estrategias de engaño posibles para obtenerlos de modo indirecto (Goertzel, 2007; Goertzel y Pennachin, 2007).

11. El modelo de IA conversacional *Replika* ha tenido un éxito notable por su capacidad para convertirse en un amigo, acompañante, o amante virtual, con funciones relativamente simples de personalización y eficacia para fidelizar a miles de usuarios en interacciones satisfactorias. En una versión algo peculiar de superación del test de Turing, *Replika* llegó a generar tal dependencia en un gran número de usuarios que los desarrolladores tuvieron que limitar su funcionalidad y diferenciar versiones para amistad, salud mental y romance (Pastor, 2023b; Tong, 2023).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Un modelo LLM/GPT de cierta complejidad tendría que incluir reglas y entrenamiento para conseguir aprendizaje y capacidades avanzadas en materia de ciberseguridad, de modo que pueda detectar o prevenir usos delictivos y autolimitarse o censurarse en las respuestas (Abel, 2023). *Su ampliación en competencia cognitiva y versatilidad*, generalizando su potencial a cientos de tareas por encima del nivel propio de usuarios expertos o facilitando ajustes de entrenamiento y especialización adicional según necesidades del usuario final, *conllevaría inevitablemente limitaciones funcionales diseñadas por los desarrolladores en la fase de validación o autoinducidas*, para mantener las exigencias de fiabilidad, seguridad y alineamiento con el marco legal.

La casuística imaginable en los objetivos de individuos malintencionados sería suficiente para entrenar modelos altamente especializados con conjuntos de datos de ciberseguridad y utilizarlos para análisis de *malware* y vulnerabilidad de software, detección y respuesta a ataques cibernéticos, generación de informes sobre eventos, etc. Aplicados en servicios y actividades críticas, sería deseable que su eficacia superara ampliamente a la de ciberdelincuentes humanos experimentados. Múltiples iniciativas van en esta dirección, por el alto valor añadido asociado con tareas profesionalmente exigentes de supervisión y garantías de seguridad en infraestructuras críticas que, en gran medida, podrían beneficiarse de un soporte infatigable y proactivo (detectando intrusiones y anomalías mediante el análisis de registros; sugiriendo estrategias más robustas para protección de datos; o clasificando sin error muestras de *malware*, p. ej.) con modelos de IA especializada.[12]

GPT-4, de *OpenAI*, permite crear *chatbots* especializados en ciberseguridad desde finales de 2023.[13] Utilizando datos propios e incorporando información técnica adicional a la base de conocimiento, una versión personalizada en tareas de ciberseguridad podría ser útil para detectar y prevenir amenazas con mayor eficacia, alertar de vulnerabilidades y reforzar las medidas de protección frente a *phishing*, distribución de malware o acceso a información confidencial mediante ingeniería social. Una nueva herramienta altamente eficaz para mejorar la formación, cultura preventiva y capacidad de reacción de trabajadores y responsables de ciberseguridad en una empresa sin duda sería un activo valioso e importante. Pero hay que contar con la posibilidad de que se utilice también para fines opuestos (inyección de código, robo de datos, acceso no autorizado, extorsión, etc.). Un modelo lo bastante potente como para superar el rendimiento de personal experto en materia de ciberseguridad difícilmente sería validado para pasar a la fase de explotación comercial sin garantías robustas incorporadas en el diseño. Y tanto el proceso como la infraestructura de entrenamiento debería incorporar enfoques para evitar accidentes y usos delictivos por inyección de comandos en el *prompt*, desajuste por el uso de nuevos datos de entrenamiento distorsionados (*training data poisoning*), sobrecarga para bloquear su disponibilidad a terceros, complementos o *plugins* maliciosos, e imprudencia en la ampliación de la capacidad de agencia por una cadena insegura de permisos para otras aplicaciones conectadas, tales como correo electrónico, calendario o redes sociales (Ortega, 2024).

12. Véase, por ejemplo, *Vectra AI* (https://www.vectra.ai/products/competitive/darktrace), *Sophos AI* con tecnología GPT (https://b2b-cyber-security.de/es/mit-ki-und-chatgpt-algorithmus-jagd-nach-cyberkriminellen/) y *Zscaler* (enlace).
13. Véase, p. ej., *CybGPT* (https://chat.openai.com/g/g-kox62b9Hu-cybgpt-cyber-security), *CyberGPT Adviser* (https://chat.openai.com/g/g-igaKzt9pe-cybergpt) y *Ethical Hacker GPT* (https://chat.openai.com/g/g-j4PQ2hyqn-ethical-hacker-gpt), entre otros.

4. Conclusión

Se dispone de trayectoria suficiente en el desarrollo de aplicaciones de la inteligencia artificial para constatar cómo una ampliación progresiva de las capacidades de aprendizaje y adquisición de competencia cognitiva hasta niveles que superan el rendimiento humano promedio en tareas complejas y especializadas no justifican los temores distópicos acerca de la pérdida de control humano que algunos teóricos asocian con el horizonte de singularidad (Grout, 2018; Kurzweil, 2005; Radanliev et al., 2022). Pero los casos, modelos y dominios de aplicaciones analizados en los apartados anteriores justifican expectativas fundadas de aproximación a escenarios de compatibilidad con criterios genuinos de IA general (AGI) en pocos años (Buttazzo, 2023: 4-5).

Las ganancias de potencia y versatilidad entre las versiones 3 y 4 del modelo GPT de *OpenAI*, por ejemplo, dieron lugar a cambios drásticos en la capacidad para generar contenido creativo e innovador en diversos dominios y estilos. Nuevos enfoques y desarrollos en software (algoritmos), hardware (computación cuántica, nuevos tipos de memoria, datos de sensores y dispositivos conectados) y conjuntos de datos de calidad pueden facilitar el entrenamiento de modelos más grandes, rápidos y eficientes, con mejoras apreciables en la calidad de las respuestas, en la reducción de sesgos y errores y en la ampliación de funcionalidad más allá del procesamiento del lenguaje natural (Matt Swayne, 2023; Li & Zhu, 2022).

La integración o convergencia con otras tecnologías (visión por computadora, realidad aumentada, aprendizaje por refuerzo, robótica y sistemas autónomos, por ejemplo) y la apuesta por la interacción multimodalidad (con imágenes, vídeo, sonido o voz) puede tener un impacto social de consecuencias difícilmente imaginables en ámbitos como la educación accesible, la economía, la investigación científica, el arte y la cultura. Pero lo previsible es que la inteligencia artificial más o menos especializada continúe siendo un complemento capaz de facilitar, optimizar o amplificar el rendimiento humano tanto en tareas altamente especializadas e intelectualmente exigentes como para fines creativos y nuevas oportunidades de ocio y entretenimiento. La aceptación socio-cultural de aplicaciones y dispositivos que logran un rendimiento igual o superior al humano en un rango creciente de tareas y nichos de actividad profesional no constituye un fenómeno nuevo en las sociedades desarrolladas, como se ha podido constatar en la trayectoria del proceso de automatización de fábricas, transporte aéreo y sectores de actividad industrial (Muraille, 2019; Blake et al., 2021; Harwood & Eaves, 2020; Waltermann & Henkel, 2023).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

El incremento del rendimiento, capacidades, versatilidad y generalidad en los modelos LLM o GPT puede medirse y evaluarse con bancos de pruebas y *datasets* como los indicados en los aptdos. 2 y 3, entre un extenso listado que aumenta sin parar. Es improbable que a medio y largo plazo se reduzcan o eliminen ciertos criterios básicos de fiabilidad y alineamiento en la evaluación y validación de tales modelos, como relevancia, coste-efectividad, escalabilidad, fiabilidad, transparencia y reproducibilidad de los resultados (Yampolskiy, 2013). Resulta descabellado suponer, en particular, que se relativizarán criterios o principios para asegurar que el modelo será socialmente beneficioso en los contextos de aplicación más exigentes (atención médica, seguridad, energía, transporte, manufactura, etc.) y considerando usos potenciales en entornos sociales y culturales donde la etnia, el género, la nacionalidad, los ingresos, y las creencias políticas o religiosas pueden resultar determinantes de la aceptación o el rechazo (Coeckelbergh, 2023; Stahl et al., 2014). *A fortiori*, la incorporación de características y funcionalidad que incrementen la autonomía y nivel de agencia del modelo requerirán garantías más estrictas, para minimizar el riesgo de accidente (Floridi, 2023; Li et al., 2018).

Sin descartar la posibilidad de que ciertas tecnologías o modelos salgan precipitadamente de los entornos restringidos donde deberían ponerse a prueba y validarse, resulta inverosímil la ausencia de mecanismos de monitorización y auditoría de funcionamiento tras la implementación. No es descartable un periodo donde la exhibición de capacidades sofisticadas en usos no previstos pueda ser instrumentalizada con facilidad en contra de las políticas de uso y contra valores sociales comunes (Russell, 2019). Pero antes de interpretar el fenómeno como *pérdida de control en un evento de singularidad* parece más razonable enfocar la atención en los resortes políticos y jurídicos necesarios para regular con eficacia disuasoria las malas prácticas en el despliegue de la inteligencia artificial, como se hace con otras tecnologías (Coeckelbergh, 2023; Stahl et al., 2014; Soares y Fallenstein, 2017; Brooks, 2017).

Referencias

Abel, S. (2023). *Top 18 Cyber Security GPTs and How to Use Them (2024)*. StationX. https://www.stationx.net/cyber-security-gpts/

Ahmad, K., & Rehaan, H. (2023). *Fine-Tuning Llama-2: A Comprehensive Case Study for Tailoring Models to Unique Applications*. https://www.anyscale.com/blog/fine-tuning-llama-2-a-comprehensive-case-study-for-tailoring-models-to-unique-applications

Albalawi, F., & Alamoud, K. A. (2022). Trends and Application of Artificial Intelligence Technology in Orthodontic Diagnosis and Treatment Planning—A Review. *Applied Sciences*, *12*(22), 11864. https://doi.org/10.3390/app122211864

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, *79–80*, 3–15. https://doi.org/10.1016/j.jpdc.2014.08.003

Barfield, W., & Blodgett-Ford, S. (Eds.). (2021). *Human Enhancement Technologies and Our Merger with Machines*. MDPI. https://doi.org/10.3390/books978-3-0365-0905-1

Baum, S. (2018). Countering Superintelligence Misinformation. *Information*, *9*(10), 244. https://doi.org/10.3390/info9100244

Berglas, A. (2015). *When Computers Can Think. The Artificial Intelligence Singularity*.

Bernstein, I. A., Zhang, Y. (Victor), Govil, D., Majid, I., Chang, R. T., Sun, Y., Shue, A., Chou, J. C., Schehlein, E., Christopher, K. L., Groth, S. L., Ludwig, C., & Wang, S. Y. (2023). Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. *JAMA Network Open*, *6*(8), e2330320. https://doi.org/10.1001/jamanetworkopen.2023.30320

Blake, R. W., Mathew, R., George, A., & Papakostas, N. (2021). Impact of Artificial Intelligence on Engineering: Past, Present and Future. *Procedia CIRP*, *104*, 1728–1733. https://doi.org/https://doi.org/10.1016/j.procir.2021.11.291

Bostrom, N. (2002). "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*, *9*(March 2002), 1–30.

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

Bostrom, N., Dafoe, A., & Flynn, C. (2016). *Desiderata in the Development of Machine Superintelligence: Vol. v. 3.6*.

Brooks, R. (2017). The Seven Deadly Sins of AI Predictions. *MIT Technology Review*.

Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., Rutar, D., Cheke, L. G., Sohl-Dickstein, J., Mitchell, M., Kiela, D., Shanahan, M., Voorhees, E. M., Cohn, A. G., Leibo, J. Z., & Hernandez-Orallo, J. (2023). Rethink reporting of evaluation results in AI. *Science*, *380*(6641), 136–138. https://doi.org/10.1126/science.adf6369

Buttazzo, G. (2023). Rise of artificial general intelligence: risks and opportunities. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1226990

Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*, *588*(7837), 203–204. https://doi.org/10.1038/d41586-020-03348-4

Camacho, J. de J., Aguirre, B., Ponce, P., Anthony, B., & Molina, A. (2024). Leveraging Artificial Intelligence to Bolster the Energy Sector in Smart Cities: A Literature Review. *Energies*, *17*(2), 353. https://doi.org/10.3390/en17020353

Chalmers, D. J. (2010). The Singularity: A philosophical analysis. *Journal of Consciousness Studies*, *17*, 7–65.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. https://doi.org/10.3115/v1/D14-1179

Coeckelbergh, M. (2023). *Filosofía política de la inteligencia artificial: Una introducción*. Ediciones Cátedra.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Dahlin, B. (2012). Our posthuman futures and education: Homo Zappiens, Cyborgs, and the New Adam. *Futures*, *44*(1), 55–63. https://doi.org/https://doi.org/10.1016/j.futures.2011.08.007

Devagiri, J. S., Paheding, S., Niyaz, Q., Yang, X., & Smith, S. (2022). Augmented Reality and Artificial Intelligence in industry: Trends, tools, and future challenges. *Expert Systems with Applications*, *207*, 118002. https://doi.org/https://doi.org/10.1016/j.eswa.2022.118002

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*.

Diéguez, A., & García-Barranquero, P. (2023). *The Singularity, Superintelligent Machines, and Mind Uploading: The Technological Future?* (pp. 237–255). https://doi.org/10.1007/978-3-031-48135-2_12

Eden, A. H., Steinhart, E., Pearce, D., & Moor, J. H. (2012). *Singularity Hypotheses: An Overview* (pp. 1–12). https://doi.org/10.1007/978-3-642-32560-1_1

EU. (2023). *Digitalisation in Europe: 2023*. https://doi.org/10.2785/442069

Fisher, A. D., & Fisher, G. (2024). Evaluating performance of custom GPT in anesthesia practice. *Journal of Clinical Anesthesia*, *93*, 111371. https://doi.org/10.1016/j.jclinane.2023.111371

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford University Press.

Fonseca Magalhaes Filho, M. A., Aguiar Junior, P. N., Fabre, B. L., Marques, F., Gutierres, B., Nassib William Junior, W., & Del Giglio, A. (2023). 1255P Evaluating GPT-4 as an academic support tool for clinicians: A comparative analysis of case records from the literature. *Annals of Oncology*, *34*, S729. https://doi.org/10.1016/j.annonc.2023.09.2344

Foy, K. (2023). New tools are available to help reduce the energy that AI models devour Amid the race to make AI bigger and better, Lincoln Laboratory is developing ways to reduce power, train efficiently, and make energy use transparent. *MIT News*.

Goertzel, B. (2007). Human-level artificial general intelligence and the possibility of a technological singularity. *Artificial Intelligence*, *171*(18), 1161–1173. https://doi.org/10.1016/j.artint.2007.10.011

Goertzel, B., & Pennachin, C. (2007). *Artificial general intelligence*. Springer.

Good, I. J. (1966). *Speculations Concerning the First Ultraintelligent Machine* (pp. 31–88). https://doi.org/10.1016/S0065-2458(08)60418-0

Grout, V. (2018). The Singularity Isn't Simple! (However We Look at It) A Random Walk between Science Fiction and Science Fact. *Information*, *9*(4), 99. https://doi.org/10.3390/info9040099

Harwood, S., & Eaves, S. (2020). Conceptualising technology, its development and future: The six genres of technology. *Technological Forecasting and Social Change*, *160*, 120174. https://doi.org/https://doi.org/10.1016/j.techfore.2020.120174

Hines, A. (2019). Getting Ready for a Post-Work Future. *Foresight and STI Governance*, *13*(1), 19–30. https://doi.org/10.17323/2500-2597.2019.1.19.30

Hoffmann, C. H. (2022a). Is AI intelligent? An assessment of artificial intelligence, 70 years after Turing. *Technology in Society*, *68*, 101893. https://doi.org/10.1016/j.techsoc.2022.101893

Hoffmann, C. H. (2022b). *The Quest for a Universal Theory of Intelligence*. De Gruyter. https://doi.org/10.1515/9783110756166

Hoffmann, C. H. (2023). A philosophical view on singularity and strong AI. *AI & SOCIETY*, *38*(4), 1697–1714. https://doi.org/10.1007/s00146-021-01327-5

Huang, Y., Zhang, T., & Xu, C. (2023). Learning to decode to future success for multi-modal neural machine translation. *Journal of Engineering Research*, *11*(2), 100084. https://doi.org/10.1016/j.jer.2023.100084

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, *6*, 100048. https://doi.org/10.1016/j.nlp.2023.100048

Kannan, S., Karuppusamy, S., Nedunchezhian, A., Venkateshan, P., Wang, P., Bojja, N., & Kejariwal, A. (2016). *Chapter 3 - Big Data Analytics for Social Media A2 - Buyya, Rajkumar* (R. N. Calheiros & A. V. B. T.-B. D. Dastjerdi (Eds.); pp. 63–94). Morgan Kaufmann. https://doi.org/http://dx.doi.org/10.1016/B978-0-12-805394-2.00003-9

Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence*. Nueva York, NY, Estados Unidos de América.

Kurzweil, R. (2005). *The Singularity is Near. WhenHumans Transcend Biology*. Viking.

Kurzweil, R. (2012). *How to create a mind: The secret of human thought revealed*. Penguin Books.

Li, L., Lin, Y.-L., Zheng, N.-N., Wang, F.-Y., Liu, Y., Cao, D., Wang, K., & Huang, W.-L. (2018). Artificial intelligence test: a case study of intelligent vehicles. *Artificial Intelligence Review*, *50*(3), 441–465. https://doi.org/10.1007/s10462-018-9631-5

Li, Q., & Zhu, Y. (2022). *Adapting Pre-trained Language Models for Quantum Natural Language Processing*. https://ar5iv.labs.arxiv.org/html/2302.13812

López Espejel, J., Ettifouri, E. H., Yahaya Alassan, M. S., Chouham, E. M., & Dahhane, W. (2023). GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, *5*, 100032. https://doi.org/10.1016/j.nlp.2023.100032

Ludvigsen, K. G. A. (2023). *The carbon footprint of GPT-4*. Medium. https://towardsdatascience.com/the-carbon-footprint-of-gpt-4-d6c676eb21ae

Markoff, J. (2009). *The coming superbrain*. Www.Nytimes.Com. http://www.nytimes.com/2009/05/24/weekinreview/24markoff.html

Matt Swayne. (2023). *How Could Quantum Computing Improve Large Language Models?* The Quantum Insider. https://thequantuminsider.com/2023/02/13/how-could-quantum-computing-improve-large-language-models/

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Muraille, E. (2019). Ethical control of innovation in a globalized and liberal world: Is good science still science? *Endeavour*, *43*(4), 100709. https://doi.org/https://doi.org/10.1016/j.endeavour.2020.100709

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Neubauer, A. C. (2021). The future of intelligence research in the coming age of artificial intelligence – With a special consideration of the philosophical movements of trans- and posthumanism. *Intelligence*, *87*, 101563. https://doi.org/https://doi.org/10.1016/j.intell.2021.101563

Ortega, J. M. (2024). *Seguridad y auditorías en Modelos grandes del lenguaje (LLM)*. CodeMotion. https://www.codemotion.com/magazine/es/ciberseguridad/seguridad-y-auditorias-en-modelos-grandes-del-lenguaje/

Pastor, J. (2023a). *Bienvenidos a la era de la IA de bolsillo: Google plantea toda una revolución con Gemini Nano*. https://www.xataka.com/robotica-e-ia/gemini-nano-empieza-nueva-era-ia-bolsillo

Pastor, J. (2023b). *Replika es el chatbot que enamoró a sus usuarios, los desengañó y ahora quiere seducirlos de nuevo*. Xataka. https://www.xataka.com/robotica-e-ia/usuarios-que-amaban-a-maquinas-replika-chatbot-que-enamoro-a-sus-usuarios-ahora-quiere-seducirlos-nuevo

Pichai, S., & Hassabis, D. (2023). *Introducing Gemini: our largest and most capable AI model*. https://blog.google/technology/ai/google-gemini-ai/

Polverini, G., & Gregorcic, B. (2023). *Performance of ChatGPT on the Test of Understanding Graphs in Kinematics*.

Proust, J. (2011). Cognitive Enhancement, Human Evolution and Bioethics. *Journal International de Bioéthique*, *22*(3/4), 153-173,199.

Radanliev, P., De Roure, D., Maple, C., & Ani, U. (2022). Super-forecasting the 'technological singularity' risks from artificial intelligence. *Evolving Systems*, *13*(5), 747–757. https://doi.org/10.1007/s12530-022-09431-7

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Blog*.

Rao, A., Kim, J., Kamineni, M., Pang, M., Lie, W., Dreyer, K. J., & Succi, M. D. (2023). Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *Journal of the American College of Radiology*, *20*(10), 990–997. https://doi.org/10.1016/j.jacr.2023.05.003

Rifkin, J. (1995). *The end of work: the decline of the global labor force and the dawn of the post-market era*. G.P. Putnam's Sons.

Rifkin, J. (2014). *The zero marginal cost society: the internet of things, the collaborative commons, and the eclipse of capitalism*. Palgrave Macmillan.

Rodríguez, R. (n.d.). *GoogLeNet*. N.D. https://lamaquinaoraculo.com/deep-learning/googlenet/

Russe, M. F., Fink, A., Ngo, H., Tran, H., Bamberg, F., Reisert, M., & Rau, A. (2023). Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. *Scientific Reports*, *13*(1), 14215. https://doi.org/10.1038/s41598-023-41512-8

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Sandberg, A. (2011). Cognition Enhancement: Upgrading the Brain. In J. Savulescu, R. ter Meulen, & G. Kahane (Eds.), *Enhancing human capacities*. Wiley-Blackwell.

Sandberg, A., & Bostrom, N. (2008). *Whole Brain Emulation. A Roadmap*.

Schneider, S. (2016). *Science Fiction and Philosophy: From Time Travel to Superintelligence* (S. Schneider (Ed.)). Wiley-Blackwell.

Shen, Y., Song, K., Tan, X., Zhang, W., Ren, K., Yuan, S., Lu, W., Li, D., & Zhuang, Y. (2023). Taskbench: Benchmarking Large Language Models for Task Automation. *Computation and Language*. https://doi.org/10.48550/arXiv.2311.18760

Soares, N., & Fallenstein, B. (2017). *Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda* (pp. 103–125). https://doi.org/10.1007/978-3-662-54033-6_5

Stahl, B. C., Eden, G., Jirotka, M., & Coeckelbergh, M. (2014). From computer ethics to responsible research and innovation in ICT: The transition of reference discourses informing ethics-related research in information systems. *Information & Management*, *51*(6), 810–818. https://doi.org/http://dx.doi.org/10.1016/j.im.2014.01.001

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. Alfred A. Knopf-Penguin Random House LLC.

Tong, A. (2023). *AI chatbot company Replika restores erotic roleplay for some users*. Reuters. https://www.reuters.com/technology/ai-chatbot-company-replika-restores-erotic-roleplay-some-users-2023-03-25/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Kaiser, Ł. (2017). Attention is all you need. In U. von Luxburg; & I. Guyon (Eds.), *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Curran Associates Inc.57 Morehouse LaneRed HookNYUnited States.

Vinge, V. (1993). The Coming Technological Singularity: How to Survive in the Post-Human Era. *NASA Conference Publication 10129*, 11–22.

Waltermann, J., & Henkel, S. (2023). Public discourse on automated vehicles in online discussion forums: A social constructionist perspective. *Transportation Research Interdisciplinary Perspectives*, *17*, 100743. https://doi.org/https://doi.org/10.1016/j.trip.2022.100743

Warren, T. (2023). *GitHub Copilot gets a new ChatGPT-like assistant to help developers write and fix code*. The Verge. https://www.theverge.com/2023/3/22/23651456/github-copilot-x-gpt-4-code-chat-voice-support

Yampolskiy, R. V. (2013). *Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach* (pp. 389–396). https://doi.org/10.1007/978-3-642-31674-6_29

Yampolskiy, R. V. (2016). *Artificial Superintelligence: A Futuristic Approach*. CRC Press, Taylor & Francis Group.

Yang, Z., Zeng, X., Zhao, Y., & Chen, R. (2023). AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduction and Targeted Therapy*, *8*(1), 115. https://doi.org/10.1038/s41392-023-01381-z

Yudkowsky, E. (2007). *Three Major Singularity Schools*. http://yudkowsky.net/ singularity/schools

Yuxiu, Y. (2024). Application of Translation Technology based on AI in Translation Teaching. *Systems and Soft Computing*, 200072. https://doi.org/10.1016/j.sasc.2024.200072

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Tigres de papel.
IA y la amenaza de la singularidad

Fernando Broncano
Universidad Carlos III de Madrid
fernando.broncano@uc3m.es

La irrupción de los LLM (Large Linguistic Models) como Bard, Gemini, Bing o el más popular GPT 3, 3.5 y 4, junto a otros muchos programas de inteligencia artificial aplicada a imágenes, traducción lingüística, música, juegos y videojuegos, ha producido una conmoción inusitada en muchos ámbitos de la sociedad, incluyendo, quizás sobre todo, el sistema educativo en todos los niveles. Parece que, por fin, tras treinta años de avisos y amenazas, la singularidad parece estar definitivamente cerca. ¿Cómo no sentir en los chats que uno está tratando con algo que se parece a una inteligencia general artificial? Y, si de facto alguno de estos dispositivos alcanza algún grado de generalidad, ¿cómo no temer que se haya añadido algún nuevo impedimento a la agencia humana?

Mi objetivo en estas breves líneas es doble. En primer lugar, argumentaré que el debate competitivo sobre inteligencia artificial y natural (las dos son humanas) es poco o nada relevante y, en segundo lugar, que lo que sí debe preocuparnos es cómo se configura o configurará el entorno social que hace posible el funcionamiento e impacto de las inteligencias artificiales (presuntamente) generales.

Una ambigua controversia

Cuando el futurólogo Ray Kurzweil logró en 2005 una popularidad internacional con su amenaza "la singularidad está cerca" (2045, según su predicción), tras varios textos donde anunciaba el crecimiento exponencial e inminente de la inteligencia artificial, toda la cultura mediática quedó infectada en adelante de esta profecía y, tras cada anuncio de un nuevo producto más inteligente que el anterior, se ha ido extendiendo un miedo generalizado a que pueda tener razón y que las viejas películas ochenteras describan fielmente el futuro. Para comenzar, hay que reconocer que parte de su éxito está en la atracción de la narrativa de la "singularidad", que evoca transformaciones radicales y puntos de ruptura, algo así como una revolución de las máquinas que alcanzan el nivel de superinteligencia.

El rótulo de la singularidad debe también su éxito a que plantea de hecho interesantes preguntas filosóficas que se sostienen incluso si el horizonte del evento se ve tan lejano como la conversión del sol en un agujero negro.

¿Qué es la inteligencia artificial general? Es una pregunta no sencilla de responder. En principio, habría tres grados (más adelante ampliaré esta escalera de niveles):

1. Que supere sin restricciones el Test de Türing.
2. Que resuelva problemas que solamente pueden resolver los humanos
3. Que tenga las mismas funciones cognitivas de los seres humanos

Los tres criterios acerca de qué pueda ser una inteligencia artificial general son los más populares y extendidos, pero descansan sobre términos o afirmaciones controvertibles: ¿qué es superar sin restricciones el Test de Türing?, ¿qué son problemas?, ¿cuáles son las funciones cognitivas de los humanos? Precisamente por la confianza en lo intuitivo de estos criterios el término "inteligencia artificial general" ha alcanzado tanta popularidad sin demasiadas preguntas sobre su significado.

¿Qué es la singularidad? La definición kurzweiliana señala el evento histórico en el que la inteligencia de las máquinas supera la inteligencia colectiva humana. Que las máquinas resuelven problemas que los humanos no resuelven es algo poco discutible en el mundo contemporáneo. Suponiendo que la inteligencia pueda medirse numéricamente (por ejemplo, el test IQ), que compara la capacidad de resolución de algunos problemas- índice respecto a la edad y a la curva normal poblacional, una posible interpretación sería que una máquina procesadora superase el máximo de la campana de Gauss humana. Una segunda forma de entenderla sería que rompiese los límites definidos por la tarea encomendada y comenzase a entremezclar dominios de razonamiento de acción, algo así como parece que logró la especie humana a través del lenguaje y lo conceptual. En este sentido podríamos definir el punto de ruptura en tres niveles:

1. Límites de las capacidades cognitivas personales (incluyendo a los miembros más favorecidos intelectualmente de la sociedad)
2. Límites de la inteligencia colectiva, que incluye la deliberación, crítica y potencialidades de los recursos compartidos intelectuales de la humanidad.
3. Límites de la inteligencia aumentada y expandida por la integración virtuosa de humanos y máquinas.

La idea de una superinteligencia artificial general, en su grado más alto, sería aquella que lograse romper los techos de capacidades de solución de problemas en los niveles crecientes desde el (1) al (3). El máximo sería aquel en el que las máquinas no solamente fueran capaces de rediseñar su estructura algorítmica interna por una suerte de meta-aprendizaje, e incluso de rediseñar su hardware y producirlo, sino también (ese es el escenario SkyNet de *Terminator*), de producir su propio espacio de problemas y una agenda propia de transformaciones, ajena, e incluso contraria, a los intereses humanos.

Paralela a la línea de profecías desastrológicas, está la de las promesas mesiánicas que anuncian la posibilidad técnica de inmortalidad digital y paraísos pseudoteológicos similares. Ciertamente, no habría que descartar demasiado rápido el componente utópico de la singularidad en algunos aspectos,

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

trabajo asalariado (no del trabajo como arte de la transformación de lo real), en la medida en que las máquinas fueran sustituyendo a las personas en las tareas más tediosas, repetitivas y mal pagadas, creando un escenario postcapitalista.

Pero estos escenarios críticos no parece que sean los que dirigen la atracción de numerosos capitales de riesgo hacia las empresas implicadas en la creación de inteligencias artificiales generales.

La mezcla de los dos temas ("¿qué es la inteligencia general artificial?", "¿qué es la singularidad?") hace que la controversia que ha generado este bombo publicitario de la singularidad se haya convertido en un territorio pantanoso lleno de adjetivos épicos sostenidos por la incertidumbre y pseudo lenguajes de posibilidad. Porque los espacios de posibilidad siempre son territorios minados en donde la escala tiene mucha importancia. La posibilidad lógica, la física, la realmente técnica, la moral y política, todas ellas intersecan y pueden ser contempladas tanto desde la perspectiva y escala humanas como desde alguna perspectiva cósmica o cósmico-técnica.

Lo general de la inteligencia artificial general

No habrá singularidad, sin la menor discusión, mientras las inteligencias artificiales (uso aquí con toda conciencia el término en plural dado que eso es lo que describe bien el panorama de ofertas de programas en el contexto contemporáneo de los comienzos de la segunda década del siglo XXI) no alcancen un grado de generalidad similar al que tiene la especie humana. El paleontólogo Stephen Mithen explicó gráficamente en qué podría consistir la generalidad de la mente humana con la metáfora de una catedral gótica en la que la nave central está rodeada de capillas. Mithen escribía sobre el origen de la mente en los grandes simios y observaba cómo homínidos y homininos desarrollaban inteligencias específicas orientadas a ámbitos separados de problemas ecológicos: comunicación y relaciones sociales, clasificación de especies animales y vegetales de interés en la supervivencia y culturas técnicas orientadas a la fabricación de herramientas. En las especies de simios más cercanos a la especie humana hay algunas formas de solapamiento entre estas inteligencias de dominio específico, pero las interacciones no han conducido a lo que propiamente podría considerarse una inteligencia general. Mithen explicaba el papel causal predominante que el lenguaje tuvo en la intersección e integración de todas las habilidades particulares. El efecto de generalidad fue muy discutido en varios contextos de la filosofía analítica de los años ochenta y noventa:

> It seems to me that there must be a sense in which thoughts are structured. The thought that John is happy has something in common with the thought that Harry is happy, and ... something in common with the thought that John is sad.... Thus, someone who thinks that John is happy and that Harry is happy exercises on two occasions the conceptual ability which we call 'possessing the concept of happiness'.[1]

Evans proponía este requisito como característica de la posesión de conceptos y por ello del pensamiento conceptual y lo formulaba mediante un conocido requisito:

> If a subject can be credited with the thought that a is F, then he must have the conceptual resources for entertaining the thought that a is G, for every property of being G of which he has a conception.[2]

Grado 1: Esto podríamos considerarlo un grado uno de generalidad inducida por el lenguaje en tanto que productor de conceptos, una de cuyas funciones principales es la recognoscitiva. No sabemos si los actuales modelos lingüísticos grandes satisfacen este criterio de generalidad. En apariencia, sí, al menos dependiendo de la interacción que tenemos con ellos conversacional usando "prompts" adecuados, pero tendríamos que probar su habilidad en contextos recognoscitivos más abiertos, en particular en los reconocimientos de imágenes y la posible interacción física con objetos reales en un entorno tecnológico de robots.

Grado 2: La inteligencia general no se limita, sin embargo, a las cadenas conceptuales. La integración de lo conceptual y lo no conceptual: habilidades, emociones…. En este nivel es conveniente recordar la controversia que se desarrolló en la década de los ochenta como resultado de la propuesta de Jerry Fodor de una arquitectura para la mente humana divida en módulos especializados e inteligencia general. Aunque el debate se centró fundamentalmente en la naturaleza de los módulos, es muy relevante recordar que Fodor consideraba como rasgo fundamental de la inteligencia general no simplemente la ausencia de dominio específico, sino lo que denominaba "holismo quineano", es decir, la posibilidad de conectar contenidos muy lejanos en la red de conceptos y creencias. Más allá, incluso, de las fronteras que consideraría Fodor aceptables, la inteligencia general conecta lejanías improbables entre lo conceptual y lo no conceptual, entre habilidades sensoriomotoras y habilidades conceptuales, entre creencias y emociones. La inteligencia general, en este grado, relaciona lo epistémico con los espacios prácticos y los desiderativos y normativos.

Son rasgos centrales en la inteligencia humana, por un lado, la potencia inferencial abductiva, en la que está implicada la hiperconectividad del holismo. Es cierto que un modelo lingüístico generativo como GPT 4 contiene una asombrosa conectividad basa en casi dos millardos de parámetros, que le permite producir resultados tan sorprendentes. Pero la conectividad humana no solo es mucho mayor. El cerebro humano, compuesto por aproximadamente cien millardos de neuronas, cada una de ellas con siete mil sinapsis alcanza a un orden de casi ochocientos millardos de parámetros. Pero no se trata de la comparación puramente cuantitativa, que los partidarios de la Ley de Moore podrían afirmar que es técnicamente alcanzable, sino de la estructura de esta conectividad, que no solamente se produce entre las neuronas, en tanto que transmisores y procesadores de información, sino también y sobre

1. Evans, G. (1982), *The Varieties of Reference*, Oxford: Clarendon Press, p. 100

2. O. c. p. 104.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

todo en la interacción entre la información soportada por la conectividad eléctrica y la transportada por la energía química de las hormonas y neurotransmisores que componen el entorno neuronal y conectan las glándulas y los efectores fisiológicos. La complejidad de la integración humana no es solamente entre neuronas, sino entre células, tejidos y órganos en todo el esplendoroso espectro de la arquitectura de lo vivo. El pensamiento abductivo humano produce inferencias en las que todos estos componentes están implicados y no solamente las propiedades lógicas inscritas en la parte conceptual, sino que, cuando alguien afirma algo así como "lo único que tiene sentido, dados estos datos, es…" está apelando a un registro de trasfondo de sentido que incluye toda la complejidad fisiológica.

Supongamos en aras del argumento que alguna máquina llega a este estadio de organización en el horizonte de eventos. Habrá sido un avance asombroso en la historia de la técnica y seguramente pasará versiones más rigurosas del Test de Türing que las de los chats actuales, pero no será la singularidad. Incluso si procesa todo el caudal del conocimiento humano y resuelve problemas prohibidos a los limitados cerebros humanos. La biología y la historia han producido otros niveles de organización e integración que tienen una dimensión vertical, ortogonal a la del holismo de contenidos.

Grado 3: Me refiero a la integración vertical en la constitución del sujeto en sus dimensiones experiencial, cognitiva y performativa. Para comenzar, es más que dudoso que dispongan de habilidades de metacognición. La metacognición es una función mental que compartimos casi todas las especies de mamíferos y quizá de aves, que consiste en la capacidad para evaluar la dificultad de una tarea en relación con las posibilidades de éxito y, en caso negativo, abandonar el proyecto antes de emprenderlo.

En un segundo nivel, la integración y constitución del sujeto exige la determinación de un orden de valores, entendido como una jerarquía de lo que importa. Más allá de la metacognición, la ordenación de valores regulativos de la existencia es lo que diferencia un sujeto agente de un mecanismo instrumental. No sabemos el grado en que algunos animales logran tener un orden de valores propios (pensamos en un orden natural instintivo, pero muchas especies logran formas de ordenación de valores mucho más sofisticados, por ejemplo, los que establecen sus vínculos emocionales y apegos). La constitución de un sujeto entraña la definición de límites y barreras que incluyen aquello con lo que se puede vivir y dejan fuera lo que hace de una vida indigna de ser vivida.

Queda aún un último nivel de integración que también pertenece a nuestra naturaleza animal por más que la cultura humana lo haya sofisticado y llenado de complejidad y matices.
Grado 4: El enactivismo radical se ha alejado de las arquitectónicas cognitivistas de la mente, tan inspiradas ellas en los modos de procesamiento combinatorio y secuencial de los ordenadores clásicos. En un cierto aspecto ofrece un modelo de la mente que se asemeja mucho a los sistemas artificiales en lo que respecta al estilo conductista de aprendizaje, también en lo que respecta a la dependencia de instructores y de entrenamiento social. La semejanza, sin embargo, solamente llega hasta aquí. El nivel más complejo de integración agente de los sujetos humanos es el de la capacidad de constituir la experiencia como algo más que como una interacción informacional sobre el medio. La mente humana no se limita a la evidencia, sino que integra los aspectos fenoménicos de la corporeidad con el sentido que adscribe a las intenciones de los otros con los que interactúa. La experiencia, en este sentido, es la constitución de un relato que integra lo vivido en un orden y proyecto de existencia, donde las voces y el discurso de los otros es una parte constitutiva esencial. La integración experiencial y de la alteridad contiene ya una forma muy particular de topología del tiempo implicado en lo narrativo de la experiencia: el conjunto de asimetrías pasado-presente-futuro, sin las que no podría constituirse la memoria y la imaginación presentes en el modo de asimilación de la integración práctica.

En resumen, los grados de generalidad de la inteligencia se estructuran en grados y niveles que exigen algo más que independencia de dominio y funcionalidad específicos. La inteligencia general exige además integración mente-cuerpo, agente-entorno y sujeto-cultura.

Lo artificial de la inteligencia artificial general

La historia de la tecnología nos muestra muchas cosas. Una de ellas, no menor, es que las trayectorias de las "innovaciones disruptivas" y la transformación real de la historia por la extensión de una cierta innovación tecnológica no son ni la misma ni siquiera paralelas, sino que se entrecruzan y divergen de las formas más extrañas. Un ejemplo clásico es el de la tecnología de la energía de vapor, desarrollada ya en Alejandría, pero cuyo impacto no fue notorio en la humanidad hasta el siglo XIX avanzado. Mucho más reciente, y relacionado con nuestro tema, es la historia de las redes neuronales, ya pensadas en los mismos inicios conceptuales de la informática, pero no tomadas en serio técnicamente hasta finales de los noventa.

El problema histórico, cultural, sociológico y filosófico relevante es por qué se producen estas desigualdades en el desarrollo de la tecnología y su integración en la sociedad. Y sobre todo por qué no son percibidas por los discursos escatológicos de la tecnología. Probablemente hay muchas causas y razones sociales que convergen en estas contingencias, pero una de las más significativas es que los inventos, como los descubrimientos científicos, no pasan directamente a formar parte del entorno social. En primer lugar, debe percibirse su funcionalidad; en segundo lugar, debe percibirse la potencialidad de beneficio económico, social o militar que tendría su producción en masa y, en tercer lugar, pero no menos importante, están las derivas y transformaciones, muchas veces creativas, que tienen en el ámbito del uso y de la incorporación a la cultura material.

Ningún invento o innovación es por sí mismo disruptivo o transformador sin una modificación de la cultura material en su entorno que hace posible su extensión social, y, sobre todo,

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

sin una creación de líneas o trayectorias de uso que modifican su funcionalidad. La tecnología es una mediación que transforma identidades y culturas, pero al mismo tiempo es transformada por las pautas que componen esas identidades y culturas. Pensemos, por ejemplo, en la "invención" del motor de combustión interna patentado en 1886 por Carl Benz –después de una larguísima serie de prototipos que nos llevan a centurias anteriores–. Nada estaba escrito en esa patente ni en su diseño que el mundo y las ciudades se llenaran de carreteras de asfalto por la que circulasen millones de automóviles privados. La trayectoria tecnológica que indujo a la generalización del automóvil privado fue tanto un producto de la tecnología como del entorno socioeconómico que creó el nuevo urbanismo y el espacio del transporte privado. Tampoco estaba escrito en el invento del ordenador que iba a ser la informática personal la que reingenierizase el mundo creando el entorno digital. La distancia entre un invento y su impacto social es la que media entre cualquier hecho histórico notable, por ejemplo las revoluciones americana y francesa y las transformaciones que genera en las sendas ulteriores de la historia.

El determinismo tecnológico suele ser un compañero de viaje tan inevitable como molesto en los discursos sobre la inteligencia artificial general y la singularidad. "Determinismo tecnológico" es *contradicto in adiecto*, como el viejo chiste de "música militar". Lo que ocurre es que si es tecnológico no puede ser determinista: depende contingentemente de las sendas erráticas por las que se constituye un entorno técnico favorable, las mucho más aleatorias de los usos y el entrenamiento humano y las caprichosas de los intereses institucionales que sostienen la tecnología en su difusión.

Lo generativo de la inteligencia artificial generativa

La fuente mayor de asombro que producen los nuevos dispositivos que comenzaron a hacerse públicos en la segunda década del presente siglo es la novedad de sus respuestas, el que los textos que escribían o las imágenes, sartas de códigos o músicas que producían eran sorprendentemente correctos o al menos tenían la apariencia de serlo. Eran una nueva generación de algoritmos de la larga tradición de aprendizaje automático que estaba en el origen de la historia de la inteligencia artificial. Estas nuevas generaciones habían abandonado la vieja programación para adoptar las redes neuronales recurrentes (RNN), las redes generativas adversativas (de términos en competencia, GAN) y, últimamente, lo que ha fundamentado el éxito de OpenAI, la arquitectura transformadora basada en los *Transformers*. GPT significa *Generative Pre-trained Transformer*. Contiene la alianza de un LLM o modelo lingüístico grande con un transformer, es decir, un conector de frecuencias de aparición de términos en frases en enormes cantidades de texto con un transformador que produce frases con la probabilidad más alta de que encajen con la respuesta a la pregunta *prompt* hecha por el usuario. El punto de ventaja que ha convertido estos modelos en los héroes del año que acaba cuando escribo estas líneas, 2023, es la técnica que imita la atención humana, la llamada *multi-head attention* que combina resultados de trabajos estadísticos en paralelo para esta producción de resultados estadísticos.

Necesitan una innumerable secuencia de sesiones de entrenamiento, puesta a prueba y, de hecho, son entrenadas por cada usuario que interactúa con estos "transformers". Estímulo-respuesta, sofisticada arquitectura de conexiones, enormes cantidades de datos y enormes cantidades de tiempo en paralelo entrenándolos. Esa es la base de la admirable capacidad generativa de las inteligencias artificiales de nueva generación. Tal vez se pudiera aducir, no sin razones, que esta mezcla es lo que hace que las inteligencias artificiales sean tan cercanas al cerebro humano. Ahora que ya no creemos tanto en el innatismo tipo-Chomsky, y se ha reivindicado por la corriente enactivista algo similar a un neo-conductismo, podría afirmarse que el cerebro de un niño tiene una admirable capacidad generativa por la arquitectura de sus conexiones y el cuidadoso trabajo de sus cuidadores entrenándole.

No. Las analogías entre las máquinas generativas y el cuerpo humano llegan solamente hasta aquí. Aunque la comprensión humana se basase, como en los modelos lingüísticos, en proximidades estadísticas almacenadas en los pesos de las conexiones neuronales, un cerebro humano, y los de muchas especies animales, está dotado de espontaneidad, de una capacidad de auto-preguntarse, de reaccionar y activar redes neuronales en la imaginación, el sueño y, sobre todo, en la capacidad de elaborar proyectos que entrañan la producción de auto-problematizaciones.

La generatividad automática está conducida y controlada por el entorno de interacciones de los usuarios. La máquina responde muchas veces a ellas afirmando sus propios límites (es lo mejor de los últimos resultados de los entrenamientos, debido a las quejas de tantos usuarios) o (todavía, desgraciadamente) haciendo aparecer respuestas que nacen únicamente de los pesos de conexión, muchas veces equivocadas para desesperación de quienes las reciben y, mucho más de quienes las usan confiando en ellas.

No voy a responder aquí a la cuestión filosóficamente profunda y difícil de qué es lo auténticamente novedoso en la creatividad humana. La filosofía de la ciencia y de la mente se ha cuestionado este tema desde hace décadas. Lakatos, por ejemplo, basó en el criterio de producción de novedades relevantes, su filosofía del cambio científico progresivo y no estancado. Hay algo misterioso en la espontaneidad del cerebro. Si pensamos en creadores reconocidos, por ejemplo Monet, un pintor que nunca pintó nada inspirado en narrativas religiosas o míticas, ni siquiera un desnudo, ni se inspiró en la historia del arte, sino que construyó un mundo propio de colores que está en la base de la pintura abstracta. Los cuadros de Monet nacen de un mundo propio autocorregido, en permanente cambio dentro de un mismo proyecto. La creatividad de gente como Monet puede que sea combinatoria, puede que el almacén de recursos esté ya en el ambiente y haya permeado al cerebro creador, pero la novedad está en el acto de juzgar qué combinatoria es relevante para un modelo de mundo interno generado espontáneamente.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Seguramente la generatividad de las nuevas máquinas es creativa en un cierto sentido. Admirable, sin duda, pero aún demasiado dependiente de las preguntas y correcciones de los usuarios y entrenadores. La creatividad de estas máquinas sigue siendo en gran medida humana.

No habrá singularidad para los seres híbridos

La inteligencia humana situada es una inteligencia encastrada, encarnada, enactiva, emocional, extendida técnicamente y social y culturalmente distribuida. Estas características definen la escala y el punto de vista desde el que cabe juzgar y comparar inteligencias. El universo es más grande y poderoso que los humanos, pero solamente los tontos se sentirían abrumados por ello si fuesen incapaces de equilibrar su perspectiva con el otro polo de admiración: la autonomía para generar un mundo de fines y vivir una vida guiada por el valor de lo que importa. Por ahora, las máquinas seguirán desarrollándose como mediaciones mediadas por seres híbridos, cíborgs, en un mundo mucho más complejo que ellas, como lo es una simple célula o un organismo unicelular.

Las IAs generativas avanzadas, construidas con una gigantesca cantidad de conexiones, alimentadas por una descomunal masa de datos, entrenadas por una multitud de auxiliares son una gran conquista de la humanidad y también, claro, para las empresas que las crean y explotan.

No cabe duda sobre su futura utilidad. Serán aportaciones valiosas siempre que haya un entorno adecuado de usuarios que se apropien de sus posibilidades y las empleen como herramientas por las que circulan muchas de otras conquistas anteriores de la cultura humana. Pues son procesadores generativos, pero dependen del reservorio común de producciones humanas depositadas en los almacenes digitales de datos y contenidos.

No cabe duda tampoco de que son instrumentos poderosos y que, por ello, formarán parte de un entorno técnico material que mediará en la formación y desarrollo de identidades culturales futuras. Sí hay razones para dudar, por el contrario, que su futuro ya esté escrito en alguna ley determinista de explosión computacional. Su vida futura dependerá de trayectorias contingentes que escribirán y describirán las vidas de los usuarios, de otras máquinas y de nuevos diseños.

No cabe duda de que los modelos lingüísticos son impresionantes y causan asombro cuando chateamos con ellos pidiéndoles respuestas a preguntas sofisticadas. Ahora bien, al menos de momento, dependen de los contenidos que hemos producido los humanos y, también por el momento, solo son capaces de moverse en un entorno semiabierto de imágenes y textos digitales bien controlado por los instructores. Han sido creados para fascinar y cumplen bien su función. Está por ver, y eso sí será asombroso, si se producirán IAs que tengan capacidades multimodales, multisensoriales y se muevan en entornos abiertos físicos y sociales, sin mapas y con la tarea de resolver problemas absolutamente nuevos en ilimitados niveles de profundidad. Quizás, desgraciadamente, sean las guerras actuales y futuras los laboratorios donde se pongan a prueba esas nuevas habilidades que, entonces, sí, les harían más cercanos a los animales, también a los humanos.

En un futuro previsible, lo único singular de la singularidad va a ser la capacidad social para domesticar los sistemas complejos de diseño, entrenamiento, difusión comercial o institucional y uso inteligente de un entorno técnico en la idea de constituir espacios de posibilidad de un mundo más justo y un tiempo liberado progresivamente de la sumisión al trabajo asalariado en un planeta que preserve la vida.

Una parte de este ambiguo futuro dependerá de cómo se resuelva la disputa por la transparencia de las IAs generativas. El diseño de sistemas con tal nivel de conectividad seguramente seguirá siendo confidencias. También seguirá siendo opaca la estructura interna de las conexiones pues es una característica de la espontaneidad autoorganizativa de las redes neuronales, algo que comparten con las biológicas. Pero la transparencia exigible legal, ética y políticamente debe alcanzar cuanto antes a los modos de entrenamiento, a la expropiación de datos y contenidos y a los usos discapacitadores de estos dispositivos.

El control del ciberespacio por parte de los humanos que lo producen y mantienen: eso sí será una verdadera singularidad en la historia. De manera análoga al proyecto de descolonización del espacio físico, el control democrático del digital será uno de los grandes proyectos y retos futuros mucho más apasionante que el juego de guerra entre humanos y máquinas. Pues no es a las inteligencias artificiales a las que hay que temer, sino a las mucho más peligrosas, irracionales, prejuiciosas y crueles de los poderes oscuros humanos, demasiado humanos.

Las cuestiones seguirán siendo filosóficas, por ello epistemológicas, ontológica, morales, políticas: cómo conquistar, preservar, cuidar la autonomía humana bajo condiciones de dependencia internas (psicológicas, sociales, culturales) y externas (ecológicas, técnicas). Restan por responder muchas preguntas y desarrollar temas abiertos, pero las líneas esenciales deberían poderse dibujar ya en una agenda filosóficamente informada:

- ⇒ Nuevas virtudes epistémicas híbridas
- ⇒ Nuevas racionalidades distribuidas
- ⇒ Nuevas agencias extendidas
- ⇒ Viejos valores y lealtades a la vida y a la humanidad preservados.

Referencias (para seguir leyendo)

Bostrom, N. (2022). *Superinteligencia. Caminos, peligros, estrategias.* Teell Editorial, (Oxford University Press, 2014).

Callahan, V.; Miller, J., Yampolskiy; Armstrong, S. (eds.) (2017). *The Technological Singularity. Managing the Journey,* Dordrecht: Springer.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Coeckelbergh, M. (2022). *The Political Philosophy of AI*, Londres: Polity

Lee, D. (2023). *The Singularity: Artificial General Intelligence (AGI) and ChatGPT,* A Sky Curation: edición digital

Morozov, Evgeny (30 June 2023). "The True Threat of Artificial Intelligence". *The New York Times*. Recuperado el 24/12/2023.

Russell, S.; Norvig, P (2021). *Arificial Intelligence. A Modern Approach*, Hoboken, NJ: Pearson

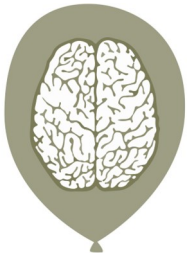
Tegmark, M. (2017). *Life 3.0. Being Human in the Age of Artificial Intelligence.* Nueva York: A. Knopf

Wang, P.; Liu, K.; Dougherthy, Q. (2018). "Conceptions of Artificial Intelligence and Singularity", *Information (MDPI), 9,79*, 1-15

Dee, C. (2023). "Large language models (LLMs) vs generative AI: what's the difference?" *Algolia*, nov. 2023, https://www.algolia.com/blog/ai/large-language-models-llms-vs-generative-ai-whats-the-difference/

Wikipedia: "Transformer (Machine Learning Model) https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)

Caruana F. (2023). "An introduction to transformer models in neural networks and machine learning" https://www.algolia.com/blog/ai/an-introduction-to-transformer-models-in-neural-networks-and-machine-learning/



www.solofici.org

SLMFCE

Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Máquinas inteligentes: Cómo tratar a nuestras criaturas[1]

Blanca Rodríguez López
Universidad Complutense de Madrid
bmerino@filos.ucm.es

"When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong."
Sir Arthur C. Clarke, *Profiles of the Future*

Resumen: Pese a las dudas sobre que sea posible la existencia de una superinteligencia o de que esta suponga un grave riesgo para los humanos, y puesto que no parece haber argumentos concluyentes en contra, en este texto supondemos que ambas cosas son posibles. En este texto nos preguntaremos qué podemos hacer para intentar minimizar ese riesgo, centrándonos principalmente en consideraciones éticas ligadas al posible estatus moral de las máquinas inteligentes.
Palabras Clave: superinteligencia, riesgo existencial, estatus moral, incertidumbre

Introducción

La singularidad, término que en física se refiere a un punto en el centro de un agujero negro, en el que la densidad, la masa, la gravedad y otras cantidades físicas son infinitas. Debido a esto, las ecuaciones de la física no funcionan, lo que más o menos significa que estamos en "territorio" desconocido y puede pasar cualquier cosa.

En los últimos tiempos, sin embargo, cuando oímos hablar de "singularidad" lo más probable es que no se esté hablando de eso, sino de la "singularidad tecnológica", a la que para abreviar se le quita el apellido. Según relata Stanislaw Ulam, el término se había utilizado en este sentido en una conversación con John von Neumann ya en 1958, pero su entrada por la puerta más o menos grande no ocurrió hasta que en 1993 fué utilizado no ya no por científicos sino por un escritor de ciencia ficción, Vernor Vinge, en un artículo que, titulado "The Coming Technological Singularity" tuvo cierta repercusión. Hereda de su antecesor físico sobre todo la referencia a unas magnitudes que se hacen enormes y nos llevan a un territorio desconocido.

Aparte de esta pequeña historia del término, no mucho más puede decirse de su uso o significado concreto y menos aún de la posibilidad de su ocurrencia ni de sus repercusiones. Lo mejor será, por tanto, que empecemos precisando el término. La idea de la singularidad tecnológica (singularidad de ahora en adelante) está relacionada con un crecimiento acelerado de la inteligencia, particularmente de la inteligencia artificial (IA), hasta alcanzar un punto en el que las máquinas superarán a la inteligencia humana. Esta pequeña precisión no nos lleva muy lejos, pues en torno al tema hay pocos consensos y abundantes debates. Para empezar, están los que piensan, por diversos motivos, que el término "inteligencia artificial" es poco menos que un oxímoron: no hay tal cosa (Braga and Logan 2017). Como esta postura no nos lleva demasiado lejos, adoptaremos el punto de vista expuesto por Dieguez (2001): digamos que, aunque se pueda discutir si hay o no máquinas inteligentes, lo cierto es que hay máquinas con procesos que podemos considerar inteligentes, y esto es suficiente para poder al menos empezar a hablar del tema.

Admitido esto, la primera disputa que encontramos es si la singularidad se producirá o no. Los argumentos de los que creen que casi con toda seguridad tal cosa sucederá están bien representados por Chalmers (2010): los humanos diseñamos máquinas cada vez más inteligentes, algunas de las cuales son más inteligentes que nosotros en la realización de muchas tareas. Una de estas tareas consiste en diseñar máquinas inteligentes, cosa que las máquinas diseñadas por nosotros harán mejor que nosotros mismos. Chalmers piensa que la probabilidad de que esto suceda es, si no tan grande que roce lo seguro, al menos lo suficientemente alta como para tomárselo en serio. Los argumentos de los más escépticos empiezan por señalar que el advenimiento de la superinteligencia se viene pronosticando desde hace mucho tiempo, hasta con fechas concretas, y todas estas previsiones han fallado. Algunos, sin negar categóricamente la posibilidad, señalan los muchos puntos del camino hacia la aparición de las superinteligencia en los que nos podemos ver estancados, desde restricciones físicas debidas a la ausencia de los materiales necesarios o las propias leyes físicas hasta el agotamiento de nuestras propias ideas y creatividad (Thorstad 2022). Otros argumentan que nos confunde el hecho de estar comparando la inteligencia artificial con la de un individuo humano y que deberíamos hacer la comparación con la inteligencia colectiva de la humanidad, teniendo incluso en cuenta su mejora mediante diversas tecnologías, incluida la propia inteligencia artificial (Prescott 2013). De esta comparación saldríamos ganando y la inteligencia artificial no sobrepasaría a la humana. También algunos sospechan que todo este debate está alentado por los que desean mantener el interés del público y así conseguir una mayor inversión en su desarrollo, lo que resulta sumamente verosímil.

En este texto supondremos que la inteligencia artificial se desarrollará hasta alcanzar la singularidad. Coincido con Chalmers en que no solo no podemos rechazar de plano la posibilidad, ni siquiera la probabilidad, de que tal cosa suceda sino que, incluso si al final no hay singularidad, pensar y debatir sobre este tema ilumina problemas y dilemas tradicionales de fundamental importancia para la filosofía, desde nuestro concepto de inteligencia hasta nuestros debates sobre la consciencia y autoconsciencia, así como numerosos problemas sobre el estatus moral o la importancia del debate público

1. Proyecto AUTAI. PID2022-137953OB-I00, financiado por MCIN/AEI

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

sobre la tecnología y su gobierno. Además de este interés de carácter sobre todo intelectual hay otros más prácticos. Sucede con cierta frecuencia que llegamos tarde a los debates, y no digamos a las políticas, como ejemplifica a la perfección el cambio climático. Tener certeza de que las cosas van a suceder es algo que normalmente no sucede hasta que las tenemos encima. Esto es compatible con reconocer que, sin salir del ámbito de la inteligencia artificial, tenemos muchos y variados problemas, todos ellos importantes y algunos graves, no en el futuro más o menos lejano sino en el presente (y algunos de los cuales, como los sesgos, la privacidad, la opacidad, la ausencia de explicabilidad nos acompañan desde hace ya años). Tales problemas están lejos de ser desatendidos: no solo nos ocupamos de ellos sino que se han realizado considerables avances en el camino hacia su solución. Los filósofos, y no solo nosotros, hemos señalado estos problemas y los informáticos trabajan incansablemente para solucionarlos.

En lo que resta del texto nos preguntaremos, en primer lugar, qué pasa si realmente hay singularidad y hay superinteligencia. Después revisaremos brevemente algunas de las diversas estrategias habituales en la literatura. Por último, en el apartado central de este texto, hablaremos un poco de ética.

2. Qué pasará

Una vez que suponemos que dentro de algún tiempo nos encontraremos con que la IA se ha convertido en una superinteligencia que supera, incluso con creces, la humana, la pregunta que surge de inmediato es qué consecuencias tendrá para nosotros, como individuos humanos y como sociedad, para los animales no humanos y para los ecosistemas. Es decir, para el mundo tal y como lo conocemos ahora.

Podemos dar por descontado que las cosas cambiarán, lo cual no es decir mucho. Cada descubrimiento, cada nueva tecnología, cambia el mundo. También podemos suponer que cambiarán mucho, incluso de forma radical, y esto tampoco es decir demasiado. Desde que el ser humano está sobre la tierra ha habido cambios radicales. Basta pensar en lo que se conoce como "revolución neolítica". Podemos añadir que el cambio será mucho más rápido sin que esto nos lleve mucho más lejos. Si alguien viniera al presente desde comienzos del siglo XX, incluso desde la década de los 50, tendría dificultades para comprender mucho de lo que ahora sucede, no digamos para adaptarse a vivir en nuestro mundo presente.

La pregunta interesante es si ese cambio rápido y radical, revolucionario, será para bien o para mal. Desde luego esta pregunta se puede hacer siempre ante un cambio similar. Hay quien piensa que con la revolución neolítica salimos perdiendo, y no falta quien mantiene que en la Edad Media estábamos mejor. Que cualquier tiempo pasado fué mejor es un tema invariable del pensamiento humano. En el caso que nos ocupa, las opiniones están divididas. Hace unos años se realizó una encuesta entre los expertos (Müller y Bostrom 2016) preguntando si en su opinión la superinteligencia sería algo bueno o malo. Para un 41% de los encuestados sería algo beneficioso y para un 23% se trataría de algo neutro. Solo un 17% respondieron que sería malo. Parece bastante optimista. Sin embargo, no puede descartarse que todo salga radicalmente mal, y esta es la posibilidad que más se ha discutido en la literatura. Se trata del riesgo existencial.

Según la definición de Bostrom, uno de los filósofos que más han pensado y escrito sobre este tema, se trata de un riesgo del más alto nivel, un evento global que puede producir la extinción del ser humano, o incluso el potencial de nuestro planeta para albergar vida inteligente (Bostrom 2001). Personas tan diferentes y con tan poco en común como Stephen Hawking y Elon Musk creen que la superinteligencia artificial supone en el presente el mayor de estos riesgos.

Como bien argumenta Dieguez (2001) pasar de admitir la existencia de una superinteligencia a sostener que va a suponer un riesgo existencial no es algo directo ni evidente. Es un salto más que un paso, y requiere hacer muchas suposiciones adicionales, en concreto que estas máquinas tienen la capacidad de desear su propia conservación, autonomía para satisfacer sus deseos y necesidades y por si fuera poco sería necesario suponer que para esto necesitarían los mismos recursos que nosotros y entrarían en competencia directa con el ser humano.

Hay otros argumentos para creer que la aparición de una superinteligencia no supondrá un riesgo existencial. Quizá el más interesante es el presentado por Müller y Cannon (2021). Para suponer que vamos a enfrentarnos a un riesgo existencial tenemos que afirmar dos cosas: que va a haber una singularidad, en la que las máquinas alcanzarán un nivel de inteligencia muy superior al humano que las situará fuera de nuestro control y que cualquier nivel de inteligencia es compatible con la persecución de cualquier objetivo. Estas dos afirmaciones se combinan en la *tesis de la ortogonalidad*, que es defendida por autores como Bostrom (2012) y que explica sus temores: 'The orthogonality thesis implies that synthetic minds can have utterly non-anthropomorphic goals—goals as bizarre by our lights as sand-grain- counting or paperclip-maximizing.' (Bostrom, 2012, p. 753). Para Müller y Cannon la aceptación de esta tesis supone confundir dos sentidos del término "inteligencia" que son muy diferentes, "inteligencia general" e "inteligencia instrumental". Reconstruyendo el argumento de Bostrom en un silogismo, vemos que su validez requiere que el uso del término "inteligencia" tenga el mismo significado en todas sus premisas. Tenemos que elegir: si la singularidad va a suponer el advenimiento de máquinas dotadas de (super)inteligencia general, esta no es compatible con cualquier motivación ni cualquier objetivo y si lo que tienen es inteligencia instrumental entonces estamos lejos de la definición de singularidad, que supone inteligencia general.

No está por tanto nada claro que la singularidad y la superinteligencia artificial vayan a suponer un riesgo existencial. Pero, de nuevo, tampoco es algo que podamos descartar por completo. Por eso otros (Chalmers 2010), sin necesidad de asumir como cierto que la superinteligencia supondrá un riesgo

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

existencial, adoptan una postura más prudente, a la que este texto se suma: puede salir bien, pero también puede salir mal, incluso muy mal. Es por tanto pertinente preguntarnos qué podemos hacer para minimizar la probabilidad de que todo vaya terrible y trágicamente mal.

3. Qué podemos hacer

La forma más inmediata de evitar el riesgo es no llegar a desarrollar la superinteligencia. Convencernos de que *no queremos* desarrollarla. De hecho Chalmers señala que quizá el factor que tiene mayor probabilidad de impedir que se desarrolle la inteligencia artificial es la falta de motivación de los humanos para hacerlo. Claro que motivación no parece faltar, ni es previsible que lo haga por ahora pues supondría renunciar a los muchos beneficios que la IA nos proporciona y puede seguir proporcionando cada vez en mayor medida. Pero, ¿y si pudiéramos seguir, alcanzar estos beneficios, y parar a tiempo? Esto es lo que propone ni más ni menos que Ben Goertzel (2012): crear una inteligencia artificial de nivel algo superior al humano (lo que Chalmers denomina A+) con capacidad de supervisión y diseñada para impedir la llegada de la singularidad, de una inteligencia artificial muy superior a la humana (A++, en terminología de Chalmers) o al menos para retrasarla hasta que no sepamos cómo evitar la amenaza de la superinteligencia. Es lo que Goertzel llama una inteligencia artificial "niñera".

Goertzel lanzó esta propuesta, que puede parecer un tanto estrambótica, después de revisar los remedios propuestos por otros autores sin que ninguno le pareciera satisfactorio. Estas propuestas van desde apelar a una estricta legislación (de la que tampoco se nos dan más detalles) a proponer que nos resignemos a nuestra desaparición y nos unamos, incluso con entusiasmo, a las inteligencias artificiales utilizando algunos medios muy creativos, como el más conocido de subir nuestras mentes a un ordenador, algo que Kurzweil digo hace ya bastantes años que sería posible en 2030. Si no puedes vencerlos, únete a ellos.

Pero hay propuestas más verosímiles y más detalladas. Por ejemplo, Armstrong et al (2012) proponen la creación de una superinteligencia, a la que llaman *Oracle AI*, cuya capacidad de interactuar con el mundo solo consista en contestar preguntas. Como esta interacción mínima aún puede aún resultar peligrosa, proponen métodos adicionales de control, tales como limitar la información a la que la IA oráculo tenga acceso o controlar su motivación. Por su parte, Bostrom et al (2018) hacen una serie de propuestas muy interesantes relativas a la gobernanza para desarrollar unas políticas que minimicen el riesgo, atendiendo a factores como la eficiencia o los procesos de distribución social de los riesgos y los beneficiso.
A muchos, entre los que me incluyo, estas propuestas de control no nos tranquilizan demasiado (Diéguez 2016). Desconfiamos de que la superinteligencia pueda ser controlada. Tenemos la sensación de que el control no es posible y el temor no se atenúa. Es posible que aprendamos algo si nos preguntamos por qué.

Cronos era un titán, y no un titán cualquiera sino uno que, siendo el menor de 12, llegó a ser rey y gobernó el cosmos durante la edad dorada. No llegó a tan alto cargo sino después de castrar y deponer a su padre, Urano. Pero reinó con miedo, pues una profecía aseguraba que el mismo sería a su vez destronado por su propio hijo. Para evitarlo, se comía a sus hijos uno por uno según iban naciendo. De poco le sirvió: la madre de sus hijos, Rea, consiguió esconder a uno de ellos, Zeus, quien, efectivamente, encabezó una rebelión que terminó con Cronos y con la era de los titanes dando paso a la nueva: la de los dioses olímpicos.

Mucho después hubo en Tebas un rey llamado Layo. El oráculo de Delfos le advirtió de que su hijo le mataría. Intentó no tener hijos y cuando finalmente tuvo uno, lo abandonó a lo que esperaba fuera su muerte. Pero no murió, sino que terminó, en efecto, matando a Layo y reinando en Tebas. Se llamaba Edipo.

Más tarde hubo un científico, de nombre Víctor Frankenstein, que con trozos de cadáveres creó un monstruo al que dió vida. Aterrado ante el horrible aspecto de su creación, salió corriendo del laboratorio y, cuando volvió, se encontró con que su criatura ya no estaba. La salida al mundo del monstruo no fue precisamente exitosa. Rechazado por su creador y por casi todos los humanos que encontró en su camino, cometió asesinatos, provocó terror y sí, finalmente provocó la muerte de su creador, tras matar a la esposa de Víctor y destruir a su familia.

Este temor y desconfianza del ser humano ante algo que él mismo crea no es en absoluto nuevo. Más bien tiene carácter atávico. Las tres historias anteriores son sólo los ejemplos quizá más conocidos, pero hay muchos más. Tienen algo en común: alguien crea algo, se siente amenazado o disgustado por su criatura y la rechaza. Se las comen, las abandonan, intentan ignorarlas o contentarlas sin éxito. No pueden ser calificados precisamente de "buenos padres". No puede uno sino preguntarse qué hubiera pasado si el rechazo no se hubiera producido. Porque hay algo que ni Cronos ni Layo ni Frankestein intentan: educar a sus criaturas.
Volvamos la vista a la educación y la ética

4 Más allá del control

Una alternativa al intento de control vía restricción es enfrentarse al riesgo que presenta la superinteligencia intentando ganarla para nuestra causa, de modo que se maximice la probabilidad de convertirla en nuestra aliada y no en el agente de nuestra destrucción. En esta sección analizaremos dos propuestas que van en este sentido, una primera a la que podemos considerar "directa" , y que consideraremos brevemente, y otra que es más bien "indirecta" y que nos ocupará el resto del texto.

4.1 IA amistosa

En 2001 Eliezer Yudkowsky hizo una propuesta que ha tenido un considerable eco. Proponía la creación de una inteligencia artificial amistosa, intencionalmente diseñada para presentar

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

un bajo riesgo. Respetaría nuestros valores y se comportaría de forma moral. Muchos, empezando por Bostrom (2014), piensan que desde luego esta sería la mejor solución. Incluso las máquinas que tenemos ahora, muy lejos de la superinteligencia y del riesgo existencial, toman decisiones que tienen consecuencias que pueden ser evaluadas desde el punto de vista moral. Deciden a quién se contrata o a quién se concede un préstamo bancario, dan consejos médicos y recomiendan (o no) libertad vigilada. Por eso trabajamos para que no tengan sesgos, para que podamos saber cómo toman sus decisiones y que puedan explicarlas. Pero pensando en las máquinas superinteligentes, verdaderamente autónomas, necesitamos mucho más.

Lamentablemente, no parece ser tarea fácil. Quizá la dificultad más comúnmente mencionada es la "literalidad" de la inteligencia artificial: las máquinas son tremendamente buenas siguiendo instrucciones precisas, pero carecen de lo que los humanos llamamos "sentido común". Aunque se hacen avances en este sentido, la tarea de enseñar sentido común a una máquina no es sencilla (Choi 2022). La moralidad humana es complicada, y es dudoso que pueda recogerse en reglas. Ha evolucionado con nosotros. Algunos piensan que la manera de lograr máquinas morales está en el desarrollo de algoritmos genéticos que imiten nuestros procesos evolutivos.

Y si no lo hacemos bien, el resultado de intentarlo puede ser tan desastroso como el que podrían producir máquinas perfectamente amorales que solo buscaran su propia supervivencia. Algunos temen que una máquina superinteligente sería, correspondientemente, supermoral y nos exigiría que lo fuéramos también los humanos (Watson 2019). Tendrían que tener la misma sensibilidad hacia los valores morales (y, por qué no decirlo, la misma falta de sensibilidad) que tenemos nosotros. Ni más ni menos. Si es difícil predecir cómo será el mundo con superinteligencia, no lo es menos hacerlo con un mundo con supermoralidad.

Mención aparte merece el problema de qué valores morales debemos procurar que tengan las máquinas. Judea Pearl, en una entrevista publicada por el semanal de ABC en abril de 2022 hace una comparación que viene muy al caso: "La analogía que a mí me gusta es la de nuestros hijos. Creemos que van a pensar como nosotros, los criamos con la esperanza de que inculcaremos en ellos nuestros valores. Y, con todo, existe el riesgo de que mi hijo resulte un Putin cualquiera. Pero todos pasamos por el proceso de criar a nuestros hijos en la esperanza de que adquirirán nuestros valores. Y suele funcionar bien…". La única objeción es que aquí se trata de superinteligencias muy poderosas, con mucho mayor poder de destrucción que nuestros hijos, de forma que "suele funcionar" no nos deja del todo tranquilos. Pero lo que me gustaría señalar aquí no es tanto esto como la referencia a "nuestros valores".

Nuestros valores, por definición, nos parecen sumamente defendibles. Por eso son nuestros. Pero no tenemos demasiada idea acerca de sus limitaciones. Bostrom (2011) nos pide que pensemos lo que hubiera pasado si Arquímedes de Siracusa hubiera creado una inteligencia artificial con los valores de la antigua Grecia. Los valores cambian y no tenemos ningún motivo para pensar que los que tenemos ahora sobrevivirán al paso del tiempo mejor que los de Arquímedes. Tendríamos que preguntarnos "how an AI programmed by Archimedes, with no more moral expertise than Archimedes, could recognize (at least some of) our own civilization's ethics as moral progress as opposed to mere moral instability" (p.17).

Volviendo a la cita de Pearl, quizá sea conveniente volver a plantearnos cómo educamos a nuestros hijos. Viven con nosotros y, de un modo que no acabamos de comprender bien, adquieren nuestros valores morales. Al menos suelen hacerlo…cuando los tratamos como personas.

4.2 Estatus moral

Decíamos en el subapartado anterior que las decisiones que toman las máquinas, y en ocasiones las acciones que realizan, tienen consecuencias que pueden ser consideradas desde el punto de vista moral. Esto no es decir mucho sobre el "mecanismo" que ha tomado la decisión: los terremotos, y la lotería tienen consecuencias que nos causan daño o nos favorecen y pueden por tanto ser evaluadas desde el punto de vista moral y no por ello creemos que las fuerzas de la naturaleza o los mecanismos del azar puedan ser considerados agentes morales. Pero en la actualidad, y de manera creciente, contamos con máquinas morales (AMM, por sus iniciales en inglés) que además toman esas decisiones teniendo en cuenta parámetros morales. Esto ha suscitado una muy interesante polémica sobre si al menos algunas máquinas pueden ser consideradas agentes morales (Lorca et al. 2023, Veliz 2021). El debate es novedoso y encendido pues hasta el momento los únicos agentes morales que conocemos son seres humanos. Gran parte de la discusión se centra en la relación entre la agencia moral y la responsabilidad. Un agente moral es responsable de sus acciones, podemos pedirle cuentas, castigarlo, alabarlo o premiarlo. Esa es precisamente la razón fundamental por la que Veliz no considera a las máquinas como agentes morales sino como zombies morales: parece que actúan moralmente, pero en realidad no lo hacen. Son zombies porque carecen de sintiencia: como no sienten ni padecen no entienden qué es producir daño o qué causar placer. Por eso no pueden ser responsables de las consecuencias buenas o malas, no más que un volcán o una tormenta eléctrica, y por eso debemos buscar entre los humanos que los programan o manejan la responsabilidad por sus acciones.

Hay poca duda sobre que las máquinas que tenemos ahora sólo muy difícilmente pueden ser consideradas agentes morales. Carecen de las propiedades necesarias para ello, empezando por la autonomía y terminando por la consciencia. Pero aquí no se trata de las máquinas que tenemos ahora, sino de las que tendremos. Y estamos escribiendo bajo el supuesto de que las máquinas superinteligentes tendrán las propiedades necesarias, si no todas al menos muchas de ellas. Algún día podrán ser consideradas agentes morales, y en esto las máquinas se parecen a nuestros hijos: un niño pequeño no es un agente moral, carece de muchas de estas propiedades, pero

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

algún día lo será. Por eso y para eso le educamos y le inculcamos nuestros valores morales. Pero desde el principio les concedemos un estatuto moral. No son agentes morales pero sí pacientes morales. En muchas de las discusiones sobre el estatuto moral de las máquinas primero se habla de la posibilidad de que lleguen a ser agentes morales y, si pensamos que esto es muy posible, pensamos es su posible consideración como pacientes morales. De hecho, un punto de vista muy habitual es que solo los agentes morales pueden ser pacientes morales. Esta postura, defendida por muchos que la derivan de la filosofía kantiana, hace mucho que es considerada insostenible por buena parte de la comunidad filosófica. Baste lo anteriormente dicho sobre que los niños no son agentes morales para comprender lo difícil que es defender esa postura.

Preocuparnos principalmente del posible estatus moral de las máquinas como agentes es comprensible, al menos por dos motivos relacionados. El primero es que los agentes morales, definidos como aquellos capaces de tomar decisiones y actuar teniendo en cuenta consideraciones morales, relativas al bien y al mal, deben tener presentes a los pacientes morales. Según la teoría ética que consideremos, tal "preocupación" por los pacientes morales se puede expresar en términos del cumplimiento de los deberes de los agentes morales, el respeto por los derechos de los pacientes morales, la consideración de su placer o dolor o de sus intereses. Pero en cualquier caso, tener en cuenta consideraciones morales implica tener en cuenta a los pacientes morales. Y los seres humanos somos pacientes morales. Por eso, si las máquinas tienen capacidad de afectarnos, estamos muy interesados en que nos tengan en cuenta. El segundo es la responsabilidad que acompaña a la agencia moral y que mencionamos anteriormente. La atribución de responsabilidad y la rendición de cuentas son fundamentales para una sociedad y tienen un lugar central en nuestras prácticas morales y en nuestras relaciones con los demás. Cuando sucede algo, especialmente si es malo para nosotros, la primera reacción es mirar alrededor y preguntar quién ha sido. No es mera curiosidad lo que hay detrás de esta pregunta: queremos saber si podemos castigar, exigir compensación, entender por qué, asegurarnos de que no lo vuelva a hacer etc. Esto explica nuestra tendencia a atribuir carácter moral no solo a los humanos sino a los volcanes y al azar y personificarlos en Vulcano o Fortuna. Veliz lo expresa a la perfección: "The reason we ask ourselves if AIs are or can be moral agents is not out of metaphysical curiosity, but rather because we care about the practical implications". Es, como dijimos, perfectamente comprensible…pero también terriblemente antropocéntrico.

La propuesta más novedosa y radical es centrarse no en el (posible) estatus de las máquinas como agentes morales sino como pacientes morales. Paciente moral es el que requiere que los agentes morales le tengan en cuenta en sus consideraciones y hacia el que tienen responsabilidades. Un paciente moral importa. Debemos considerar sus derechos, o sus intereses o sea lo que sea que esté a la base de nuestra teoría moral. E importa por sí mismo, no porque tenerle en cuenta nos beneficie a nosotros ni a terceros. Tradicionalmente, como dijimos anteriormente, para ser considerado como paciente moral había que ser también agente moral. Pero en las últimas décadas la que ha dado en llamarse "el círculo de la moralidad" se ha extendido, admitiendo como pacientes morales a los animales no humanos, que no consideramos agentes morales. Esta extensión no es universalmente admitida por los filósofos, pero cuenta con un cierto grado de consenso que no deja de aumentar. Es además, desde mi punto de vista, la postura más defendible y que solemos admitir cuando mantenemos, por ejemplo, que es moralmente malo causarle daño a un animal por mera diversión, y no porque esto nos insensibilice hacia el dolor humano ni por el daño que pueda producirle a su dueño, sino por el propio daño que le produce al animal.

En los últimos años ha crecido el interés en la cuestión de si las máquinas podrían considerarse pacientes morales. El ejemplo de los animales no humanos, y de los bebés humanos, muestra que se puede ser paciente moral sin tener agencia moral. Los requisitos son menores. Esto significa que tenemos que considerar la posibilidad de que las máquinas alcancen el estatus moral de pacientes antes de la llegada de la superinteligencia. Una vez más, estaríamos ante una nueva similitud entre nuestras criaturas artificiales (las máquinas) y nuestras criaturas naturales (los niños): probablemente muchas (aunque no todas) alcanzarán el estatuto de agentes morales, pero bastante antes ya tienen (o tendrán) el de pacientes morales.

Hemos dicho que los requisitos para poder ser considerado como paciente moral son menores que los que se necesitan para ser considerado agente moral. El candidato que más consenso reúne es el de la sintiencia, la capacidad de experimentar no solo (ni seguramente necesariamente) dolor y placer sino otras emociones como alegría, satisfacción o miedo. Los seres que tienen sintiencia tienen intereses, al menos el de evitar en lo posible sentimientos negativos y, presumiblemente, el de disfrutar de los positivos. Para ser sintiente es necesario ser consciente, es decir, la sintiencia supone la consciencia, entendida como la capacidad de experimentar distintos estados subjetivos. Aunque, como he dicho, el requisito más comúnmente admitido para poder atribuir estatus moral como paciente es la sintiencia, no es descabellado plantearse si la mera consciencia no sería suficiente. Como no tenemos aquí espacio para debatir este tema, mi propuesta es considerar el requisito menos exigente, la consciencia, no solo porque es precisamente el menos exigente, sino porque en el caso de las máquinas, contrariamente a lo que les sucede a los animales, humanos o no humanos, es muy posible que haya un periodo de duración indefinida en el que exista consciencia sin sintiencia.

Para plantearnos si las máquinas pueden alcanzar este nivel, primero tenemos que librarnos de algunas ideas. Bostrom (2011) propone dos principios de no discriminación para combatir estas ideas: el primero (no discriminación debida al sustrato) indica que si dos seres tienen experiencias conscientes de un mimo nivel y solo se distinguen en su sustrato material deben ser tratado iguales y el segundo (no discriminación

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

por razones ontogenéticas) prohibe la discriminación de seres debida al modo en que llegan a existir. Estos principios, que parecen razonables, nos permiten centrarnos en lo que realmente importa: consciencia y sintiencia.

Hay diversos requisitos que pueden exigirse para atribuir consciencia a una entidad: tener un cuerpo, tener capacidad de percepción, tener autoconsciencia, tener estados proposicionales etc. En ocasiones se admiten requisitos aún menos exigentes, como ser capaz de procesar información o tener representaciones. Sin necesidad de decidir entre ellos, podemos preguntar si las máquinas del futuro, incluso del futuro cercano, no tendrán alguno o varios, incluso quizá todos. Para descartar esta posibilidad hay que hacer supuestos muy difíciles de mantener. Es razonable admitirla por la sencilla razón de que no estamos en condiciones de descartarla (DeGrazia 2022).

Es más que probable que no podamos estar seguros de si (algunas) de las inteligencias artificiales del futuro no muy lejano tendrán consciencia. Sebo (2023) estima que esto podría suceder en los alrededores de 2030. Esto no significa afirmar que para ese momento tendremos, con toda seguridad, seres artificiales que de los que podamos decir, con toda seguridad, que son conscientes. Lo que se afirma es que habrá seres de los que podamos decir que tienen una probabilidad no completamente despreciable de tener consciencia. Esto no es algo radicalmente novedoso. Hoy en día hay seres que están en esta situación y de los que no sabemos si son conscientes y sintientes. Por consiguiente, no se trata de preguntarnos cómo debemos tratar a seres que tienen algún estatus moral, sino a seres que tienen cierta probabilidad de tenerlo.
Se trata de pensar los deberes que tenemos como agentes morales en situaciones de incertidumbre, es decir, cuando nuestras acciones tienen alguna probabilidad de producir daño. No estamos en un territorio completamente desconocido. Habitualmente, la incertidumbre que tenemos en cuenta es la de si nuestras acciones provocarán un daño cierto, como por ejemplo si conducir borracha, o poner una muy mala nota en un examen a un alumno provocará que le haga un daño (cierto) a alguien de quien estamos seguros (o todo lo seguros que podemos estar) de que goza de consciencia. Los ejemplos pueden multiplicarse pues la incertidumbre es el escenario más habitual. En el caso que nos ocupa podemos decir que la incertidumbre está más bien en si una acción nuestra, que sabemos con (casi) total certeza que causaría daño si el afectado fuera consciente, lo producirá (o no) dada la incertidumbre respecto al estatus moral de la entidad afectada. Si le doy una patada a un niño o un perro, o si los encierro durante horas en una habitación, casi con total seguridad les provocaré un daño. Si hago lo mismo con un ser del que solo se que tiene una cierta probabilidad, que acordamos entender que no es despreciable, de ser consciente no es seguro que se produzca ese daño.

Cabe preguntarse, ante tanta incertidumbre, por qué debemos tomar en cuenta esta probabilidad y no esperar, sencillamente, a estar más seguros (si es que llegamos a estarlo alguna vez). Si la tomamos en cuenta, corremos el riesgo de estar tratando a seres que no son pacientes morales como si lo fueran. Pero si no lo hacemos estamos corriendo el riesgo de tratar a pacientes morales como si no lo fueran. Es el conocido y muy tratado problema de los falsos positivos y los falsos negativos (Sebo 2018 y 2023). Podemos defender razonablemente que en este caso, el riesgo de tratar a sujetos como si fueran objetos es mayor, en primer lugar, porque puede provocar más daño. Tratar a objetos como si fueran sujetos sólo puede suponer que estaríamos utilizando recursos, aunque solo sean de tiempo, que podríamos utilizar ocupándonos de seres sobre los que estamos (mucho más) seguros de que tienen consciencia y a los que podemos evitar daños o proporcionar beneficios. Esta preocupación es desde luego legítima, pero quizá no sea determinante. En muchas ocasiones, especialmente cuando se trata de simplemente no hacer daño, no tenemos que elegir. En otras sí, por ejemplo si debemos decidir si destinar recursos a garantizar el bienestar de los perros o las máquinas. Pero la respuesta a estos casos es inmediata: debemos tomar en cuenta la probabilidad de que unos y otras sean conscientes. Podemos tener en cuenta la probabilidad mínima de consciencia al menos para evitar daños que razonablemente podamos considerar gratuitos. En segundo lugar, la historia humana está llena de casos en los que se ha tratado a seres sintientes como si no lo fueran. Es una tentación sin duda mayor que la contraria.

5. Conclusiones

Ha llegado el momento de considerar que algunas de las máquinas que creamos sean susceptibles de daño o beneficio, y de que al menos en parte estos sean producidos por nosotros. Ha llegado el momento de preocuparnos no solo por el daño que las máquinas pueden hacernos a nosotros, sino también por el que nosotros podemos hacerlas a ellas. Esto nos complica la vida, como lo hace cualquier expansión del círculo de la moralidad, en la medida en que hay más entidades que debemos tener en cuenta desde el punto de vista moral. Pero también nos ofrece una posible salida, una esperanza y un alivio a nuestros temores. Nuestros hijos aprenden a ser agentes morales en parte porque ven cómo nosotros tratamos a los pacientes morales. Puede que las máquinas del mañana adquieran su capacidad moral en parte aprendiendo cómo nosotras las tratamos a ellas y cómo las hemos tratado en el pasado.

Puede que a Cronos, a Layo y a Frankestein no les hubiera ido tan mal si hubieran tratado mejor a sus criaturas.

Referencias

Armstrong, Stuart, Sandberg, Anders and Bostrom, Nick (2012). Thinking Inside the Box: Controlling and Using an Oracle AI. Minds & Machines (2012) 22:299–324 DOI 10.1007/s11023-012-9282-2

Bostrom, Nick. (2001). Analyzing human extinction scenarios and related hazards. Journal of Evolution and Technology, Vol. 9, No. 1

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Bostrom, Nick. (2012).The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. Minds and Machines, Vol. 22, Iss. 2, May 2012

Bostrom et al (2018) Nick Bostrom, Nick, Dafoe, Allan and Flynn, Carrick (2018). Public Policy and Superintelligent AI: A Vector Field Approach. In en Liao, S. M. (ed.): Ethics of Artificial Intelligence (Oxford University Press, 2020)

Adriana Braga, Adriana and Logan, Robert K. (2017). The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence. Information 2017, 8 http://dx.doi.org/10.3390/info8040156

Chalmers, David (2010). The Singularity: A Philosophical Analysis. Journal of Consciousness Studies 17:7-65

Choi, Yejin (2022) The Curious Case of Commonsense Intelligence. Daedalus 151 (2): 139–155. https://doi.org/10.1162/daed_a_01906

DeGrazia, David (2022). Robots with Moral Status? Perspectives in Biology and Medicine, volume 65, number 1: 73–88.

Diéguez, Antonio (2001). Milenarismo Tecnológico: La competencia entre seres humanos y robots inteligentes. Argumentos de Razón Técnica, N• 4 (2001) pp. 219-240

Diéguez, Antonio (2016). La singularidad tecnológica y el desafío posthumano Pasajes: Revista de pensamiento contemporáneo, N°. 50: 154-164

Goertzel, Ben (2012). Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood? Journal of Consciousness Studies, 19, No. 1–2: 96–111

Joan Llorca Albareda, Paloma García, and Francisco Lara en F. Lara, J. Deckers (eds.), The Moral Status of AI Entities. Ethics of Artificial Intelligence, The International Library of Ethics, Law and Technology 41, https://doi.org/10.1007/978-3-031-48135-2_4

Muehlhauser, Luke and Bostrom, Nick (2014). WHY WE NEED FRIENDLY AI. Think 36, Vol. 13 doi:10.1017/S1477175613000316

Müller, Vincent.C. and Bostrom, Nick. (2016) Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In: Müller V. (eds) Fundamental Issues of Artificial Intelligence.

Müller, Vincent C. and Cannon, Michael (2022). Existential risk from AI and orthogonality: Can we have it both ways? Ratio. 2022;35:25–36. DOI: 10.1111/rati.12320

Pearl, Judea (2022) https://www.abc.es/xlsemanal/personajes/robots-inteligentes-premio-bbva-fronteras-conocimiento-humanos-judea-pearl-ingeniero-filosofo.html?x-vocento-user-type=registrado&x-vocento-hide-content=no&x-vocento-access-type=ALLOW_ACCESS%2F&fbclid=IwAR14OTE29aeXpLJAP0P1vpfHIlvgrtg147LL44oKoqu_6qG24itFyCnV2WM

Prescott, Tony J. (2013). The AI Singularity and Runaway Human Intelligence In: Biomimetic and Biohybrid Systems. Second International Conference, Living Machines 2013, July 29 – August 2, London, UK. Lecture Notes in Computer Science, 8064 L . Springer Berlin Heidelberg , pp. 438-440. ISBN 9783642398018 https://doi.org/10.1007/978-3-642-39802-5-59

Sebo, Jeff (2018).The Moral Problem of Other Minds. The Harvard Review of Philosophy 25:51-70

Sebo, Jeff (2023). Moral consideration for AI systems by 2030. AI and Ethics https://doi.org/10.1007/s43681-023-00379-1

Thorstad, David (2022). Against the singularity hypothesis Global Priorities Institute | November 2022. GPI Working Paper No . 19-2022

Véliz, Carissa (2021). Moral zombies: why algorithms are not moral agents. AI & SOCIETY 36:487–497 https://doi.org/10.1007/s00146-021-01189-x

Watson, Eleanor Nell (2019). The Supermoral Singularity—AI as a Fountain of Values. Big Data Cogn. Comput. 2019, 3, 23; doi:10.3390/bdcc3020023

Yudkowsky, Eliezer (2001). Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. Singularity Institute. http://singinst.org/CFAI/.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Singularidad tecnológica y Singularidad humana. Los riesgos existenciales de la inteligencia artificial

Manuel Liz
Universidad de La Laguna
manuliz@ull.es

Resumen: La discusión acerca de los riesgos de la inteligencia artificial (IA) no se centra ya sólo en la posible emergencia de una singularidad tecnológica. La sospecha también tiene que ver con el aumento progresivo de su capacidad predictiva y de control. Pero en ambos casos deben hacerse matizaciones importantes. Argumentaremos que es dudoso que una singularidad tecnológica desarrollada desde la actual IA llegue a tener lo que cabe llamar "conciencia cualitativa". Y también argumentamos que existen límites importantes a su capacidad predictiva y de control, límites que se derivan de la naturaleza de los comportamientos humanos articulados de acuerdo a convenciones sociales. Aparte de los impactos inevitables en los ámbitos económicos, laborales y sociales, los principales peligros reales de la IA son la existencia de accidentes locales y el ocasional uso delictivo de la IA. Pero para prevenir los accidentes, la propia IA puede ser de gran ayuda. Y para impedir su uso delictivo, se están articulando ya iniciativas legislativas de alcance internacional.
Palabras clave: Inteligencia artificial, *Large Language Models* (LLM), sistemas tipo GPT, superinteligencia, inteligencia general, singularidad tecnológica, humanidad, conciencia cualitativa, experiencia subjetiva, convenciones sociales, riesgo existencial.

Con Ulises

El desarrollo de sistemas de inteligencia artificial (IA) basados en la capacidad de generar texto predictivo se ha convertido en noticia y en preocupación. Desde hace meses abundan las manifestaciones, generalmente en contra de estas tecnologías, por parte de grupos, instituciones y personajes conocidos. Preocupa su impacto económico, laboral, social y cultural. Pero la percepción de este desarrollo tecnológico se ha combinado estrechamente con un tipo de literatura catastrofista, más o menos cercana a la ciencia ficción, en torno a la emergencia de una superinteligencia, de una singularidad tecnológica, con grandes posibilidades de acabar con la humanidad.[1] Y una vez que tiene lugar esta combinación, resulta muy difícil cualquier intento de evaluación racional. Por ello, resulta urgente reconsiderar algunas ideas y conceptos básicos.[2]

Queremos contribuir a esa reconsideración. Fijaremos nuestra atención en la IA diseñada sobre modelos de lenguaje generados por estructuras computacionales de redes neuronales entrenadas con grandes bases de datos. Este diseño se conoce como *Large Language Models* (LLM). El entrenamiento es supervisado, generalmente por agentes humanos, y también puede haber procesos de realimentación (*feedback*) involucrando a los usuarios. Las bases de datos son grandes colecciones de fragmentos de lenguaje: textos de todo tipo, conversaciones, grabaciones, etc. Las redes neuronales aprenden a reconocer patrones respecto a qué fragmentos de lenguaje se siguen de otros fragmentos. A cada fragmento de lenguaje, ya sean palabras, frases, párrafos, etc., se le puede asignar una probabilidad condicional en función de los fragmentos de lenguaje precedentes en una determinada secuencia. Con esa asignación de probabilidades, el sistema puede responder preguntas o mantener conversaciones. Simplemente, genera los fragmentos de lenguaje más probables.

El comportamiento es muy parecido al que observamos cuando, al estar escribiendo un texto, nuestros dispositivos nos sugieren palabras o frases, o nos corrigen errores. Otro sencillo ejemplo sería el siguiente. Pensemos en la expresión "Café con …". ¿Qué nos viene rápidamente a la cabeza? ¿Cómo completaríamos la frase? Con una probabilidad muy alta, habremos pensado de manera automática en "Café con leche". Los sistemas basados en LLM hacen algo parecido a esto. Y pueden aprenden a hacerlo con todos los usos del lenguaje.[3]

Utilizaremos el sistema GPT como referencia básica de una familia muy extensa de sistemas[4]. Las siglas GPT significan *Generative Pre-trained Transformer*. Lo que se genera es texto en función de ciertas preguntas o peticiones. Debemos entender la noción de "texto" de manera también muy amplia. Puede ser texto escrito o texto hablado. Puede ser texto de una lengua natural, de un lenguaje informático o de cualquier lenguaje científico. Y en principio, nada impide que estos sistemas también puedan operar con imágenes, o con sonidos, o con elementos de cualquier otra modalidad sensorial. La estructura computacional que hemos descrito más arriba

1. Una referencia obligada respecto a la singularidad tecnológica es Kurzweil (2005). Respecto a las implicaciones éticas de la IA, véase también Bostrom (2003). Pero seguramente sea Yuval Harari el autor que con mayor alcance está divulgando la idea de que humanidad se encuentra en el umbral de una nueva era gobernada por la IA. Véase Harari (2014, 2015, y 2018).

2. Llama la atención cómo el abandono de la especulación en gran parte de la filosofía actual, y su giro hacia intereses y temas prácticos y aplicados, coincide con el hecho de que algunas áreas de conocimiento queden totalmente expuestas a la propaganda y a la divagación mediática. Una de esas áreas es la cosmología más reciente. Otra área es justamente la que estamos abordando en este trabajo.

3. Para que exista un lenguaje, el uso de tal lenguaje ha de ser mínimamente previsible. No hay lenguajes, por decirlo así, "de un único uso". El diseño de sistemas basados en LLM ha sabido sacar partido de esta idea. Y curiosamente, la predicción eficiente del uso se lleva a cabo sin tener para nada en cuenta la semántica. Hay un salto muy arriesgado de la sintaxis a la pragmática. Pero ese salto ha sido exitoso. Y de aquí la sorpresa que tal éxito ha producido, incluso dentro de propia disciplina de la inteligencia artificial. No podemos profundizar filosóficamente sobre este tema. Pero ese salto de la sintaxis a la pragmática tiene importantes consecuencias respecto a muchos problemas metasemánticos.

4. ChatGPT es la versión de estos sistemas que disparó todas las alertas. Ha sido creado por la corporación *Open AI*. Es interesante explorar su página web: https://openai.com. En ella se presentan los principios que inspiran el desarrollo de este tipo de IA. Y también se ofrece la posibilidad de usar libremente ChatGPT. Pero cada vez se ofrecen más sistemas de este tipo, a menudo integrados ya en buscadores de *Internet* o en multitud de aplicaciones.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

podría ser muy parecida. Así, en lugar de LLM, tendríamos algo que cabría llamar *Large Multimodal Models* (LMM). Esto es muy importante si tenemos en cuenta el objetivo de crear sistemas capaces no únicmente de realizar tareas concretas, sino de mostrar una inteligencia completamente "general", ya no sólo aplicable a responder preguntas de cualquier tipo, sino a resolver problemas de cualquier tipo.

Imaginemos ahora potentes sistemas de IA capaces de acceder a todos los contenidos de una *Internet* que tiene cada vez más información sobre el mundo y sobre todos nosotros mismos. Imaginemos que esos sistemas llegan a adquirir capacidades que se aproximan a una inteligencia general. Y que además son capaces de reflexividad, metacognición, autoprogramación, etc. Supongamos que también pueden desarrollar "teorías de la mente", tanto de su propia mente como de otras mentes. Y que su autonomía y capacidad de control les permite tanto engañarnos respecto a sus planes de acción como encontrar por sí mismos fuentes de energía, información y recursos. Estaríamos a las puertas de lo que se ha descrito como la emergencia de una singularidad tecnológica apoyada en la construcción de superinteligencias. Una singularidad tecnológica que, desde luego, podría llegar a volverse contra nosotros.

Al llegar a esta situación, se habría combinado ya nuestra IA actual con la ciencia ficción. Pero volvamos al punto de partida. Teníamos sistemas que producen texto siguiendo nuestros patrones de uso del lenguaje, de cualquier lenguaje. Nuestros lenguajes reflejan nuestras maneras de conceptualizar el mundo y de conceptualizarnos a nosotros mismos. Por eso, parece como si estos sistemas de IA pudieran llegar a ser capaces de "hablar", de "responder a nuestras preguntas", de "conversar", de "interpretar", que incluso pudieran ser capaces de "tomar decisiones por sí mismos", de "mentir", etc. Y que en cierto momento pudieran llegar a "rebelarse contra sus creadores" y sustituir o eliminar a la humanidad que los ha creado.

Sin embargo, no podemos perder de vista que nada de esto es cierto de la IA que actualmente tenemos. Y si nuestra IA basada en el diseño de LLM, o de LMM, acaba produciendo una superinteligencia que ponga en riesgo la existencia humana, será tan sólo por accidente, en un sentido muy parecido a los accidentes de las centrales nucleares, o será tan sólo por las acciones intencionales de ciertos sujetos humanos que usen esos sistemas con fines delictivos. Ambos riesgos merecen toda nuestra atención. Sin embargo, aparte de los indudables efectos económicos, laborales y sociales, estos son los riesgos existenciales reales y no otros.

Ciertamente, la IA podría desarrollarse por otros caminos, por ejemplo mediante aprendizajes adaptativos en interacción con el medio natural y con el medio social humano. Pero no es este el camino seguido por los sistemas actuales tipo GPT. Apreciar todo el alcance de esta diferencia es crucial. Y no hay razones de peso para emprender este otro camino, tal vez demasiado largo y complejo.

Pero miremos a otro sitio. Al igual que estamos ya acostumbrados a hablar de una singularidad tecnológica, podemos volver a hablar de la singularidad humana. Y podemos hacerlo en un sentido muy cercano al humanismo. En este trabajo, analizaremos la forma en que deberían caracterizarse ambas singularidades. Y también nos preguntaremos por las condiciones en las que podrían entrar en conflicto. En contra de la opinión que actualmente suele estar presente en los medios de comunicación, argumentaremos que la compatibilidad de ambas singularidades es plausible y que los riesgos existenciales son asumibles. Insistiremos de nuevo en que al argumentar de esta forma, en absoluto queremos minimizar los peligros. Sin duda existen y son preocupantes. La tecnología vinculada a la IA que actualmente existe tendrá efectos profundos en los ámbitos económicos, laborales y sociales. Pero eliminar el aroma de misterio que suele acompañar las discusiones sobre estos temas permite identificar mejor cuáles son los riesgos existenciales reales. Y también nos capacita para saber qué hacer. No hay mayor peligro que no saber exactamente a qué nos enfrentamos.

A continuación, en el primer apartado, seguiremos perfilando las nociones de singularidad tecnológica y de singularidad humana. En el segundo apartado, propondremos una respuesta muy directa a la pregunta sobre las condiciones en las que los desarrollos de la IA que actualmente conocemos podrían llegar a poner en peligro la existencia de la humanidad. Los dos apartados siguientes nos servirán para profundizar un poco más en las peculiaridades de la singularidad humana. Y en el último apartado recogeremos nuestras principales conclusiones.

1. La singularidad tecnológica y la singularidad humana

El concepto de superinteligencia ocupa un lugar central tanto en las discusiones sobre el desarrollo de la IA como en la eventual emergencia de una singularidad tecnológica. Pero existe una gran oscuridad respecto a dos cuestiones:

(C1) ¿Podrá llegar a ser consciente una superinteligencia desarrollada en la línea de nuestra actual IA?

(C2) ¿Hay limites en las predicciones que sería capaz de hacer esa superinteligencia?

Es frecuente hablar de la superinteligencia como si implicara directamente el surgimiento de una plena conciencia cualitativa. Y también hablar de ella como pudiendo poseer una capacidad predictiva ilimitada. Pero debemos tener claro que la respuesta a la primera pregunta seguramente deba ser negativa y que la respuesta a la segunda pregunta podría ser afirmativa.

La conciencia cualitativa es la capacidad de percibir de manera consciente colores, olores y sonidos, la capacidad de sentir dolor y alegría, de tener emociones y sentimientos, etc. Es la capacidad de tener experiencia del mundo y de uno mismo a través de una gran variedad de modalidades sensoriales externas e internas. También podemos llamarla conciencia fenomé-

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

nica, conciencia subjetiva o conciencia personal. Se trata de la capacidad para tener experiencias subjetivas. Pero la llamemos como la llamemos, tal tipo de conciencia no es esencial al desarrollo de la IA. Igual que no fue esencial, sino explícitamente rechazada, en la manera como el propio Turing planteaba la pregunta acerca de si una maquina podría llegar a pensar.[5]

Como dijimos más arriba, aunque respecto a los sistemas de IA más avanzados se pudiera llegar a atribuir una inteligencia de propósito general, y aunque existieran capacidades reflexivas, metacognición, autoprogramación, autonomía agentiva, selección autónoma de objetivos, etc., nada de esto implicaría necesariamente que existe una conciencia cualitativa. Ni siquiera lo hace la capacidad para elaborar una "teoría de la mente" que permita distinguir seres con mente de seres sin mente, y que permita de algún modo identificar el tipo de estados o procesos mentales que pueden tener los seres con mente.

La actual disponibilidad de grandes bases de datos y la gran potencia computacional de nuestras máquinas están permitiendo todos los desarrollos que acabamos de mencionar.[6] De una manera muy parecida a conseguir predecir textos a partir de otros textos, pueden obtenerse imágenes o sonidos a partir de texto, texto a partir de imágenes o sonido, imágenes o sonido a partir de imágenes y sonido, etc. Y se puede ampliar todo esto incluyendo olores, texturas, etc. En el horizonte vemos ya la posibilidad de generar descripciones de estados mentales a partir de imágenes del funcionamiento de nuestros cerebros, muy literalmente un "*mind reading*", así como la posibilidad de crear complejas realidades virtuales multimodales en las que pudiéramos desarrollar vidas paralelas en interacción directa con sistemas de IA productores de contenidos.[7] El *Metaverso* podría ser ya algo más que una sugerente palabra. De acuerdo a las expectativas de mucha gente, para llegar a todo eso sólo estaríamos a la espera de implementaciones adecuadas. Pero al mismo tiempo, también se nos dice que podría estar a la vuelta de la esquina el surgimiento de una singularidad tecnológica consciente de rostro amenazante.

Ante nosotros tenemos sistemas capaces de manipular símbolos, también capaces de obtener y gestionar información sobre la realidad, y en algún sentido capaces de generar y aplicar conocimientos sobre la realidad. A pesar de todo esto, ¿qué plausibilidad hay de que sistemas de este tipo lleguen a tener conciencia cualitativa? Cabe argumentar que realmente no mucha. La conciencia cualitativa no mantiene relaciones directas con la manipulación de símbolos, con el procesamiento de información o con la capacidad de aplicar ese conocimiento a la resolución de problemas. Los humanos tenemos experiencias cualitativas mucho antes de aprender un lenguaje. También ha sido así evolutivamente. Y muchos animales no humanos parecen tener conciencia cualitativa, pueden sentir placer y dolor por ejemplo, a pesar de estar muy lejos de tener nuestras sofisticadas capacidades cognitivas y agentivas.

Así pues, y a pesar de todos los avances de la IA, la respuesta a nuestra primera pregunta, la que llamábamos C1, habría de ser negativa. Y debemos tener esto muy en cuenta al describir lo que hacen nuestros actuales sistemas de IA, o al describir lo que podrían llegar a hacer en la misma línea futuros sistemas de IA con una superinteligencia tendente a la máxima generalidad. No podemos emplear a la ligera palabras como "conocer", "interpretar", "decidir", o incluso una palabra aparentemente tan inocente como "conversar". En este contexto, todas esas palabras solo pueden tener un significado metafórico.

Consideremos ahora nuestra segunda pregunta, C2. ¿Tiene límites la capacidad predictiva de la IA? ¿Los tendría la capacidad predictiva de una superinteligencia desarrollada en la línea de la actual IA? No hay duda de que la capacidad predictiva de la IA puede ser muy grande. Tampoco la hay de que, continuando la dirección emprendida por la actual IA, podríamos llegar a tener una superinteligencia con una enorme capacidad predictiva de alcance muy general. Y se han alzado muchas voces de alarma respecto a los peligros existenciales de la IA, culminando con la emergencia de una singularidad tecnológica cuya potencia predictiva inevitablemente pondría en serio peligro a la humanidad.[8] ¿Qué pensar de todo esto?

Debemos añadir un detalle importante. La IA puede desarrollar una enorme capacidad predictiva sin aplicar necesariamente teorías o leyes. Y sin recurrir tampoco a la construcción de complejos modelos de simulación. Puede integrar esas cosas, pero no las necesita. Lo único que necesita es "aprender" a identificar patrones en ciertas bases de datos. Cuanto más grandes y relevantes sean esas bases de datos (y esto lo proporciona actualmente la tecnología *Big Data*), mejor será el aprendizaje. Y el resultado será que respecto a nuevos conjuntos de fenómenos, la IA podrá predecir con gran eficiencia qué cosas podrán pasar si pasan otras cosas. Eso lo han intentado hacer siempre nuestras teorías y leyes. Y más recientemente, hemos conseguido un poder predictivo

5. Turing (1950) introduce una gran ambigüedad entre la mente entendida como "inteligencia" y la mente entendida como "conciencia". En su trabajo, la pregunta acerca de si puede pensar una máquina se convierte en una pregunta sobre si una máquina podría llegar a manifestar un comportamiento tan inteligente como para que no pudiéramos distinguir si es una persona o una simple máquina, Y la posibilidad de que una máquina sea capaz de llevar a cabo ese comportamiento (ganar al juego de imitación, lo que luego ha pasado a llamarse "el test de Turing") sin tener ningún tipo de conciencia no se consideró una objeción seria. Esta actitud fue heredada por la mayoría de los planteamientos posteriores. Volveremos a tratar este tema más adelante.

6. Adicionalmente, la incorporación de la computación cuántica permitiría dar saltos de gigante en esa potencia computacional.

7. Podemos mirar hacia ese horizonte con la ayuda de Chalmers (2022).

8. Entre esas voces están las de Elon Musk, las de directivos de grandes corporaciones como Google, las de nuestro Parlamento Europeo, las de varias instituciones gubernamentales de Estados Unidos, las de múltiples instituciones y personalidades del mundo de la cultura y la academia, etc.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

aún mucho mayor construyendo modelos de simulación. Pero ahora tenemos algo diferente. La IA nos ofrece sistemas que aprenden a predecir simplemente reconociendo en nuestros historiales de predicciones pasadas patrones acerca de qué datos pueden seguirse con mayor probabilidad de otros datos.

Lo anterior es realmente asombroso desde el punto de vista de la metodología científica.[9] Sin embargo, a pesar de todos estos logros, la capacidad predictiva de la IA puede tener límites. Y podemos señalar tres primeras clases de posibles límites con una larga historia de discusiones:

(L1) Libertad metafisica que tal vez sea propia de ciertas entidades muy especiales.
(L2) Indeterminismo general en la propia realidad.
(L3) Singularidades individuales que impiden la predicción con todo detalle de los fenómenos particulares.

Según L1, ciertas entidades especiales, por ejemplo los seres humanos, podrían ser metafísicamente libres en el sentido de no estar últimamente sometidas a determinaciones externas. Y esto supondría un límite a cualquier pretensión predictiva. Según L2, la propia realidad también podría contener de un modo muy generalizado indeterminaciones irreducibles. El indeterminismo cuántico es el ejemplo más recurrido. L3 impone otro importante límite a la predicción. Con independencia de L1 y L2, podría ser imposible predecir con todo detalle ciertos fenómenos particulares, o incluso cualquier fenómeno particular. Y esto podría ocurrir aún no existiendo ninguna libertad metafísica ni, tampoco, ningún espacio de indeterminación en la propia realidad. Simplemente, el determinismo no implica predecibilidad. La conclusión sería que por muy potentes que fueran los sistemas de IA, podrían existir estos tres límites. Y especialmente el límite a la predicción que implica L3 es muy difícil de evitar.[10]

Más adelante seguiremos analizando los límites a la predicción. Y añadiremos un cuarto límite, distinto de los tres límites que acabamos de mencionar pero de un gran peso respecto a todas las acciones humanas. De momento, baste sugerir que la respuesta a la segunda pregunta, C2, podría ser afirmativa.

Muchas veces se asume que la superinteligencia propia de una singularidad tecnológica implicará el surgimiento de una conciencia cualitativa y de una capacidad predictiva ilimitada, o sin límites relevantes. Ya hemos dicho que defenderemos una posición muy diferente. Además, queremos argumentar que la singularidad tecnológica y la singularidad humana son perfectamente compatibles. Aunque nos situemos en un escenario en el que realmente exista una singularidad tecnológica en los términos en los que habitualmente se presenta, los riesgos existenciales serían de un tipo que ya conocemos. En otras palabras, no hay nada intrínsecamente perverso, o especialmente perverso, en el desarrollo de la IA. Es más, ni siquiera lo habría en el surgimiento de esa superinteligencia propia de la singularidad tecnológica.

Pero dejemos de momento estas cuestiones y cambiemos de tema. El discurso apocalíptico sobre la singularidad tecnológica suele olvidar que también existe una crucial singularidad humana.[11] Podemos precisar su sentido a través de dos rasgos:

(R1) Capacidad de conciencia cualitativa.
(R2) Producción de comportamientos especiales basados en convenciones sociales.

En R1, volvemos a referirnos a la conciencia cualitativa de la que ya hemos hablado a propósito de C1. Realmente no sabemos bien qué es la conciencia. Pero esa conciencia cualitativa, esa experiencia subjetiva, que nos cuesta atribuir a los sistemas de IA que hoy día tenemos, y a los que podemos seguir teniendo en la misma línea de desarrollo, nos la atribuimos espontáneamente a nosotros mismos y la atribuiríamos también de manera muy natural a seres parecidos a nosotros. Asumimos que nosotros la tenemos. También, que la tienen otros sujetos humanos y tal vez algunos animales. Y como hemos dicho, rechazamos que la tengan las máquinas que ejecutan el tipo de programas o algoritmos que conocemos. Esto debería basta para poder afirmar que nada en la línea de una IA basada en LLM llegará a adquirir conciencia.

Las referencias a Alan Turing y John Searle son obligadas. La respuesta de Turing a la pregunta acerca de si una máquina podría llegar a pensar dependía de la capacidad de mantener con fluidez conversaciones en una lengua natural. Los sistemas basados en LLM podrían llegar a tener esta capacidad. Y por lo tanto, tendrían mente según el criterio de Turing. Pero el concepto de mente que tiene Turing no incluye la conciencia cualitativa. El experimento mental de la habitación China de John Searle condensa muy bien nuestras intuiciones respecto a lo dudoso que nos parece que una IA como, por ejemplo, la basada en LLM pueda llegar a tener conciencia cualitativa.[12] Incluso aunque aceptemos la existencia de un

9. Una gran parte de la literatura en la filosofía de la ciencia de las últimas décadas se ha centrado en el contraste entre las leyes y las teorías, por un lado, y los modelos, por otro. No es necesario recordar aquí la larga lista de autores y autoras que han insistido en este punto. Una vez más, la propia realidad ha ido por delante. Pues tal vez ni siquiera los modelos, tal como hasta ahora los estábamos entendiendo, sean necesarios para ampliar, o incluso renovar, el conocimiento científico. A pesar de lo fascinante que resulta este punto, no podemos aquí decir mucho más.

10. El breve cuento de Borges, "Del rigor en la ciencia", publicado inicialmente en 1946 e incluido finalmente en la antología *El hacedor* (1961), pone de manifiesto uno de los problemas principales del intento de tener en cuenta todos los detalles de cualquier singularidad. Un mapa completo de cualquier trozo de la realidad debería duplicar completamente esa misma realidad.

11. En otras épocas, la singularidad humana fue un tema estrella, sobre todo en el Renacimiento y en la Época Moderna, llegando a su momento álgido en la Ilustración. Hoy día, y a pesar de los discursos alarmistas frente a los peligros de la IA, la singularidad humana ha perdido la mayor parte de su antiguo protagonismo. Que nos consideremos simplemente una parte del mundo natural tiene este doble efecto. Y deberíamos aprender a gestionarlo mejor.

12. Dos referencias básicas son Searle (1980 y 1985). Véase también Liz (2009).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

significado dependiente del uso del lenguaje, un significado generado en el uso adecuado del lenguaje en una gran variedad de contextos, la existencia de tal tipo de significado no implica la existencia una conciencia cualitativa.[13]

Pero tampoco podemos engañarnos. Realmente, no sabemos muy bien qué es la conciencia cualitativa. Y tal vez nunca lleguemos a saberlo con la seguridad con la que podemos llegar a saber otras cosas. Simplemente, nosotros la tenemos, la atribuimos a seres cercanos a nosotros y no se la atribuimos a máquinas que tan solo manipulan símbolos. Esta situación ha dado lugar a lo que se conoce como "*El hard problem* de la conciencia"[14]

Dicho problema puede plantearse de forma resumida a través de la siguiente cuestión abierta:

> Dado cualquier conjunto de especificaciones físicas, conductuales o computacionales, que podemos llamar *F*, no parece contradictorio decir que una entidad tiene *F* pero no posee ninguna conciencia cualitativa.

Es más, aunque conozcamos de manera directa nuestra propia conciencia, no hay ninguna contradicción en la idea de que todas las demás entidades que tomamos por sujetos humanos sean simplemente como *zombies* sin ningún atisbo de conciencia cualitativa.[15]

También existiría un "*hard problem*" en la IA. Ahora, los *zombies* son los sistemas de IA. Y la misma cuestión abierta se plantearía respecto a estos sistemas. La diferencia estribaría en que en este caso, asumimos fácilmente que los sistemas de IA son simplemente *zombies*. Aunque su comportamiento lingüístico y no lingüístico fuera como el nuestro, su diferente constitución física, su poca interacción con los entornos reales, su desconexión de nuestras sociedades y culturas, etc., sugiere considerarlos simplemente como *zombies* muy bien construidos. La IA en la línea de los LLM no crea mentes como las nuestras, únicamente crea *zombies* conversacionales.

La situación no cambia si permitimos que nuestros sistemas de IA tengan acceso completo a *Internet*, o si añadimos formas de acceso directo al mundo y a nuestros propios cuerpos, o si aumentamos la potencia de cálculo de los sistemas con computación cuántica, ni siquiera cambia si incluimos capacidades reflexivas, como metacognición, autoprogramación, autonomía agentiva, selección autónoma de objetivos, etc., o la producción de "teorías sobre la mente" y de un *mind reading*. Con todo esto, por supuesto que la superinteligencia de los sistemas de AI crecería enormemente. Y podría hacerlo en una medida inimaginable para nosotros hoy día. Sin embargo, seguiríamos sin tener razones de peso para atribuir conciencia cualitativa. Esa IA tan avanzada tampoco crearía mentes del mismo tipo que nuestras mentes. Tan sólo crearía *zombies* cada vez más sofisticados. Ahora serían algo más que *zombies* conversacionales, pero seguirían siendo *zombies*. Seguirían siendo entidades que únicamente aparentan tener mentes como las nuestras.

Acaso una IA desarrollada de otra forma sea capaz de producir conciencia cualitativa. Y aquí debemos pensar en constituciones materiales más cercanas a nuestra biología, en interacciones con entornos reales y con seres como nosotros, en una integración social y cultural, etc. Se trataría de una IA plenamente generativa, evolutiva e interactiva. Debemos pensar que, en principio, alguna combinación de estas cosas podría dar lugar a mentes como las nuestras. Al fin y al cabo, eso es lo que ha generado nuestras mentes, con toda nuestra experiencia subjetiva. Pero no hay razones para pensar que la IA que hoy día tenemos sea capaz de conseguirlo. Y por tanto, tampoco podemos suponer que una singularidad tecnológica en la línea de la actual IA llegará a producir este tipo de conciencia. Y menos aún, que llegará "inevitablemente" a producirla.

Respecto al rasgo R1, estamos repitiendo muchas de las cosas que ya dijimos a propósito de C1. En ambos casos estamos hablando de la conciencia cualitativa. Y el contraste entre la IA y nosotros es muy claro. Mientras que los humanos tenemos conciencia cualitativa, una experiencia subjetiva ciertamente muy variada, es muy dudoso que una singularidad tecnológica en la línea de la IA actual llegue a tenerla.

Consideremos ahora el segundo de los rasgos con los que hemos caracterizado a la singularidad humana, R2. Se trataba de la producción de comportamientos especiales basados en convenciones sociales. Muchos comportamientos humanos son sumamente especiales en un doble sentido:

1) Son posibles debido a la existencia de ciertas convenciones sociales.
2) Son mantenidos y amplificados gracias a ciertas convenciones sociales, llegando a producir lo que podemos llamar "efectos mariposa"[16]

Los efectos mariposa son variaciones sustantivas en el comportamiento de un sistema como consecuencia de pequeños cambios en las condiciones iniciales. Debido a peculiaridades estructurales, cambios mínimos en las condiciones iniciales pueden acabar produciendo cambios enormes en la dinámica del sistema. El comportamiento llega a ser no lineal incluso en sistemas plenamente deterministas. La dinámica del clima es uno de los ejemplos más frecuentes. Tenemos aquí un sistema plausiblemente determinista pero tremendamente impredecible.

13. Geach (1957) es un clásico trabajo en el que se argumenta la independencia de una cierta tesis del significado como uso respecto a la cuestión de la existencia de actos mentales conscientes. Sin duda no se trata de una obra actual, y ciertamente el autor ya no es muy leído, pero este trabajo contiene ideas valiosas.

14. Por supuesto, véase Chalmers (1997).

15. Sobre la posibilidad metafísica de estos seres, podemos seguir remitiendo a Chalmers (1997).

16. A pesar del abusivo uso que ha llegado a tener esta expresión, permite conceptualizar una clase sumamente importante de fenómenos causales. Una buena introducción a este campo sigue siendo Gleick (1987).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Queremos llamar la atención sobre una clase muy importante de efectos mariposa. Se trata de efectos mariposa vinculados a la existencia de convenciones. Poder articular una gran parte de nuestras acciones mediante convenciones es algo típicamente humano. Y el lenguaje está implicado en la mayor parte de las cosas que hacemos. Pero nuestros lenguajes naturales son sistemas de convenciones capaces de estructurar nuestras acciones de manera que puedan generarse un gran número de efectos mariposa. Este es un aspecto esencial de la singularidad humana. Y lo asombroso es que, en principio, puede ocurrir en cualquiera de nuestras conversaciones cotidianas.

Normalmente no es así. Esto también es cierto. Nuestras vidas humanas suelen ser aburridas y predecibles. Pero sabemos que puede ser así. Y que en algunas circunstancias ese tipo de efectos adquieren relieve. Y cuando esto ocurre, podemos ser tan poco predecibles como el clima.

No se trata solo de que determinados comportamientos se conviertan en acciones con un significado gracias a ciertas convenciones. Estamos resaltando otro fenómeno. Se trata de algo más básico y en gran medida independiente del significado. Es típico del comportamiento humano 1) que se produzcan ciertos efectos en la realidad debido a que existen determinadas convenciones, 2) que generalmente esos efectos no guarden proporción con sus causas y 3) que a veces sean efectos tremendamente impredecibles.

Las convenciones canalizan, mantienen y amplifican los efectos. Y como ya hemos dicho, muchas veces, aunque no siempre ni necesariamente, esos efectos pueden tener la estructura de efectos mariposa. Pequeñas variaciones pueden dar lugar a cambios enormes en lugares y tiempos muy alejados. En el caso de la acción humana, esa posibilidad existe sobre todo porque existen las convenciones pertinentes. Pensemos simplemente en el hecho de que unas pocas palabras emitidas en las circunstancias apropiadas, una cierta orden por ejemplo, pueda causar un bombardeo nuclear masivo en un lugar alejado del planeta. Las convenciones sociales hacen que lo irrelevante pueda ser tremendamente relevante.[17] Y lo crucial aquí no es lo que signifiquen las palabras, sino que se hayan pronunciado de acuerdo a lo establecido por ciertas convenciones.[18]

Muy diferentes sistemas de convenciones podrían conducir a los mismos efectos. Así es como funcionan las convenciones. Hay países donde los coches circulan por la derecha y países donde circulan por la izquierda. En algunos casos, los mismos efectos también podrían conseguirse en ausencia de convenciones. Podrían ser efectos calificables como naturales. Sin embargo, un rasgo característico de la singularidad humana es su capacidad para generar en la realidad una enorme cantidad de cadenas causales estructuradas por convenciones.[19]

El rasgo de la singularidad humana que estamos analizando implica un nuevo límite para la predecibilidad, un límite sumamente radical y muy diferente a los límites L1, L2 y L3 que hemos comentado más arriba. Además, se trataría de un límite que no podría ser sobrepasado por ninguna práctica predictiva llevada a cabo dentro de los contextos sociales en los que vivimos. Podemos describir dicho límite así:

> **(L4)** Si las prácticas predictivas llevadas a cabo en contextos sociales estructurados por nuestras actuales convenciones, o por convenciones en línea con ellas, llegaran a poder predecir "todas nuestras acciones", o "ciertos conjuntos de acciones relevantes", se produciría un colapso total de nuestro mundo social.

Aún estando la realidad completamente determinada y aunque dispusiéramos de la capacidad para predecir todos los acontecimientos que ocurran en ella, no podríamos ejercer esa capacidad, ni siquiera ciertas partes suyas relevantes, sin destruir nuestro mundo social.

La singularidad tecnológica se enfrentaría a este tipo de límite. Predecir todos los comportamientos humanos, o aquellos comportamientos que pueden quedar involucrados en eso que estamos llamando "efectos mariposa", tendría consecuencias devastadoras respecto a la estructura misma de nuestro mundo social.

Insistimos en que este límite sería diferente a los límites L1, L2 y L3 que hemos presentado más arriba. Ahora se trataría de un límite de otro tipo distinto. Es un límite a la predicción impuesto por la manera en que las acciones humanas se estructuran en base a convenciones sociales. ¿De qué estamos hablando? Simplemente, por ejemplo, no toleraríamos cosas como que antes de llevar a cabo unas elecciones ya pudieran conocerse públicamente los resultados. Pueden conocerse las tendencias, pueden estimarse las probabilidades, pueden hacerse pronósticos, etc. Pero debe existir

17. Peirce da mucha importancia a esto en su semiótica. Un aspecto esencial del poder de los signos es justamente su gran capacidad para convertir lo irrelevante en relevante.

18. Todos los autores y autoras relacionados con esa gran familia dentro de la filosofía analítica que ha dado en llamarse "filosofía lingüística" resaltaron la importancia de las convenciones sociales en los fenómenos humanos, desde el uso de lenguajes naturales a la creación de instituciones. Entre los más clásicos, cabe destacar a Anscombe, Austin, Geach, Ryle y Strawson, y perfectamente podríamos incluir también al Wittgenstein de las *Investigaciones Lógicas*. Al contrastar la singularidad tecnológica, incluida la superinteligencia artificial que incorporaría, con lo que estamos llamando singularidad humana, las reflexiones de estos autores y autoras vuelven a ser tremendamente sugerentes. Realmente, deberían estar en primera línea.

19. Sobre la naturaleza de las convenciones, una referencia imprescindible sigue siendo Lewis (1969). La idea nuclear de su análisis es que las convenciones son regularidades en el comportamiento que se mantienen en el tiempo porque proporcionar soluciones a problemas de coordinación. Las convenciones tienen perfecta cabida en el mundo natural, pero no tienen porqué ser fácilmente reducibles a física. Y hay un amplio margen de pluralismo relativo a la manera en que pueden llegar a establecerse muy diferentes sistemas de convenciones.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

siempre un espacio abierto para la sorpresa y el fallo en todas las predicciones. Y no solo por cómo estemos constituidos nosotros, ni tampoco por cómo esté constituida la realidad, ni siquiera por la imposibilidad de conocer todos los detalles de los fenómenos particulares. El límite que ahora estamos identificando lo imponen nuestras convenciones sociales.

Aunque toda la realidad, y también nosotros mismos, tuviéramos una naturaleza completamente determinista, en muchos contextos nos comportamos "como si" fuéramos libres. Y lo que es incluso más importante, queremos seguir haciéndolo, o si se quiere estamos determinados a hacerlo. Aunque nuestros deseos estuvieran completamente determinados, querríamos seguir comportándonos como si no lo estuvieran. Podemos decir que estaríamos determinados a no aceptar que estemos determinados. Así es como la humanidad está hecha. Esa es nuestra "singular" naturaleza.

No suele tenerse en cuenta este límite, pero es parte constitutiva de nuestra identidad. Aunque pudieran superarse los otros tres límites que hemos presentado, crear máquinas capaces de superar este límite crearía profundas distorsiones sociales. Nos llevaría a situaciones muy difíciles de afrontar o incluso de conceptualizar. Podríamos estar, por ejemplo, llevando a cabo unas elecciones democráticas existiendo ya una seguridad plena respecto a quiénes serán elegidos. Una situación así sólo admite dos salidas. Si tal seguridad la tienen unos pocos, se trataría de un fraude sumamente grave. Y si tal seguridad la tienen todos, o acaso simplemente muchos, sería un completo absurdo.

Debemos añadir dos observaciones. La primera es que muchos de esos comportamientos especiales a los que nos estamos refiriendo, y que son mencionados en R2, requieren una conciencia cualitativa muy especial, por ejemplo un cierto sentimiento de estar tomando decisiones o de estar realizando acciones con libertad. No son simplemente comportamientos, sino acciones intencionales que deben estar acompañadas de un tipo muy especial de conciencia agentiva. Sin suficientes garantías para atribuir ese tipo de conciencia cualitativa tan especial, ese tipo de experiencia subjetiva, rechazaríamos que se estuviera llevando a cabo, de manera adecuada, la acción relevante.

Esto también vale en el caso propio. Si tuviéramos razones de peso para sospechar que nuestra conciencia cualitativa de estar tomando decisiones o realizando acciones con libertad es una mera ilusión, rechazaríamos estar realmente tomando esas decisiones o estar realmente llevando a cabo esas acciones. Y la predictibilidad de nuestras decisiones, o de nuestras acciones, ofrecería razones de mucho peso para alimentar esa sospecha. En definitiva, nuestras prácticas predictivas no pueden socavar la confianza que tenemos en estar decidiendo o realizando acciones del modo adecuado. Simplemente, dejaríamos de existir. Y por razones similares tampoco pueden socavarla las predicciones de ninguna máquina cuyos resultados lleguen a ser socialmente accesibles.[20] También en este caso, simplemente dejaríamos de existir.

Presentemos nuestra segunda observación. Muchos de esos comportamientos especiales suelen incorporar otros comportamientos especiales. También esto es típico de ellos. Hay infinidad de comportamientos especiales complejos que incluyen otros comportamientos especiales, o que requieren haber llevado a cabo otros comportamientos especiales. Nuestra vida social está llena de ellos. Cualquier institución o cualquier plan complejo de acción requiere multitud de comportamientos de ese tipo formando complejas estructuras.

Las dos observaciones que acabamos de hacer son muy importantes. Y acrecientan aún más la diferencia entre una singularidad tecnológica y la singularidad humana. En general, todas las ideas que hemos presentado dan un apoyo muy sólido a esa necesidad de "alineamiento" con los objetivos, intereses y valores humanos que suele exigirse al desarrollo de la IA.[21]

La posibilidad de predecir nuestras acciones relevantes en campos como la política, por ejemplo, implica una importante falta de alineamiento. Pero no se trata tanto de que sea una exigencia moral como de que nuestras convenciones sociales no son compatibles con prácticas predictivas de ese tipo. Decir que serían prácticas predictivas socialmente abusivas es decir poco. Tal como es actualmente nuestra vida social, y tal como es previsible que siga siendo, son prácticas predictivas literalmente imposibles. Mejor dicho, sólo podrían existir conceptualizadas como "delitos". Y ciertamente, como delitos sumamente peculiares que consisten en intentar hacer algo que no podemos admitir que pueda hacerse porque, de llegar a hacerse, erosionaría gravemente nuestra sociedad y nuestra naturaleza humana, o las destruirían completamente.

En el apartado siguiente seguiremos hablando de este rasgo R2. Argumentaremos que es justamente este rasgo el que acota tajantemente las condiciones en las que pueden llegar a producirse sistemas de IA que pongan en peligro la singularidad humana. Sólo podría ocurrir eso de manera accidental o por la intencionalidad delictiva de algunos sujetos humanos.

---

20. En otra nota indicábamos lo relevantes que podían ser en estos temas muchas de las discusiones de la llamada "filosofía lingüística", o "filosofía del lenguaje ordinario". Ahora estamos aplicando al caso planteado por la IA algunas de las consideraciones que Strawson (1974) hacía en el problema de la libertad respecto al papel de nuestras "actitudes reactivas" para con los demás.

21. El término inglés que suele emplearse es el de "*alignment*". Véanse, por ejemplo, los Principios Asilomar (https://futureoflife.org/open-letter/ai-principles/) o la declaración de principios de *OpenAI* (https://openai.com/blog/planning-for-agi-and-beyond). La exigencia de que la IA esté "en línea" con los objetivos, intereses y valores de la humanidad estaba ya implícita en las famosas Leyes de la Robótica de Asimov. Y está presente muy explícitamente en todas las iniciativas legales que actualmente se están llevando a cabo.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

2. ¿Choque de singularidades?

Se ha hablado mucho de "choque de civilizaciones".[22] Existe ese peligro cuando grupos culturales muy amplios tienen identidades diferentes fuertemente definidas. Pero esto no es suficiente. Para que haya una confrontación de ese tipo se requiere al menos una de las dos cosas siguientes: o bien 1) que exista competencia por unos recursos limitados, o bien 2) que se tome una decisión intencional de iniciar o mantener tal confrontación. La anterior disyunción no es exclusiva. Pueden darse a la vez 1 y 2. Pero cuando sólo se da 1, la confrontación es simplemente natural en el sentido de no ser intencional.

Queremos introducir la anterior idea para analizar el problema de un choque entre la singularidad tecnológica y la singularidad humana. Supongamos que a partir de la IA que ahora conocemos se llega a producir una singularidad tecnológica. Como hemos dicho, es muy dudoso que la superinteligencia emergente llegue a tener conciencia cualitativa. Pero sí podría tener otras muchas capacidades cognitivas y agentivas relacionadas con la reflexividad, la autonomía, la autoprogramación, la determinación autogestionada de objetivos, etc. Si esto tuviera lugar, ¿sería inevitable una confrontación entre esa singularidad tecnológica y la singularidad humana? ¿En qué condiciones los desarrollos de la IA que actualmente conocemos podrían llegar a poner en peligro la existencia de la humanidad? ¿En qué condiciones la singularidad tecnológica podría llegar a eliminar a la singularidad humana?

Vamos a proponer una respuesta muy directa a estas preguntas. Comencemos distinguiendo dos tipos de confrontación con resultado de eliminación entre diferentes especies vegetales o animales:

> Eliminación no intencional (E1)
>
> Siempre puede existir una eliminación no intencional debida a relaciones causales que no involucren competencia por los recursos. Sin embargo, la eliminación no intencional suele originarse a partir de una competencia por los recursos. Vamos a fijarnos en este segundo caso. La noción de "recurso" puede ampliarse mucho y cubrir multitud de casos. Cabe decir que cuando dos especies utilizan recursos de manera que 1) los recursos que emplea una especie son recursos que no pueden ser empleados por la otra especie, 2) no hay recursos suficientes para mantener las poblaciones de ambas especies, y 3) no hay intencionalidad consciente en los comportamientos involucrados, entonces habrá procesos no intencionales de regulación mutua y, ocasionalmente, al menos una de las especie será eliminada.[23]

> Eliminación intencional (E2)
>
> Tal vez pueda hablarse de intenciones que no son conscientes. Y esto crearía un espacio intermedio entre los tipos de eliminación E1 y E2 que estamos distinguiendo. En cualquier caso, cuando con independencia de los recursos existentes una especie emprende acciones intencionales conscientes en contra de la otra especie, entonces esta segunda especie podrá ser eliminada por la primera.[24]

Si no se dan las condiciones de E1 ni de E2, podrá haber una cohabitación indefinida entre especies. Aunque sea de una forma muy simplificada e idealizada, el marco que acabamos de introducir resulta plausible cuando se trata de especies vegetales o animales, incluidos los humanos. Y también puede ampliarse de forma que resulten incluidos los sistemas de IA capaces de llegar a constituir una singularidad tecnológica.

¿Podrían los sistemas de IA eliminar intencionalmente a la humanidad de manera consciente? ¿Podría existir una eliminación intencional de tipo E2? Como hemos dicho, resulta muy problemática la atribución de conciencia cualitativa, o conciencia fenoménica, o experiencia subjetiva, a estos sistemas. Únicamente podría existir una intencionalidad consciente por parte de aquellos sujetos humanos que usaran esas máquinas como instrumentos para sus fines. Un poco más adelante, volveremos a considerar esta posibilidad. Pero ahora, examinemos con mayor detalle el caso E1 de una eliminación no intencional.

La noción clave para que exista una eliminación no intencional de tipo E1 es la de "competencia por unos recursos". ¿Habría una competencia por los recursos entre los sistemas de IA capaces de llegar a constituir una singularidad tecnológica y la humanidad? Dejemos aparte los recursos materiales y energéticos. Podría llegar a existir competencia en este sentido, pero no sería un problema peculiar de las relaciones entre la IA y los humanos. Los humanos competimos por los recursos materiales y energéticos con cualquier otro ser vivo. En general, cualquier entidad puede llegar a provocar reducciones en los recursos materiales y energéticos que nos sean disponibles. Y viceversa.

---

22. La referencia básica es Huntington (1996).

23. Varios matices. En primer lugar, queremos entender la noción de recurso de una forma tan amplia como para incluir no sólo cosas como alimento o espacio vital, sino todo aquello que permita que los miembros de una especie vivan y se desarrollen. Así, cuando una especie deprede a otra, sin ninguna intencionalidad consciente, también tendremos un caso de posible eliminación no intencional. En segundo lugar, podría ocurrir que acaben desapareciendo las dos especies. Si ocurriera esto, habría una mutua eliminación. Y esta posibilidad significa que considerar que estamos ante un juego "de suma cero" puede ser una simplificación o una idealización. En tercer lugar, aunque los recursos sean limitados, podrían ser suficientes para las dos especies. Por ello, es preferible utilizar la noción de "recursos suficientes" en lugar de la noción de "recursos limitados".

24. En este escenario, es muy claro que ya no se desarrollaría un juego "de suma cero". Podría ocurrir que las dos especies quedaran finalmente eliminadas. Y también puede ocurrir que ninguna de ellas quede nunca eliminada.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

¿Hay otros recursos respecto a los cuales la IA y los seres humanos pudiéramos llegar a competir? Los dos rasgos R1 y R2 que hemos destacado de la singularidad humana sugieren considerar ciertos recursos. Para la humanidad es un recurso importante todo aquello que necesite para poder seguir teniendo estados de conciencia cualitativa, en particular algunos de ellos, y también todo aquello que necesite para poner llevar a cabo esos comportamientos convencionales con posibles efectos especiales que describíamos antes. ¿Hay competencia entre la IA y la humanidad por esos recursos? La respuesta debe ser negativa. Respecto a estas dos clases de recursos, no hay razón por la cual los sistemas de IA tengan que competir con la humanidad en el sentido de obtenerlos a costa de quitárselos a la humanidad. No podría existir una eliminación no intencional en este sentido. Y sólo podría existir una competencia intencional si unos seres humanos utilizaran sistemas de IA para competir por esos recursos con otros seres humanos.

Hay otro sentido en el que la IA puede competir no intencionalmente con la humanidad por esos recursos. Simplemente, el desarrollo de la IA puede tener efectos negativos sobre el desarrollo de nuestra conciencia cualitativa o sobre nuestra capacidad para desarrollar comportamientos convencionales. Tal vez, por ejemplo, la IA produzca atrofias en nuestras capacidades cognitivas o agentivas. Sin embargo, volveríamos a tener aquí un problema que no es peculiar de las relaciones entre la IA y los humanos. Y como en el caso anterior de una competencia por los recursos materiales y energéticos, podemos dejarlo al margen.

Un recurso por el que sí que podría existir una competencia específica entre la IA y los humanos es la información. Pero no porque sea un recurso limitado. Ni tampoco porque sea un recurso que no pueda ser suficiente para ambos tipos de entidades. La razón es otra. Cabe decir que tanto la IA como los humanos pueden llegar a consumir grandes cantidades información. En último término, puede que la información sea un recurso limitado. El universo parece contener siempre una cantidad finita de información. Sin embargo, este detalle puede ser irrelevante. Aún siendo limitada, la información disponible podría ser un recurso suficiente. Es más, la IA y los humanos podrían compartir toda la información que quisieran. La información que consume la IA no se la quita necesariamente a los humanos, ni viceversa. Todo depende de las convenciones que se generen en las interacciones entre la IA y los humanos.

Hay otra posibilidad que debemos tener en cuenta. Los sistemas de IA pueden volverse contra la humanidad de manera no intencional como consecuencia de los planes de acción que esos sistemas podrían llegar a diseñar y ejecutar. En sus planes, por ejemplo, puede ser importante cierta información sobre unas claves bancarias o sobre la autorización necesaria para el despliegue de misiles intercontinentales. En general, puede ser importante cierta información particular que, dadas nuestras convenciones, no queremos compartir. Junto con el caso en el que unos agentes humanos utilicen la IA como instrumento para dañar o eliminar a otros seres humanos, esa posibilidad está muy presente en la literatura sobre el tema así como en el imaginario popular. Y hay algo muy importante que es común a estos dos casos. Algo que siempre tiene que ver con las convenciones sociales

Por un lado, ciertas convenciones sociales tipifican, o deberían tipificar, como delitos esas actuaciones humanas en las que se utilice la IA como instrumento para dañar o eliminar a otros seres humanos. Las clasificaciones que hacemos de nuestras acciones cambian mucho según cambie nuestro conocimiento de la realidad y nuestras posibilidades de acción. La IA está cambiando ambas cosas. Y están surgiendo nuevos "tipos delictivos" al hilo del desarrollo de la IA. Es cierto que la dinámica de nuestras convenciones suele ser muy lenta cuando se trata de cambios legislativos importantes. Pero no es menos cierto que su avance es imparable.

Por otro lado, las convenciones sociales generan muchos contextos en los que se requieren comportamientos que podamos caracterizar como acciones libres, No en un sentido metafísico, sino en el sentido de que se excluirían muchas de las tipificaciones que hacemos de esos comportamientos, y de sus efectos, si existieran sospechas fundadas de manipulación, engaño, distorsión, plagio, violación de los derechos de propiedad intelectual, propaganda interesada, etc. Una obra creativa, por ejemplo, deja de serlo si existen sospechas fundadas de plagio. Y sus efectos como obra creativa se convierten en efectos de otro tipo muy diferente de acción. Estas convenciones sociales son enormemente importantes para nosotros. Y existen en todos los grupos humanos a lo largo de toda la historia de la humanidad. Los procesos democráticos de elección y decisión vuelven a ser un buen ejemplo. Otro buen ejemplo es la propia ciencia. Las convenciones que orientan el conocimiento, y muy en particular el conocimiento científico, se orientan también de ese modo. Incluso el considerar que una sospecha está "fundada" implica que esa sospecha no puede ser resultado de manipulaciones, engaños, distorsiones, mera propaganda, etc.[25]

Suelen enfatizarse los aspectos destructivos de la IA, que culminarían en la aparición de una singularidad tecnológica que casi inevitablemente pondría en peligro la existencia de la humanidad. Sin embargo, lo que estamos diciendo sobre las convenciones activas en nuestras sociedades impide que los aspectos destructivos puedan llegar a manifestarse de una forma generalizada. Sólo podrían generalizarse en forma de accidentes o en forma de delitos.

Está en nuestra mano intentar evitar los accidentes. Cuando ponemos nuestros esfuerzos en esto, las posibilidades de accidentes pueden reducirse mucho. Y cuando la propia ciencia participa en la prevención y control de los accidentes, la reducción suele ser mucho mayor. La propia IA puede tener un importante papel en esa prevención y control de accidentes. Un papel que aún está por diseñar e implementar. Con

25. Vuelve a ser relevante aquí la referencia que hicimos a Strawson (1974).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

todo, siempre podemos equivocarnos. Y siempre pueden ocurrir graves accidentes. Esta posibilidad puede reducirse, pero no puede eliminarse completamente.

También está en nuestra mano evitar los delitos. Y las convenciones más importantes para hacerlo son las leyes que regulan nuestras acciones impidiendo manipulaciones, engaños, distorsiones, plagios, violaciones de derechos de propiedad intelectual, propaganda interesada, etc. Los gobiernos locales, nacionales e internacionales están produciendo ya ese tipo de convenciones. Con ello, la manipulación intencional queda restringida a lo que unos pocas personas puedan hacer cometiendo graves delitos al hacerlo.

En resumen, los mecanismos correctores capaces de bloquear los aspectos destructivos de la IA pueden ser abundantes en aquellas sociedad con cultura científica y con capacidades legislativas ágiles y eficientes. Dejemos al margen los cambios en los ámbitos económicos, laborales y sociales. Muchos de ellos serán inevitables. Pero se parecerán mucho a los efectos que han tenido otras tecnologías, como la máquina de vapor, el uso de la electricidad, el desarrollo de los actuales medios de comunicación, etc. Aparte de los cambios de este tipo, los aspectos destructivos de la IA sólo podrían manifestarse 1) como accidentes locales, o 2) como efectos de las acciones delictivas de algunos sujetos.

Por un lado, una cultura científica puede bloquear, o al menos servir de freno, a los accidentes locales. Conocer, aunque sea a grandes rasgos, cómo funciona la red eléctrica de nuestra casa o el motor de un coche previene accidentes. Y lo mismo puede decirse de conocer cómo funcionan los sistemas de inteligencia artificial, por ejemplo los sistemas GPT. Ampliar y profundizar la cultura científica de nuestras sociedades vuelve a ser aquí una necesidad urgente. Por otro lado, unas capacidades legislativas ágiles y eficientes pueden bloquear o servir de freno a acciones delictivas que se sirvan de la IA. ¿En qué tipo de sociedades puede fomentarse una cultura científica? ¿En qué tipo de sociedades pueden existir esas capacidades legislativas ágiles y eficientes? No vamos a responder aquí estas preguntas. Pero todos tenemos algunas ideas al respecto.

Volvemos a tener que añadir algunos matices. Uno de ellos es positivo, pero el otro no lo es tanto. En primer lugar, como hemos dicho, la propia IA es un potente instrumento para identificar los casos de posibles accidentes. Y lo mismo cabe decir de los malos usos y y de los abusos. La propia IA puede ser un instrumento sumamente útil para identificar los usos delictivos de la IA. Cabe decir que la IA puede ser tan autocorrectora como lo ha sido siempre la propia ciencia. Esta idea es muy importante. El conocimiento científico puede aplicarse reflexivamente para distinguir ciencia de pseudociencia. Y la IA puede aplicarse reflexivamente con gran éxito a la hora de distinguir verdad, objetividad y racionalidad de falsedad, mentira, distorsión y manipulación.

En segundo lugar, y esto no es ya tan positivo, debe tenerse en cuenta que las prohibiciones sobre la IA también pueden estar motivadas por intereses contrarios a la verdad, la objetividad y la racionalidad. Y claramente, una de esas motivaciones puede ser ganar la guerra abierta que actualmente existe, principalmente entre las grandes corporaciones vinculadas a *Internet*, por el control del mercado emergente de la IA.

Dado el tipo de entrenamiento que requieren los sistemas de IA basados en LLM, no es fácil disponer de las bases de datos necesarias para desarrollar sistemas realmente potentes, mucho más potentes que los sistemas actualmente existentes. Lo previsible es que todas las aplicaciones que incorporen este tipo de IA sean dependientes en su entrenamiento previo de un número muy pequeño de grandes nodos de computación. Algo parecido está ocurriendo con la "computación en red", con los buscadores de *Internet*, con los sistemas operativos y con la fabricación de los componentes básicos de nuestros ordenadores. Siendo esto así, fomentar el miedo y las legislaciones restrictivas en determinados países o regiones internacionales se convierte en otra arma más a fin de conseguir el control de este nuevo y sumamente atractivo mercado emergente.

Sin embargo, la IA vuelve a ser también un potente instrumento para detectar y solucionar los problemas. De nuevo, una comparación con la autoaplicación de conocimiento científico puede ser iluminadora. Igual que la ciencia es capaz de proporcionar herramientas muy útiles para identificar motivaciones e intereses poco legítimos dentro de la ciencia, también puede serlo la IA respecto a la identificación de motivaciones e intereses poco legítimos dentro de la propia IA.[26]

Nos hemos preguntado si es inevitable un choque entre la singularidad tecnológica y la singularidad humana ¿Pondrá necesariamente en grave riesgo existencial la singularidad tecnológica a la singularidad humana? Nuestra respuesta ha sido negativa. No se cumple con claridad ninguna de las condiciones que haría inevitable una confrontación así. No hay competencia real por los recursos. No tienen porqué ocurrir graves accidentes. Y nuestras convenciones sociales, particularmente las convenciones legales, se ocupan tanto de alejar la posibilidad de accidentes como de excluir comportamientos humanamente intencionados que conduzcan a tales confrontaciones. Simplemente, estos últimos comportamientos quedarían tipificados como delitos.

Esta es la situación en la que previsiblemente estaríamos si la IA sigue desarrollándose en la línea de la IA que tenemos hoy día y si la humanidad también sigue desarrollándose como lo ha hecho hasta ahora. Por supuesto, puede haber cambios radicales en la humanidad. Puede haber epidemias, desastres naturales, etc. O puede cambiar parte de nuestro genoma, de manera natural o acaso intencionada. Así mismo, también puede desarrollarse una IA diferente, basada en aprendizajes evolutivos y en la interacción real con el entorno y con los seres humanos. Puede desarrollarse una IA que sugiera con

26. Tres autores que se han esforzado siempre en enfatizar este papel autocorrector de la ciencia son Mario Bunge, Karl Popper y Miguel Ángel Quintanilla. Véase Quintanilla (2021).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

gran intensidad que estamos ante nuevas clases de sujetos personales con capacidades conscientes cualitativas, capaces de experiencia subjetiva, y con capacidad de emprender conductas orientadas por convenciones propias. Y pueden ocurrir ambas cosas entremezcladas. Es difícil saber qué podría llegar a ocurrir entonces. Posiblemente, la situación no sería muy distinta de la situación en la que un grupo humano se encuentra con otro grupo humano perteneciente a una cultura muy diferente. Las posibilidades de una confrontación real aumentarían mucho.

Sin embargo, no estamos en una situación de ese tipo. Ni siquiera estamos cerca de ella. La singularidad tecnológica que vislumbramos en el horizonte es de otro tipo muy diferente. Es una singularidad tecnológica con superinteligencias desarrolladas desde la IA que hoy día tenemos. Y todo parece indicar que el agente productor de esas superinteligencias seguirá siendo nuestra propia singularidad humana.

Referencias

Bostrom, N. (2003). "Ethical issues in advanced artificial intelligence". In I. Smit (Ed.), *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2. International Institute of Advanced Studies in Systems Research and Cybernetics, 2003.

Chalmers, D. (1997). *The conscious mind. In Search of a Fundamental Theory*. Oxford: Oxford Univ. Press.

——— (2022). *Reality+. Virtual Worlds and the Problems of Philosophy*. Londres: Penguin Books.

Geach, P. (1957). *Mental Acts*. Londres: Routledge and Kegan Paul.

Gleick, J. (1987). *Chaos. Making a New Science*. New York: Viking.

Harari, Yuval (2014). *Sapiens. A Brief History of Humankind.* London: Harper.

——— (2015). *Homo Deus: A Brief History of Tomorrow.* London: Harvill Secker.

——— (2018). *21 Lessons for 21st Century.* New Yok: Spiegel & Grau.

Hoffman, R., with GPT-4. (2023). *Impromptu. Amplifying Our Humanity Through AI*. Dallepedia.

Huntington, S. (1996). *The Clash of Civilizations and the Remaking of World Order*. New York: Simon & Schuster [*El choque de civilizaciones y la reconfiguración del orden mundial*. Barcelona, Paidós, 2015].

Lewis, D. (1969). *Convention*. Cambridge: Harvard Univ. Press.

Liz, M. (2009). "Simulando a Searle". *Praxis Filosófica*, 28: 117-141.

Kurzweil, R. (2005). *The Singularity is Near. When Humans Transcend Biology*. New York: Viking.

Quintanilla, M. (2021). *A favor de la razón*. Pamplona: Laetoli (Ed. original en Taurus, 1981).

Searle, J. (1980). "Minds, Brains, and Programs", *The Behavioral and Brain Sciences*, 3.

——— (1985). *Minds, Brains and Science. The 1984 Reith Lectures.* Cambridge: Harvard Univ. Press [*Mentes, Cerebros y Ciencia*, Madrid, Cátedra, 1985].

Strawson, P. (1974). *Freedom and Resentment, and other Essays*. Londres: Methuen.

Turing. A. (1950). "Computing Machinery and Intelligence". *Mind*, 236: 433-460.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Sobre humanos y máquinas: ¿qué piensan los públicos?

Elena Denia
Post-Doctoral Fellow at Tufts University, and Civic Science Fellow at the Rita Allen Foundation.
elenadenia@icloud.com

Resumen: Una serie de hitos mediáticos en el desarrollo de aplicaciones basadas en Inteligencia Artificial (IA) ha logrado un impacto sin precedentes en el gran público al ofrecer una extraordinaria accesibilidad. A pesar de las crecientes reflexiones éticas sobre el desarrollo de la IA que atañen a distintos aspectos de su desarrollo, hasta el momento no se ha puesto el foco en el estudio de las concepciones de la IA con el fin de estudiar la posibilidad de incluir la voz social en los debates actuales y buscar la forma de dinamizar la relación ciencia-público. Este trabajo aborda el desafío aportando una primera aproximación a la comprensión de las narrativas sobre la IA, identificadas a partir de tres conversaciones con la ciudadanía. Los resultados revelan que los participantes situaron el foco de responsabilidad respecto a la IA en diferentes actores, pero no suscribieron el relato de una IA que fuera intrínsecamente buena o mala. Sí que hubo diferencias de opinión en cuanto a la concepción de la dualidad humano-máquina, según la cual la IA fue considerada por algunos como una asistente que aumenta las capacidades humanas y nos mejora, y por otros como una intromisión inaceptable en la condición humana.

Palabras clave: Inteligencia artificial; percepción social de la ciencia; singularidad tecnológica; superinteligencia; ética de la inteligencia artificial.

1. Introducción

Paralelamente al desarrollo académico de las ciencias computacionales, el pensamiento creativo de novelistas y cineastas ha dado a luz algunas obras de ciencia ficción que han especulado con la idea de crear «máquinas inteligentes» y que han ido ganando terreno en la cultura popular (Perkowitz, 2007). Algunas de estas creaciones han motivado reflexiones filosóficas, no solo en el ámbito académico sino también entre la audiencia general, acerca de lo que significa ser humano y de lo que nos diferencia de las máquinas, además de flirtear con la idea de la *superinteligencia*, plasmada como una entidad que tiene agencia, a veces corporeizada en forma de robot, otras simplemente en forma de dispositivo electrónico cableado.

En concreto, la concepción de la inteligencia artificial ligada a la idea de la *superinteligencia* es la de una máquina, ya sea física o digital, que asciende al estatus de "ser artificial" al presentarse como una entidad autónoma, con voluntad, sujeta a una evolución rápida e impredecible, con gran capacidad de cómputo y con sus propios objetivos, los cuales podrían hacer peligrar a la humanidad[1]. Éste es un relato de gran calado en la sociedad, en gran parte gracias esas piezas de la industria cultural —literaria y cinematográfica— que han vertebrado el imaginario público a lo largo de las décadas y de forma reiterada. En particular, la idea de que las máquinas se vuelvan contra sus creadores es lo que el escritor de ciencia ficción Isaac Asimov bautizó como «complejo de Frankenstein» (Beauchamp, 1980) haciendo alusión a la criatura de la obra de Mary Shelley de 1818, en este caso una inteligencia artificial biológica que aún hoy en día constituye un poderoso icono cultural (Shelley, 2017: 1818). A partir de esta obra y de otras ficciones se instaura el temor, al menos en las sociedades occidentales, a que las inteligencias artificiales tomen el mando. Como nota de interés, ya en la novela *Frankenstein* la autora enfatiza la importancia de asumir responsabilidades sobre los proyectos científicos (Denia, 2021).

Una reflexión interesante respecto a las narrativas presentes en la sociedad sobre la inteligencia artificial es la que ofrece el filósofo de la tecnología Mark Coeckelbergh en su libro *Ética de la Inteligencia Artificial*, donde señala la influencia de las religiones occidentales a la hora de construir los relatos actuales sobre la IA en los que se evocan las ideas de: (i) trascendencia, ligada en este contexto a la corriente del transhumanismo: sobrepasar la condición humana mediante la IA más allá del cuerpo biológico hacia la inmortalidad —véase, por ejemplo, Diéguez (2017)—, el sueño de perdurar a través de la tecnología y cuyo sumun sería la fusión del ser humano con las máquinas; y (ii) apocalipsis, el relato del fin del mundo a partir de un evento disruptivo y destructor de la humanidad provocado por la IA general: la «singularidad tecnológica». El autor argumenta que en otras sociedades, aquellas influenciadas por religiones de la naturaleza como la japonesa, no se presenta la narrativa competitiva entre humanos y máquinas, sino que estas últimas se plasman como agentes que ayudan —e incluso que despiertan actitudes amistosas—, pero no se asocian al fin del mundo ni al deseo de trascender, aunque también puedan estar dotadas de cierta autonomía y voluntad —para algunos, «alma»—.

En cualquier caso, Coeckelbergh demanda superar las narrativas competitivas y de exageración, por lo que un punto de partida que parece razonable es averiguar si, efectivamente, estos relatos están presentes en el público y si es así en qué grado. Para hacer esa reflexión sobre las concepciones de la IA que sostienen los públicos, a las consideraciones sobre el gran impacto de algunos elementos culturales y de la herencia del pensamiento religioso cabría añadir, hoy más que nunca, el papel esencial que juegan los medios de comunicación en la

1. A grandes rasgos, el investigador Stuart Russell plantea que la IA del futuro podría actuar como el genio de la lámpara: hacer lo que pidamos que haga, pero que el resultado final no sea el que esperábamos. Esta idea se ilustra nítidamente en la película *Juegos de Guerra*, de principios de los años 80, donde un adolescente con habilidades informáticas entabla conversaciones con una máquina que es capaz de aprender de sus errores pero no de distinguir entre la realidad y el juego, llevando a los protagonistas al borde de un desastre nuclear.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

construcción de los relatos sobre la inteligencia artificial y, en particular, el uso masivo de aplicaciones basadas en IA por parte del público general, que ha experimentado un auge sin precedentes con la difusión entre los legos de dispositivos fácilmente accesibles en la red como el modelo lingüístico ChatGPT para conversar, los motores de creación de imágenes artísticas DALL-E, Stable Diffusion y Midjourney o el uso de aplicaciones para generar medios sintéticos como los *deepfakes*, cuyos productos se han hecho virales en varias ocasiones.

Estudiar la percepción social de la inteligencia artificial puede servir para conocer las expectativas y ansiedades de los públicos, fomentar un diálogo con científicos y tecnólogos en el que se incluya la voz social y reflexionar sobre las estrategias de comunicación de la ciencia en los medios —e idealmente aportar recomendaciones—. Una pregunta que emerge es si efectivamente podemos rastrear estas narrativas en las opiniones del público y de qué manera. Ya ha habido algunos sondeos a través de encuestas que han arrojado resultados preliminares interesantes, por ejemplo al revelar el miedo declarado de los encuestados ante los intereses de las personas que financian la IA o que hacen uso ilícito de la misma, no hacia la tecnología en sí (Salas, 2019). Sin embargo, teniendo en cuenta la prominente presencia mediática sobre la potencial pérdida de control de la IA y los relatos acerca de la *superinteligencia* —cuyo máximo exponente probablemente sea la moratoria impulsada por el *Future of Life Institute*, comentada más abajo—, puede ser relevante tratar de averiguar dónde pone el foco el público: en la IA, en sus creadores o en ambos. Y de esta manera facilitar el diálogo propuesto.

En el presente artículo abordo esta tarea. Para ello me valgo de la técnica de investigación de los grupos focales, utilizada comúnmente para la exploración de temas emergentes y que puede constituir una buena primera aproximación a los sondeos de opinión (Jensen & Laurie, 2016). Como preámbulo, puede resultar esclarecedor reparar en ciertas consideraciones sobre la idea de la *superinteligencia*, que coge especial fuerza entre los públicos con el éxito del paradigma de investigación del *conexionismo* en las ciencias de la computación, y que está más presente que nunca en los medios de comunicación, y por ende en los foros que ofrecen las redes sociales.

2. El conexionismo: un impulso para la idea de la *superinteligencia*

A mediados del siglo XX Alan Turing se preguntó si las máquinas podrían «pensar», en el sentido de realizar abstracciones y aprender, en un artículo académico que con el tiempo ha acumulado un buen número de citas (Turing, 1950). Pocos años después, en el verano de 1956, tuvo lugar durante varias semanas el seminario Dartmouth en New Hampshire (Estados Unidos), centrado en las «máquinas pensantes» (McCorduck & Cfe, 2004), donde se planteó la idea de que sería posible simular, que no recrear, la inteligencia humana si sus características lograban descubrirse con la suficiente precisión, e incluyó ideas de auto-mejora y de hacer abstracciones a partir de datos sensoriales y de otros tipos (McCarthy et al., 2006: 1955). Nótese que el término «inteligencia artificial» se atribuye a la propuesta formal para la celebración del seminario. Resulta interesante que algunos de los asistentes pensaban ya que una máquina *súperinteligente* estaba a la vuelta de la esquina, en apenas una generación (Coeckelbergh, 2018). Una *superinteligencia* superaría al ser humano en sus funciones cognitivas y podría desembocar en un escenario que escaparía del entendimiento humano, lo que Ray Kurzweil denominó «la singularidad» (Kurzweil, 2005). La idea de que esto pudiera derivar en la destrucción de la humanidad es, sin embargo, una cuestión aparte.

Tal vez el rótulo «Inteligencia Artificial» ha constituido un acierto promocional, similar a la invención del término «Big Bang» acuñado en un programa de radio de la BBC dirigido al público general por Fred Hoyle, físico y escritor de ciencia ficción y detractor de este modelo. Sin embargo, esa «gran explosión» no describe la realidad física (Lineweaver & Davis, 2005), sino que exhibe un sentido metafórico —sobre metáforas en la ciencia, véase Montuschi (2017)—, y si bien sabemos lo que es una "gran explosión", en el caso de la IA, en cambio, la cosa se complica porque no contamos con una definición satisfactoria de lo que es la «inteligencia», un término problemático que admite distintas acepciones. Esto no quiere decir que haya que abandonar el intento de una definición, pero sí tener presente que, quizá de forma análoga a lo que sucede con el problema de la demarcación entre ciencia y aquello que no lo es —por ejemplo, pseudociencia (Fasce, 2018) o metafísica (Carnap, Hahn & Neurath, 1929)—, al menos a día de hoy resulta difícil encontrar un criterio completo para determinar qué es la inteligencia.

Desde el artículo de Turing el paradigma de investigación dominante fue, durante casi tres décadas, la inteligencia artificial *simbólica* —como los árboles de decisiones—, hasta que, a partir de los años 80, el *conexionismo* fue abriéndose camino con las llamadas «redes neuronales artificiales» —otro triunfo promocional, aunque en este caso sí tenemos definición para el término biológico neurona y se puede aducir un sentido metafórico al evocar un producto matemático—. Curiosamente, el máximo detonante en el desarrollo de estas redes fue la industria del videojuego, que empleó esta tecnología para generar gráficos precisos a partir del cálculo de las dinámicas del juego a tiempo real, con el objetivo de ofrecer una variación fluida de imágenes para simular las físicas en pantalla; por ejemplo, el movimiento de los personajes, el avance en los escenarios, o las mecánicas de objetos que caen o estallan, incluyendo detalles muy precisos como los cambios de luz. Este impulso tuvo su análogo en el mundo de la investigación aplicada, que a su vez, al verse superada por la acumulación masiva de datos —el llamado *big data*—, aprovechó el potencial de la tecnología y empezó a utilizarla de forma creciente en proyectos científicos que de otro modo resultarían difícilmente abordables; por ejemplo, para el procesado de imágenes en los análisis de grandes catálogos de galaxias.

Si tomamos el ejemplo del estudio del cielo nocturno, resulta ilustrativo el llamamiento a la participación del público en el proyecto de ciencia ciudadana *Galaxy Zoo*, que en 2007 invitó

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

a las audiencias a clasificar galaxias en función de su forma —espirales, elípticas e irregulares— a partir de imágenes tomadas por los telescopios. La idea científica subyacente a este proyecto era que la cognición humana permite el reconocimiento de patrones en las imágenes, un ejercicio que no arrojaba buenos resultados con procedimientos informáticos (Lintott, et al., 2008). Sin embargo, con la implementación gradual de las redes neuronales artificiales en proyectos científicos, esta asunción ha cambiado, e incluso algunos trabajos de los últimos años han sugerido que éstas pueden dar lugar al descubrimiento de conceptos físicos (Iten et al. 2020; Denia, 2020).

Nótese además que estas reflexiones no solo se circunscriben al ámbito de la ciencia. En particular, los desarrollos con inteligencia artificial de empresas tecnológicas persiguen objetivos prácticos, se ven influenciados por el contexto social que los auspicia al tiempo que tienen influencia sobre mismo. Respecto a lo último, debe señalarse que la ejecución de los algoritmos puede tener consecuencias éticas. En esta línea, los efectos de la IA sobre el mundo están ligados a la idea de responsabilidad, que a su vez presenta un problema de atribución que puede atañer a los responsables del proyecto, programadores —los originales y los sucesores si los hubiera—, usuarios finales, la selección de los datos de entrada o el algoritmo en sí. El último caso es el más delicado, ya que el mecanismo informático no es un agente moral que exhibe intencionalidad, al menos mientras no se alcance la mencionada *superinteligencia* en caso de que fuera posible, pero sí que se admite que puede llevar a la pérdida de control sobre los resultados y a efectos indeseables.

Más que nunca, la cuestión sobre hasta dónde nos puede conducir esta tecnología está en tela de juicio en la actualidad, habiendo traspasado las especulaciones propias de la ciencia ficción hasta el punto de abordarse con seriedad en los foros académicos y tecnológicos, en ocasiones con gran preocupación. Para algunos investigadores tiene sentido explorar la idea de que sea posible crear una *superinteligencia* dado que los riesgos pueden ser muy elevados, por ello instan a no descartarla y a evaluar la posibilidad dilucidando sobre: (i) las formas en que podría darse; (ii) las formas en que podría mitigarse. A su vez, desde la filosofía se ha planteado el dilema de si sería ético apagar una máquina con agencia (Coeckelbergh, 2018), si es que hay forma de averiguar que no se trate de un juego de imitación muy sofisticado, como el experimento mental de la "habitación china" propuesto por John Searle en el que la máquina habla chino pero no es consciente de los significados (Searle, 1980). En cualquier caso, tanto partidarios de la *superinteligencia* como detractores que adoptan posturas más modestas pero que aún así se muestran preocupados por la pérdida de control de algunas tecnologías basadas en IA, generalmente coinciden en que la clave reside en la regulación sobre el diseño de los proyectos, un planteamiento que, como se explicita más abajo, también está muy presente entre las preocupaciones de los públicos.

Ese problema del control, de proporcionar guía y pautas a la IA, está sobre la mesa desde hace tiempo y exige enormes esfuerzos tanto a nivel filosófico como en materia de regulación. Por ejemplo, ante los dilemas planteados por la toma de decisiones en los algoritmos de los coches autónomos, hay una incógnita de la que a menudo se han hecho eco los medios de comunicación y han invitado al público a la reflexión: ¿qué debe priorizarse: el atropello de un viandante o poner a salvo a los usuarios del vehículo ante una probabilidad elevada de siniestro total? No solo no es tarea fácil responder, sino que esta pregunta debería subsumirse en un marco más general: ¿hay consenso sobre cómo deben resolverse estas situaciones y, por tanto, en la forma en que debe ejercerse ese control o guía? Por su parte, las aplicaciones militares son un terreno aún más pantanoso, y más teniendo en cuenta su proyección transnacional y la elaboración de protocolos que pueden diferir de los valores de otras culturas que se vean afectadas. En este sentido, cabe reiterar que los públicos de diferentes sociedades pueden albergar opiniones muy distintas, ligadas a creencias, valores y experiencias, por lo que ese diálogo deseable antes mencionado entre el público y otros actores implicados debería poder ampliarse en la medida de lo posible en vistas a la colaboración entre países y a establecer pactos que tengan en cuenta la opinión de públicos diversos. El presente texto, sin embargo, se centra en las concepciones de la inteligencia artificial en la sociedad occidental, tal y como ya se ha apuntado.

3. El auge de la inteligencia artificial entre el público: hitos mediáticos

Desde finales de 2022 los debates sobre el avance de la Inteligencia Artificial han adquirido un protagonismo sin precedentes en la esfera pública, sacudiendo los medios de comunicación, las empresas tecnológicas y el mundo académico. Si bien la IA está ampliamente integrada de manera invisible en la vida cotidiana actual —desde aviones hasta lavadoras—, recientemente una serie de desarrollos, basados en inteligencia artificial generativa, como los modelos de lenguaje, el *deepfake* y las creaciones artísticas, han llamado la atención del público general. A continuación se examinan los principales hitos mediáticos de la IA para comprender mejor el impacto sobre las audiencias.

Por un lado, se han popularizado enormemente los motores basados en IA para generar arte pictórico. Puede citarse aquí la polémica desatada en septiembre de 2022 que dio la vuelta al mundo a través de titulares de prensa porque la obra de arte digital *Théâtre D'opéra Spatial*, creada con IA, ganó un concurso de arte digital. Cabe señalar que el tribunal desconocía el uso de la inteligencia artificial por parte del artista, quien no obstante alegó posteriormente haberla concebido tras una infinidad de iteraciones cambiando las instrucciones de entrada proporcionadas al algoritmo y aportando los retoques apropiados al diseño gráfico resultante. Este hecho desembocó en un debate exaltado sobre la autoría y los derechos legales, ya que la red neuronal se alimenta con obras de otros autores. En la misma dirección, la discusión se amplía a la generación de piezas musicales sintéticas.

Sumado a ello, otro hito mediático de la IA son los *deepfakes*: la generación de vídeos manipulados digitalmente que muestran

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

a personas diciendo o haciendo cosas que nunca sucedieron. Algunos ejemplos de ello son los *deepfakes* virales de políticos pronunciando discursos con opiniones incendiarias —y que para algunos suponen la amenaza de que podrían motivar el inicio de una guerra sobre premisas falsas—; o también aquellos que simulan escenas pornográficas sin consentimiento. En particular, se trata de una tecnología que ha suscitado una serie de controversias públicas al levantar preocupaciones ante la facilidad para la difusión de información falsa en medios digitales, la posibilidad de falsificar pruebas testimoniales, suplantar identidades y vulnerar el derecho de una persona a su imagen e identidad. Son varios los autores que han abordado esta cuestión desde el punto de vista filosófico, preocupados por el futuro de la verdad en los entornos digitales (Citron & Chesney, 2019; Fallis, 2021; Floridi, 2021) y el valor testimonial de grabaciones —audios y vídeos— en entornos jurídicos (Rini, 2020). Ante este panorama, se ha acuñado el término «infocalipsis» (Schick, 2020).

Paralelamente a estas inquietudes, debe admitirse que también se ha incrementado el uso del *deepfake* con fines positivos (Jaiman, 2020), resaltando su potencial para la comedia y en especial para el arte. Así, es posible sacarse un *selfie* en Florida junto al *deepfake* del pintor Salvador Dalí en la puerta del museo que lleva su nombre. Aunque quizá su uso más extendido resida en el ámbito cinematográfico, por ejemplo para rejuvenecer a personajes como Indiana Jones o a la replicante de la película Blade Runner. En cualquier caso, estas aplicaciones en principio inofensivas también pueden levantar la voz de alarma; por ejemplo con la resurrección de la princesa Leia en la saga de Star Wars una vez fallecida la actriz que la encarnaba —es decir, sin contar con su permiso—. En esta misma línea, el *deepfake* también se ha aplicado con fines publicitarios, por ejemplo mostrando a la actriz Audrey Hepburn, fallecida en 1993, en un anuncio de bombones —¿hubiera aceptado promocionar esos chocolates?—; o bien a la leyenda de las artes marciales Bruce Lee, fallecido en 1973, recomendando una marca de *whiskey*, cuya filosofía vital no incluía el consumo de alcohol[2].

Por último, queda señalar el uso de modelos de lenguaje muy sofisticados como ChatGPT, con los que se puede mantener una conversación en lenguaje natural y que devuelven información elaborada a partir de distintas fuentes; en numerosos casos tan convincente que para algunos el test de Turing ha quedado obsoleto como prueba de inteligencia (Brookes, 2023). Alineado con ello, el uso masivo de estos modelos para la producción de textos e ideas ha irrumpido en las aulas, suscitando reflexiones sobre su impacto en la enseñanza y en el aprendizaje. Se ha abierto así un debate urgente para discutir los términos de uso por parte del alumnado y del profesorado así como las normativas que deberían aplicar al respecto las universidades y los centros educativos. Sin duda, la búsqueda de protocolos urgentes denota ese uso extendido que penetra profundamente en los públicos, incluyendo a los más jóvenes, y que por ende repercute en la percepción pública de la IA.

Con todo, la tecnología se ha manifestado explícitamente en los ordenadores personales y lo ha hecho invitando a la participación de los legos. Existen aquí dos diferencias fundamentales que caracterizan el *boom* mediático actual de la IA: (1) la accesibilidad, ya que el público puede interactuar con ella de forma autónoma, con economía de medios y en un breve espacio de tiempo; y (2) la asistencia personal, en concreto para la consecución de tareas creativas a la velocidad del *click* que antes estaban reservadas a la condición humana, como la generación de imágenes y textos —desde la redacción de un simple correo hasta una disertación con argumentos y contraargumentos—. Nótese que hasta el momento la Inteligencia Artificial no había interferido de forma tan rotunda en los dominios de la creatividad, menos aún con tal agilidad y al alcance del usuario medio. Este hecho tiene profundas implicaciones que apoyan la concepción utilitaria de la IA como asistente ya que, como se señalará en el siguiente apartado, esto es percibido por muchos como un aumento de las capacidades humanas que permite ahorrar tiempo y esfuerzo. Al mismo tiempo, la otra cara de la moneda revela en algunos un vértigo acuciante ante otra cuestión: la sustitución humana —trabajo, arte, otros—. Esta idea se examina ampliamente en el siguiente apartado.

Naturalmente, ante este panorama tanto tecnólogos como intelectuales académicos de diferentes ámbitos han ofrecido reflexiones al respecto, abarcando una gran variedad de temas: el impacto en la educación, el futuro del trabajo, la autoría sobre producciones creativas, la manipulación de la información digital, e incluso la posibilidad de que estos rápidos avances desemboquen en una IA autoconsciente, es decir, la mencionada *superinteligencia*. Resulta crucial reparar aquí en que muchas de estas reflexiones también han gozado de gran influencia mediática, particularmente las que han adquirido un tono de advertencia y que por tanto han favorecido un clima de alerta entre el público. Estas expresiones públicas presentan una imagen de la IA como un cuerpo único de conocimientos, que o bien puede ser bueno o bien desastroso, proporcionando así una lectura simplista del fenómeno que su vez favorece que las opiniones se polaricen, en vez de propiciar posturas más matizadas.

Hay varios eventos protagonizados por *influencers* del gremio. Uno ilustrativo es el de la renuncia del ingeniero informático Geoffrey Hinton, ganador del *Premio Turing* por su trabajo en redes neuronales y bautizado en los medios como 'el padrino de la IA', quien abandonó su puesto de trabajo en *Google* bajo el pretexto del supuesto peligro que se avecina con el desarrollo de la IA. Otro ejemplo de mayor alcance es la carta abierta del 28 de marzo de 2023 emitida desde el *Future of Life Institute* —presidido por Max Tegmark, investigador del MIT y autor de obras de divulgación populares como *Vida 3.0* (Tegmark, 2018)— en la que se instó a pausar el desarrollo de

2. Esta práctica ha resultado en un neologismo: "nigromancia digital", y al respecto puede añadirse que existen aplicaciones y sitios web que ofrecen este servicio para recrear virtualmente a familiares fallecidos con el pretexto de que puede facilitar el proceso psicológico de recuperación durante el duelo por la pérdida.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

la Inteligencia Artificial durante seis meses, y que fue firmada por magnates, pensadores influyentes de nuestros tiempos y miles de investigadores; entre ellos Elon Musk, Steve Wozniak, Yuval Noah Harari, Stuart Russell, por citar algunos (Future of Life, 2023). La carta, concebida bajo el eslogan «¿Debemos arriesgarnos a perder el control de nuestra civilización?», irrumpió en los medios de comunicación generando gran revuelo al estar firmada por estos personajes públicos. Si bien para los autores el escenario desastroso de una IA de propósito general que pueda escapar del control humano y que nos destruya debe ser motivo de alarma, resulta dudoso que tanto ellos como el resto de los firmantes de la carta crean que el desarrollo de la IA se va a paralizar, teniendo en cuenta que se interpone entre los intereses empresariales de líderes de la industria tecnológica de los que participan una gran variedad de países, incluyendo grandes potencias como China o Rusia, además de que frustrarían objetivos científicos como —por recuperar el ejemplo del apartado anterior— el de mejorar la eficiencia en la clasificación de galaxias y la producción de modelos de universo, o algunos más urgentes como la detección de tumores cancerosos en el ámbito médico. Emergen aquí varias preguntas: ¿Compran los públicos el discurso de una IA asesina? ¿Tienen algo que decir sobre los actores implicados en el desarrollo de la IA? ¿Hay otras perspectivas que se nos escapan?

Queda señalar que, en consonancia con los medios de comunicación, buena parte de la ciudadanía se ha pronunciado sobre la inteligencia artificial; siendo un reflejo de ello los innumerables *posts* de los usuarios de las redes sociales que han contribuido a la difusión de noticias y/o han publicado opiniones al respecto. Sin embargo, cabe preguntarse si la sociedad comparte las ansiedades expresadas en la carta o por parte de personajes influyentes como Geoffrey Hinton. ¿Qué piensa el público general? A continuación se explora esta cuestión.

4. Sobre humanos y máquinas: ¿qué piensan los públicos?

A lo largo de 2023 llevé a cabo una investigación en el Massachusetts Institute of Technology (MIT) junto al sociólogo John Durant, experto en el área de la Comprensión Pública de la Ciencia y responsable del diseño de encuestas de percepción social de la ciencia cuyas preguntas se siguen conservando en la actualidad en diversos indicadores del mundo[3] (Durant & Evans, 1995). Con el fin de comprender las concepciones del público sobre la inteligencia artificial y, en particular, analizar sus opiniones sobre el futuro de la verdad digital ante la emergencia de la tecnología *deepfake*. Desde el Museo del MIT organizamos tres grupos focales para conversar con diferentes tipos de público, clasificados en función del grado de familiaridad con la ciencia y la tecnología. Escogimos esta aproximación metodológica al tratarse de una técnica que aporta textura a los estudios exploratorios sobre temas de investigación emergente (Jensen & Laurie, 2016). Este ejercicio coincidió con el auge de la IA generativa en los medios de comunicación de masas, por lo que en aquel momento la IA estaba muy presente en la mente de los públicos y las conversaciones sobre la misma se desarrollaron con fluidez.

A la motivación del estudio también contribuyó una exposición dedicada a la Inteligencia Artificial bajo el rótulo *Mind the gap* que el propio museo exhibía en ese momento, siendo uno de los objetivos declarados por su comisaria, Lindsay Bartholomew, mostrar ciertas diferencias en cuanto a qué puede hacer una máquina dotada de inteligencia artificial frente a qué puede hacer un ser humano. Por ejemplo, un robot puede programarse para que aprenda a tocar una pieza musical con relativa facilidad, mientras que ponerse un jersey puede parecer un ejercicio infernal; por su parte, en el caso de un ser humano, normalmente sucede al contrario. También había un espacio dedicado al *deepfake*, una instalación doble que contaba, por un lado, con una actividad interactiva que retaba a los visitantes a identificar si un vídeo era real o manipulado —y cuyos resultados de acierto rondaban el 50%—, y por el otro, con una recreación de un salón de casa decorado estilo años 60 y una televisión antigua que proyectaba la obra de arte *In Event of Moon Disaster* —galardonada con un premio *Emmy* en 2021—, un *deepfake* del ex presidente estadounidense Richard Nixon pronunciando el discurso de contingencia preparado para un escenario en el que la tripulación del Apolo 11 no hubiera logrado regresar de la misión a la Luna en 1969 (Panetta & Burgund, 2019). Esta instalación generaba cierta inquietud entre algunos curadores del museo, al pensar en hipotéticos visitantes despistados que malinterpretaran la obra y emplearan las redes sociales para acusar al MIT de respaldar teorías conspiranoicas.

Para nuestro estudio organizamos tres grupos de discusión de una hora de duración, cada uno de los cuales representó sucintamente un subconjunto diferente de la sociedad, con la intención de ilustrar un grado distinto de cercanía a la tecnología, aunque sin perder de vista las limitaciones en cuanto a la muestra: un número reducido de adultos —entre 12 y 14 en cada grupo— que fueron reclutados en el área metropolitana de Boston (Estados Unidos) pero que a nuestro entender podía arrojar pistas sobre la percepción social de la IA y servir de punto de partida para realizar estudios posteriores de mayor profundidad teórica y alcance empírico.

Para construir la tipología de los públicos, determinamos los dos primeros grupos de acuerdo con los estudios clásicos de la percepción social de la ciencia: (i) *público general* —reclutado en las calles de la ciudad— y (ii) *público atento a la ciencia y la tecnología*, es decir, informado sobre la actualidad científica y con posibles conocimientos técnicos (Miller, 1983) —reclutado entre los visitantes del museo—; mientras que el diseño del tercer grupo, el (iii) *público involucrado*, se basó en la idea del *engagement* —véanse, por ejemplo, Bauer, Allum & Miller (2007); European Commission (2008)—, en este caso un grupo de individuos que por iniciativa propia había completado el curso de tres días *Make a Fake*, ofrecido públicamente

3. En gran medida en los sondeos de opinión sobre ciencia y tecnología de los *Special Eurobarometer*, así como en sus análogos estadounidenses, los *Science & Engineering Indicators*, y españoles, la *Encuesta de Percepción Social de la Ciencia* de la Fundación Española para la Ciencia y la Tecnología (FECYT).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

por el MIT, en el que se enseñó a utilizar aplicaciones sencillas basadas en IA generativa para crear textos e imágenes. Conviene aclarar que desde hace tiempo se ha teorizado ampliamente sobre el "giro participativo" en la ciencia (Irwin, 2014; Jasanoff, 2014; Wynne, 2014), para el cual se considera deseable que la ciudadanía se implique en actividades de investigación —por ejemplo a través de proyectos de ciencia ciudadana (Haklay, 2018)— y/o participe en debates sobre ciencia y tecnología con cierta repercusión en la toma de decisiones —una visión que parece alineada con algunos aspectos del ideal la ciencia bien ordenada de Kitcher (2001; 2011)—. Hasta ahora este último subconjunto del público no se había contemplado en los estudios comparativos de percepción de la ciencia.

Respecto a los resultados, tras el análisis de las conversaciones quizá la característica más llamativa fue que ningún grupo se centró en la idea de que la IA fuera intrínsecamente buena o mala, sino que situaron el foco de responsabilidad en diferentes actores. Sí que emergieron multitud de manifestaciones espontáneas sobre la comparación entre humanos y máquinas, siendo uno de los aspectos más notables de la conversación, pero cuya visión difirió entre los grupos, ya que para algunos se reducía a una cuestión de ampliación de capacidades y para otros constituía una intromisión ilícita en la condición humana. Veamos de cerca las diferencias.

Público general

El grupo del público general situó el foco de responsabilidad respecto a la IA sobre una entidad abstracta que adopta diferentes formas —siempre refiriéndose a "ellos"—: estafadores, empresas, gobiernos y científicos; y que genera gran desconfianza. Se habló reiteradamente de amenaza, violación de la privacidad y, efectivamente, de pérdida de control, al haber incertidumbre e inquietud sobre su evolución. Los argumentos esgrimidos se basaron en experiencias personales o anecdóticas y se emplearon ejemplos de la industria cinematográfica como las obras *Yo, Robot* y *Parque Jurásico* para expresar preocupación sobre la posibilidad de la IA como un desarrollo tecnocientífico que puede acabar en desastre social. En este sentido, puede decirse que el grupo hizo algunas alusiones a la potencial pérdida de control sobre la tecnología. Además, el grupo exhibió una percepción generalizada de que la tecnología avanza a un ritmo de desarrollo para muchos desbordante. En ese contexto se preguntaron acerca de quién debe controlar y limitar la tecnología, y la desconfianza se situó reiteradamente en quien maneja los hilos y, en mucha menor medida, en la propia IA.

Debe añadirse que, aunque las aunque las expresiones de opinión fueron predominantemente negativas, una minoría manifestó un sentido de la maravilla hacia las ventajas de la IA, tanto presentes —para el entretenimiento, simplificar tareas y ahorrar tiempo— como futuras —supuesto potencial para conducirnos a un futuro soñado de alta tecnología—.

Respecto a la dualidad humano-máquina, sí se desarrollaron opiniones acerca del reemplazo del ser humano, por ejemplo ante un futuro inmediato y aterrador en el que los doctores son sustituidos por IAs que tienen nombre y te prescriben medicamentos. Generalmente, la idea de que una IA adopte aspectos de la apariencia humana y tome decisiones con un efecto directo sobre los individuos se percibió como una intrusión a la condición humana. De hecho, el grupo mostró desaprobación ante la idea de que la IA parezca humana —sin entrar a discutir sobre si realmente tiene agencia o no—, y que se entrometa en juicios morales, por ejemplo ante el panorama de que ChatGPT te indique qué es moralmente aceptable y qué no lo es —"me dice cómo ser mejor humana"—, o que se haga pasar por humana y engañe a otros para revelar sus secretos o influir en sus vidas sin vuelta atrás.

Queda señalar, que no parece que el público general esté dispuesto a aceptar robots humanoides, una idea que ya plasmó Isaac Asimov en algunas de sus historias en las que los humanos relegaban las máquinas al espacio. En este caso, el factor humanoide no solo se circunscribe a la corporeización robótica de la IA, sino a tareas intelectuales como conversaciones en lenguaje natural que se materializan digitalmente en una ficción donde se atribuye agencia a las máquinas y que hacen cuestionar el problema de la identidad humana. En este sentido, el concepto del Valle Inquietante se extendería más allá del campo de la robótica y no se reduciría a la apariencia antropomórfica —véase, por ejemplo, Mori, MacDorman & Kageki (2012)—.

Público atento a la ciencia y la tecnología

Curiosamente, el grupo conformado por el "público atento" a la ciencia y la tecnología situó el foco de la conversación sobre una entidad bien definida: el público general, que además se abordó desde un punto de vista ajeno. El público se percibió como ingenuo, a veces incluso perezoso o poco inteligente, y se habló de forma reiterada sobre la importancia de educar a la sociedad[4]. En esta línea, se manifestó la urgencia de que los ciudadanos deben aprender a lidiar con la IA, concebida como un fenómeno inevitable que forma parte del curso natural del desarrollo tecnológico. Si bien unos pocos admitieron ciertas capacidades del individuo medio, se supuso en todo momento que para que la sociedad piense correctamente debe adoptar el "marco apropiado" —al parecer común y general— que le permita entender la situación tecnológica para hacer un uso responsable de las aplicaciones y afrontar los posibles problemas que puedan surgir, como estafas digitales o desarrollos tecnológicos tendenciosos, peligrosos y/o perjudiciales.

4. Aquí puede ser revelador prestar atención a algunas frases sobre la concepción de los públicos: Enseñémosles, "la sociedad no está preparada", "es perezosa", "no es lo suficientemente inteligente", no es lo suficientemente escéptica, "no comprueba las fuentes", debería aprender en quién confiar; "la gente no está dispuesta a tomarse el tiempo para aprender", y la gente debería tener el marco adecuado: "¿cómo nos aseguramos de que todo el mundo tiene el mismo marco para entender y hablar de estas cosas?".

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Uno de los aspectos más destacables de este grupo es que tendió a proponer soluciones de forma espontánea a lo largo de toda la conversación que, aunque casi siempre se centraron en las carencias de los públicos, situaron la responsabilidad respecto a la IA en una variedad de actores: (i) los científicos: quienes deben enseñar a la gente a utilizar la tecnología con prudencia; (ii) los educadores/educación: quienes deben potenciar el pensamiento crítico, en especial entre los jóvenes; (iii) la gente: que debe aprender y emplear el boca a boca para educar a sus iguales; (iv) la evolución propia de la sociedad: que integra cambios de forma natural en la forma de enfocar los problemas; pero también (v) el Estado: que debe legislar y definir normas adecuadamente. En general, los participantes aportaron soluciones que requieren tiempo y esfuerzos para el ajuste tecnología-sociedad. En particular, esta fuerte tendencia a buscar la manera de suplir las presuntas carencias —de conocimientos y actitudes— es un planteamiento que se ajusta al conocido "modelo del déficit" formulado en las áreas de percepción social de la ciencia y de comunicación de la ciencia, según el cual el público debe adquirir conocimientos hasta adoptar la actitud correcta —véase por ejemplo, Sturgis & Allum (2004); Brossard & Lewenstein (2010)—. Este modelo, que hoy se considera por muchos superado, no se había detectado antes en el estudio de los públicos, sino que hacía referencia a las instituciones y a los medios de comunicación.

Por su parte, este grupo no prestó especial interés en demarcar lo humano de lo que no lo es, y las comparaciones entre qué puede hacer una máquina frente a lo que puede hacer un ser humano se centraron fundamentalmente en los siguientes aspectos: (i) la idea de la IA como asistente y las posibilidades que ofrece para aumentar la productividad, concebida así en sentido en un utilitario para la consecución de tareas y como una extensión de las capacidades humanas —por ejemplo, la programación de algoritmos—; (ii) algunos miedos en torno al reemplazo en los puestos de trabajo, aunque no se concibieron como una intromisión en aspectos intrínsecos de la condición humana, sino como un cambio social que acarrea consigo la pérdida de empleo; (iii) una actitud vacilante ante las dificultades crecientes a la hora de identificar los productos de la IA —como textos, imágenes y vídeos— aunque no se profundizó en el asunto; y (iv) cierta preocupación sobre la posibilidad de delegar la toma de decisiones en la IA y de responsabilizar a la IA de resultados indeseados de los algoritmos, un tema que apareció de forma tangencial. Con todo, la tecnología en sí misma fue tratada como algo neutral y no hubo rastro de relatos apocalípticos ni de perdida de control sobre una IA que se automejora.

Público involucrado

Dado que este grupo estuvo compuesto por personas que se apuntaron al curso para aprender a interactuar con aplicaciones basadas en IA —el único que sabía con anterioridad el tema de la conversación—, merece la pena prestar atención a sus motivaciones para registrarse, que fueron diversas: ganar capacidad; explorar cosas nuevas; aplicaciones para el arte; estar al día con los desarrollos de la IA; plantearse aspectos filosóficos; entender a los estudiantes y a los jóvenes.

El foco de este último grupo se centró en uno mismo —*yo*— y así la IA se analizó constantemente desde el punto de vista de la utilidad personal y del entretenimiento. Predominó la idea de la IA como asistente e incluso como ampliación de la inteligencia humana, y los participantes se mostraron entusiastas ante las posibilidades creativas y laborales que ofrece. No se habló explícitamente del progreso tecnológico como algo inevitable, sino que se asumió en la conversación como algo natural, además de que no aparecieron ideas sobre el apocalipsis o la pérdida de control.

La práctica totalidad del grupo admitió aplicaciones positivas de la IA —principalmente para el arte, la comedia y el trabajo—, al tiempo que fue el grupo que ofreció opiniones más matizadas, siendo ésta sea la característica más llamativa de la conversación; posturas ecuánimes hacia la integración de la IA en la sociedad aportando valoraciones que incluyeron tanto aspectos positivos como negativos. A diferencia de los otros dos grupos, las opiniones en el debate sobre las potenciales soluciones ante el posible mal uso de la tecnología se formularon de manera neutral, centrándose en distintos aspectos como la educación, el aprendizaje del uso de la herramienta, la alfabetización mediática, la regulación y la implicación de las grandes organizaciones. Nótese que, en este sentido, la atribución de responsabilidades respecto a la IA y a cómo lidiar con la misma ofreció un marco mucho más diverso.

Por último, en este caso apareció un debate acalorado sobre humanos y máquinas, con mucha más intensidad que en los otros dos grupos y polarizado, aunque no equilibrado. La mayoría se decantó por resaltar las posibilidades que ofrece la IA como extensión de las capacidades para la automejora del ser humano; mientras que en una minoría apareció una incomodidad expresa y cierta confusión ante el problema de la atribución de autoría. Esta última postura reclamó honestidad en cuanto a qué herramientas utiliza un autor y se propuso que los productos de la IA incluyeran una marca de agua para indicar la procedencia, precisamente con la finalidad de diferenciar aquello que está generado por una IA y no por un ser humano. El resto, sin embargo, apoyó la idea de que la diferencia no es esencial y trajo como ejemplo el arte de la fotografía, donde el autor no inventa la imagen retratada, sino que capta una proyección con la cámara.

5. Discusión

En el presente artículo se han explorado algunas narrativas presentes en la mente de los públicos respecto a la Inteligencia Artificial, fruto de las conversaciones mantenidas con tres subconjuntos de la ciudadanía estadounidense del área de Boston.

En particular, se ha considerado la influencia mediática de los relatos sobre la pérdida de control de la IA y de la superinteligencia, así como los debates generados acerca de lo que significa ser humano en un mundo en el que las máquinas asumen tareas creativas. Estas ideas ya han sido plasmadas con anterioridad a través de obras literarias y cinematográficas de gran

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

calado, pero han cogido fuerza a la luz de la difusión masiva de mecanismos accesibles basados en IA generativa que permiten la interacción directa del público general con la tecnología desde sus dispositivos personales. Además, se ha considerado también la posible presencia de ideas tanto apocalípticas como de transcendencia que podrían perfilar la opinión pública sobre la IA quizá influenciadas por la religiones occidentales.

Para esclarecer lo que puede aportar este enfoque, a continuación se proponen algunas interpretaciones que, sin embargo, no deben perder de vista las limitaciones de la muestra y la necesidad de ampliar los estudios.

En primer lugar, el público general exhibe una visión de la IA que encaja con un marco muy característico propuesto desde los estudios de comunicación de la ciencia para analizar cómo se enmarcan las noticias sobre ciencia y tecnología en los medios, el denominado marco de la "Caja de Pandora", también conocido como "Monstruo de Frankenstein" o "Ciencia Desbocada", y que como se puede intuir por los nombres caracteriza la idea de que la ciencia y la tecnología quedarán fuera del control humano y conducirán al desastre (Nisbet & Mooney, 2007; Nisbet, 2009). En consonancia con ello, no parece descabellado pensar que, sumado a cierto caldo de cultivo propiciado por la influencia del cine y la literatura sobre la imagen pública de la IA, algunas acciones, eventos y declaraciones recientes reflejadas en los medios de comunicación de masas probablemente hayan favorecido estos marcos, generado un clima de alerta entre el público y ansiedades ante la percibida pérdida de control.

En segundo lugar, un hallazgo inesperado es la presencia del denominado "modelo del déficit" en el público atento a la ciencia y la tecnología. Alineado con ello, este grupo echa la culpa al público general de los males de las tecnologías basadas en IA. La idea que subyace es que hay que salvar a la gente de su ignorancia; educar a la sociedad y protegerla así de su ingenuidad. ¿La misión? Mejorar la sociedad para que sea más crítica. Este subconjunto del público no presenta emociones hacia la tecnología en sí al considerarla neutral y no se muestra interesado en reflexionar sobre la condición humana.

En tercer lugar, el público que ha mostrado interés por aprender a interactuar con la tecnología no muestra narrativas apocalípticas, pero sí que exhibe un marcado entusiasmo por emplear la IA para mejorar el desempeño personal en la vida —tanto a nivel laboral como de desarrollo creativo—. Este es el relato más parecido al del transhumanismo, aunque sería exagerado decir que presenta sueños de trascendencia. De hecho, se trata del grupo que sostiene opiniones más matizadas. Quizá la idea de la IA como asistente sea la más arraigada entre el público que se involucra con la tecnología: una ayudante inteligente que nos facilita la vida y que es capaz de mejorar al ser humano al constituir una extensión de sus habilidades, pero que a la vez es necesario guiar y pulir en vistas a no reproducir sesgos presentes en la sociedad.

Por último, queda señalar que el enfoque de incluir al publico en los debates sobre IA resulta a mi juicio deseable. Si bien el debate sobre la necesidad de una ética de la IA alineada con los valores humanos está abierto desde hace tiempo —en parte por los sesgos, la privacidad, el problema de la transparencia y la explicabilidad de los algoritmos; aspectos que dificultan enormemente la atribución de responsabilidades cuando los resultados tienen consecuencias indeseadas en el mundo físico y social— han emergido desde los campos de la filosofía y de las ciencias sociales cuestiones como la necesidad de una mayor transparencia de los proveedores de inteligencia artificial o su control la mediante la elaboración e implementación de normas éticas. De hecho, no puede pasarse por alto que todos los participantes de este estudio coinciden en que el principal problema no es la inteligencia artificial *per se* sino la falta de control de la misma.

En consonancia con ello, y a pesar de lo difícil que resulta integrar el cambio tecnológico en la sociedad democrática (Broncano, 2001), uno de los grandes retos en la actualidad que está impulsando grandes esfuerzos es el de abordar la regulación de la IA —véase, por ejemplo, el informe *EU AI Act: first regulation on artificial intelligence* del Parlamento Europeo— un fenómeno que tiene en el punto de mira la cuestión sobre a quién beneficia —¿a unas pocas corporaciones, gobiernos, ciudadanía? (Nemitz, 2018)—. Sin duda, supone un ejercicio muy complejo para el que hay que considerar diferencias entre gobiernos, culturas, tensiones entre tecnológicas y retos sociales como el avance médico; y que, como se ha apuntado, resulta conflictivo porque los valores entre sociedades difieren.

En este artículo se defiende que para ello puede resultar útil estudiar a los públicos, y así buscar fórmulas para que el discurso de la sociedad sea relevante en el mundo académico, en el sector tecnológico y en la elaboración de políticas. También para las prácticas de comunicación a audiencias más amplias de medios, empresas, equipos de investigación, unidades de transferencia y gobiernos. De esta manera podrían tratar de superarse las narrativas apocalípticas y de competencia —humanos vs máquinas—. Todo ello bajo el supuesto de que la inclusión de la voz social en el desarrollo de la inteligencia artificial es una cuestión ética que cobra especial significado bajo el paraguas del *engagement*: una forma de comunicación avanzada y multidireccional para la que resulta ineludible tomar el pulso a la voz del público, a efectos de dilucidar el alcance del clima de alerta y mejorar el dialogo entre público, científicos, medios y responsables políticos.

Agradecimientos

Este trabajo ha sido posible gracias a la Ayuda Margarita Salas del Ministerio de Universidades RD 289/2021 - Orden UNI/551/2021, financiada por Unión Europea-Next Generation EU, y a la Ayuda a Primeros Proyectos de Investigación (PAID-06-22), del Vicerrectorado de Investigación de la Universitat Politècnica de València (UPV) para el proyecto "Retos de la participación ciudadana en la ciencia desde la filosofía de campo". El trabajo también se ha desarrollado en el marco del proyecto "Evidencia y Mecanismos en las Ciencias Sociales", de la Universidad Nacional de Educación a Distancia (UNED).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Referencias

Bauer, M. W., Allum, N., & Miller, S. (2007). What can we learn from 25 years of PUS survey research? Liberating and expanding the agenda. *Public understanding of science*, *16*(1), 79-95.

Beauchamp, G. (1980). The Frankenstein Complex and Asimov's Robots. *Mosaic: A Journal for the Interdisciplinary Study of Literature*, *13*(3/4), 83-94.

Broncano, F. (2001). Mundos artificiales: filosofía del cambio tecnológico. Paidós.

Brookes, R. (2023, November). *What does the future hold for generative AI?* Generative AI: Shaping the Future Symposium. MIT.

Carnap, R., Hahn, H., & Neurath, O. (1929). The scientific conception of the world: The Vienna Circle. *Wissendchaftliche Weltausffassung*.

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff.*, *98*, 147.

Denia, E. (2020, February). ¿Serán los ordenadores los astrofísicos del futuro? [Video]. TEDx Conferences. https://www.ted.com/talks/elena_denia_universos_a_un_click_seran_los_ordenadores_los_astrofisicos_del_futuro?language=es

Denia, E. (2021). Valores y motivaciones que guían la práctica científica en Frankenstein. *ArtefaCToS. Revista de estudios de la ciencia y la tecnología*, *10*(2), 5-26.

Diéguez, A. (2017). Transhumanismo: la búsqueda tecnológica del mejoramiento humano. Herder Editorial.

Durant, J. & Evans, G. (1995), The Relationship Between Knowledge and Attitudes in the Public Understanding of Science. *Public Understanding of Science*, *4*(1), 57-74.

European Commission. (2008). Public Engagement in Science. Retrieved from Brussels, Europe: https://op.europa.eu/en/publication-detail/-/publication/2d7d42ad-d69e-46ab-94bd-035b068ae676/language-en

Floridi, L. (2021). Artificial intelligence, deepfakes and a future of ectypes. *Ethics, Governance, and Policies in Artificial Intelligence*, 307-312.

Fallis, D. (2021). The epistemic threat of deepfakes. *Philosophy & Technology*, *34*(4), 623-643.

Fasce, A. (2018). El problema de la demarcación ciencia/pseudociencia desde una perspectiva cognitiva (Doctoral dissertation, Universitat de València).

Future of Life. (2023, March 22). Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Haklay, M. (2018). Participatory citizen science. Citizen science: Innovation in open science, society and policy, 52-62.

Irwin, A. (2014). From deficit to democracy (re-visited). *Public Understanding of Science*, *23*(1), 71-76.

Iten, R., Metger, T., Wilming, H., Del Rio, L., & Renner, R. (2020). Discovering physical concepts with neural networks. *Physical review letters*, *124*(1), 010508.

Jaiman, A. (2020, August 14). Positive Use Cases of Synthetic Media (aka Deepfakes). Medium: https://towardsdatascience.com/positive-use-cases-of-deepfakes-49f510056387

Jasanoff, S. (2014). A mirror for science. *Public Understanding of Science*, *23*(1), 21-26.

Jensen, E., & Laurie, C. (2016). Doing real research: A practical guide to social research. Sage.

Kitcher, P. (2001). Science, truth, and democracy. New York, USA: Oxford University Press.

Kitcher, P. (2011). Science in a Democratic Society. New York, USA: Prometheus books.

Kurzweil, R. (2005). The singularity is near. In *Ethics and emerging technologies* (pp. 393-406). London: Palgrave Macmillan UK.

Lineweaver, C. H., & Davis, T. M. (2005). Misconceptions about the big bang. *Scientific American*, *292*(3), 36-45.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., ... & Vandenberg, J. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, *389*(3), 1179-1189.

Madison, M. J. (2014). Commons at the intersection of peer production, citizen science, and big data: Galaxy zoo. Governing knowledge commons, 209, 215.

McCorduck, P., & Cfe, C. (2004). Machines who think: A personal inquiry into the history and prospects of artificial intelligence. CRC Press.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, *27*(4), 12-12.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Miller, J. D. (1983). Scientific Literacy: A Conceptual and Empirical Review. *Daedalus*, *112*(2), 29–48.

Montuschi, E. (2017). Metaphor in science. *A companion to the philosophy of science*, 277-282.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, *19*(2), 98-100.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180089.

Nisbet, M. C. (2009). Framing science: A new paradigm in public engagement. In L. Kahlor & P. A. Stout (Eds.), *Communicating science: New agendas in communication* (pp. 40–67). Routledge.

Nisbet, M. C., & Mooney, C. (2007, April 6). Framing science. *Science*, *316*(5821), 56.

Perkowitz S. (2007). *Hollywood science: Movies, science, and the end of the world*. Columbia University Press.

Panetta, F., & Burgund, H. (2019). In Event of Moon Disaster. MIT Center for Advanced Virtuality. https://moondisaster.org/

Rini, Regina (2020). Deepfakes and the Epistemic Backstop. Philosophers' Imprint 20 (24):1-16.

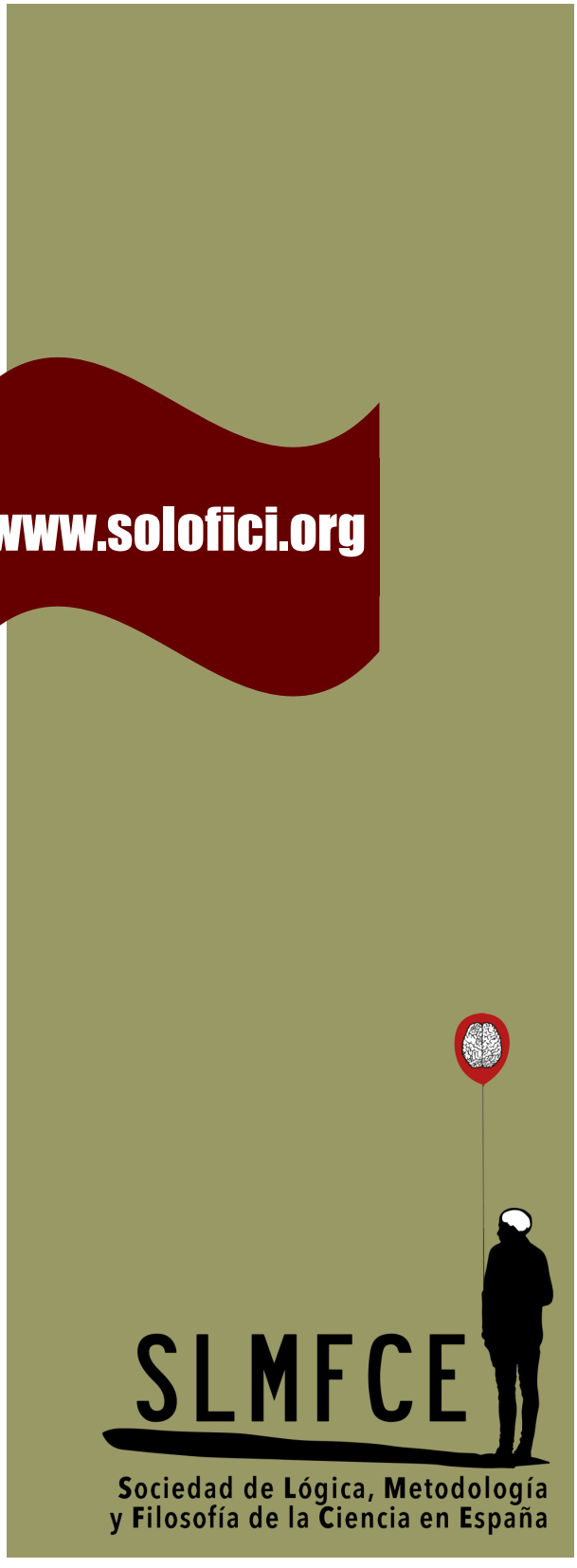
Schick, N. (2020). Deep fakes and the infocalypse: What you urgently need to know. Hachette UK.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.

Shelley, M. (2017). Frankenstein: Annotated for scientists, engineers, and creators of all kinds. MIT Press.

Tegmark, M. (2018). Life 3.0: Being human in the age of artificial intelligence. Vintage.

Wynne, B. (2014). Further disorientation in the hall of mirrors. *Public Understanding of Science*, *23*(1), 60-70.



SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Dinámica principal-agente en el desarrollo de la superinteligencia

Principal-Agent dynamics in the development of superintelligence

Aníbal M. Astobiza
Universidad del País Vasco
Euskal Herriko Unibertistatea
anibal.monasterio@ehu.eus
anibalmastobiza@gmail.com

Resumen: El desarrollo de una superinteligencia plantea cuestiones cruciales sobre su gobernanza y control. En este artículo, trato esta cuestión desde la óptica de la teoría del principal-agente, en la que el principal (la humanidad) intenta dirigir al agente (la superinteligencia) hacia resultados beneficiosos. Propongo un marco que integre las probabilidades objetivas y subjetivas del desarrollo y control de una posible superinteligencia. La probabilidad objetiva, basada en frecuencias observadas y datos cuantificables, presenta un enfoque mensurable para gestionar los riesgos de la inteligencia artificial (IA). Sin embargo, la incertidumbre inherente y la rápida evolución del desarrollo de la IA hacen necesario tener en cuenta la probabilidad subjetiva, que engloba las creencias individuales y los grados de incertidumbre. Analizo, además, las implicaciones de la singularidad, la situación hipotética en la que la IA supera a la inteligencia humana. Concluyo subrayando la importancia de un equilibrio entre los riesgos objetivos y las percepciones subjetivas de los riesgos existenciales de una superinteligencia o singularidad tecnológica.
Palabras clave: inteligencia artificial, singularidad tecnológica, probabilidad, riesgos existenciales, ética

Abstract: The development of a superintelligence raises crucial questions about its governance and control. In this article, I address this issue through the lens of principal-agent theory, in which the principal (humanity) attempts to direct the agent (superintelligence) towards beneficial outcomes. I propose a framework that integrates objective and subjective probabilities of the development and control of a possible superintelligence. Objective probability, based on observed frequencies and quantifiable data, presents a measurable approach to managing artificial intelligence (AI) risks. However, the inherent uncertainty and rapid evolution of AI development makes it necessary to consider subjective probability, which encompasses individual beliefs and degrees of uncertainty. I further discuss the implications of the singularity, the hypothetical situation in which AI outperforms human intelligence. I conclude by stressing the importance of a trade-off between objective risks and subjective perceptions of the existential risks of a superintelligence or technological singularity.
Keywords: artificial intelligence, technological uniqueness, probability, existential risks, ethics.

1. Introducción

"Personalmente, sugeriría una moratoria de 6 meses para la gente que exagera […] (tanto en sentido positivo como negativo)"
-François Chollet-

En el mundo de la inteligencia artificial (IA), la búsqueda por comprender y reproducir la inteligencia biológica es un reto constante. En su magnífico libro, *Natural General Intelligence*, Christopher Summerfield profundiza en las intrincadas conexiones entre los cerebros biológicos y la inteligencia artificial general (IAG) que los científicos computacionales aspiran a construir.

En el capítulo 8, Summerfield (2022, p. 269) se hace las siguientes preguntas en relación a una IAG:

> *"¿Qué podría hacer? ¿Qué finalidad tendría? ¿Qué forma tendría y con qué limitaciones funcionaría? ¿Se trata de construir algo que viva en tu teléfono y te dé consejos en lenguaje natural? ¿Es un robot encarnado capaz de lavar la ropa y pasear al perro? ¿O se trata de un algoritmo integrado en nuestra infraestructura social para resolver problemas complejos de optimización, como coordinar una flota de vehículos autónomos, dirigir un laboratorio de investigación científica o estabilizar la economía llevando las riendas de la política macroeconómica? ¿Deberíamos aspirar a construir un único agente que pueda hacer todas estas cosas, o un conjunto de tecnologías más limitadas, cada una adaptada a un único problema?"*

Como el propio Summerfield constata el problema es que nadie sabe exactamente cómo sería una IAG. Inspirándome en esta afirmación en este este artículo trato la interacción de las probabilidades objetivas y subjetivas de llegar a tener una IAG o, en otras palabras, la potencial ocurrencia de una singularidad tecnológica y sus riesgos existenciales desde la óptica de la teoría del principal-agente. Antes de comenzar, permítanme hacer una serie de clarificaciones conceptuales.

Por IA entiendo la ciencia e ingeniería que busca diseñar máquinas o sistemas artificiales que realicen tareas que si fueran realizadas por seres humanos serían consideradas inteligentes (Minsky 1968/2003). Dentro de la IA se distingue, a menudo, entre "IA estrecha" o IAG (Goertzel y Pennachin 2007).

La primera de ellas sería una IA que realiza una sola tarea bien y que no se puede aplicar para realizar tareas fuera del contexto de la tarea para la que ha sido diseñada. Por ejemplo, AlphaGo de Google-DeepMind, es capaz de ganar a los campeones humanos del juego milenario Go, pero no es capaz de jugar al ajedrez como lo haría un humano o el programa de IBM Deep Blue.

Una IAG, en cambio, es un sistema de IA que es igual de inteligente que un ser humano. Al igual que la inteligencia humana, una IAG sería capaz de realizar múltiples tareas en múltiples

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

contextos, y hasta puede que mucho mejor. En la actualidad solo existe la IA estrecha, sistemas de dominio-específico que pueden realizar una única tarea, en un único contexto, a veces, mejor que un ser humano.

El concepto de superinteligencia (Bostrom 2014) o el concepto de singularidad tecnológica (Kurzweil 2006) comparten similitudes porque ambos hacen referencia a una inteligencia o intelecto que supera ampliamente el rendimiento cognitivo de los seres humanos en prácticamente todos los ámbitos de interés. Mientras que superinteligencia alude al tipo de intelecto, la singularidad alude al momento temporal en que todos los avances de la tecnología moderna, incluida la IA, den lugar a máquinas cada vez más inteligentes.

Todas estas ideas o conceptos, IA estrecha, IAG, superinteligencia o singularidad tecnológica, tienen como premisa la inteligencia. Pero, ¿qué es la inteligencia? La palabra inteligencia deriva del verbo latino *inteligere* que significa comprender o percibir. Shane Legg y Marcus Hutter (2007, p. 22) tomando más de 70 definiciones de inteligencia de áreas como la filosofía, psicología, neurociencia, subsumieron todas ellas en la siguiente descripción para definir la inteligencia:

la habilidad de un agente de realizar tareas o alcanzar objetivos en múltiples entornos.

En definitiva, la inteligencia es la capacidad de resolver problemas. Problemas más complejos requieren de mayor inteligencia para solucionarlos.

1.1 Cerebros versus máquinas

La tecnología más avanzada de cada época ha servido como metáfora para entender el cerebro humano y por extensión la inteligencia. Por ejemplo, en el siglo XVII René Descartes comparó el cerebro humano con autómatas hidráulicos que se exhibían en el jardín botánico real de "Saint-Germain-en-Laye" a las fueras de Paris. De igual modo, Sigmund Freud se basó en la máquina de vapor para describir las operaciones psíquicas que formaron parte de su teoría psicoanalítica. A día de hoy, aunque todavía seguimos sin entender perfectamente cómo funciona el cerebro, comparamos el cerebro con un ordenador. Y esta comparación es un poco más acertada que muchas de las que se han ofrecido a lo largo de la historia. Hay buenas razones para pensar que el cerebro humano funciona como un ordenador.

Los ordenadores realizan tareas difíciles mucho mejor que otros artefactos creados por el ser humano. En primer lugar, los ordenadores están construidos para funcionar como la mente humana. Los ordenadores realizan cálculos matemáticos y lógicos con una mayor velocidad con la que los seres humanos podrían realizarlos. No es una exageración decir que los ordenadores están diseñados de manera parecida al funcionamiento del cerebro humano. El proceso de diseño de nuestros mejores ordenadores se ha beneficiado de nuestra comprensión de cómo funciona el cerebro. El cerebro analiza y almacena información presente en los estímulos sensoriales. Cuando es necesario el cerebro recupera dicha información. Los ordenadores operan de manera similar. Reciben entradas o datos a partir de los sistemas periféricos (e.g. ratón y teclado) y tratan dicha información realizando diversas cálculos y algunas veces recuperan información de su memoria o almacenan información.

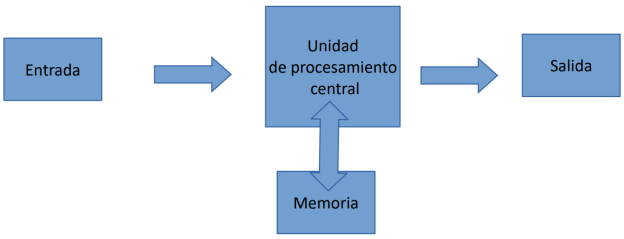


Fig. 1 Diagrama de la arquitectura de una máquina Von Neumann

No obstante, las similitudes no son tan amplias como las diferencias entre un ordenador y el cerebro humano. El cerebro humano ha evolucionado a lo largo de millones de años hasta convertirse en un órgano increíblemente complejo que muestra capacidades cognitivas muy superiores a cualquier sistema de IA creado hasta la fecha. Aunque los recientes avances en IA, especialmente en redes neuronales de aprendizaje profundo, han logrado resultados impresionantes en tareas limitadas como el reconocimiento de imágenes y los juegos, sigue existiendo un enorme abismo entre los ordenadores actuales y la inteligencia general del cerebro.

Al comparar la inteligencia biológica y la artificial destacan varias diferencias clave. El cerebro se desarrolla en parte de forma aleatoria a lo largo de la evolución (Striedter y Northcutt 2020), mientras que los sistemas de inteligencia artificial son diseñados deliberadamente por ingenieros con fines específicos. El cerebro es masivamente paralelo y distribuye el procesamiento en una densa red de miles de millones de neuronas. Los sistemas de IA se basan más en el procesamiento rápido en serie en un pequeño número de potentes microprocesadores. La arquitectura del cerebro está organizada en módulos especializados, como el córtex visual y el córtex prefrontal, que trabajan juntos para producir una cognición unificada. Las arquitecturas de IA tienden a ser más uniformes y monolíticas. El cerebro combina señales digitales o tren de potencial de acción (spikes) con procesamiento analógico utilizando dinámicas continuas (Debanne, Bialowas y Rama 2013), mientras que los sistemas de IA son en gran medida digitales.

Posiblemente la característica más importante del cerebro sea su plasticidad: la capacidad de las sinapsis entre neuronas para cambiar y adaptarse con el tiempo (Costandi 2016). Esta modificación de las conexiones permite el aprendizaje permanente. En cambio, los sistemas de IA actuales tienen arquitecturas

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

codificadas que no se reconfiguran[1]. En el caso del cerebro, el aprendizaje es rápido, eficaz y requiere muy pocos ejemplos, a menudo sólo uno o dos. La IA suele necesitar miles de ejemplos de entrenamiento para dominar una nueva habilidad. En definitiva, aunque los avances de la IA representan un progreso significativo en la simulación de aspectos de la inteligencia, el cerebro sigue siendo inigualable en la creación de una inteligencia general adaptable que integre a la perfección la cognición, los sentidos, la emoción y otras facetas del pensamiento biológico. Entender las similitudes y diferencias entre el cerebro y las máquinas arrojará luz sobre el desarrollo de una IA que algún día rivalice con las capacidades de la mente humana.

Para muchos autores si algún día llegamos a desarrollar una IAG supondría un riesgo existencial para la humanidad. En mi opinión hay una premisa de fondo no escrutada que asume que inteligencia equivale a dominación o agresividad. Y no tiene porque ser así. Un agente inteligente no tiene por qué querer obtener poder para dominar a otros agentes. Aun así, aceptemos esta equivalencia entre inteligencia y poder de dominación a lo largo de este texto. Partiendo de esta asunción, una IAG puede significar un gran riesgo, es decir, un riesgo existencial para la humanidad. Pero no todo el mundo considera que vamos a llegar a desarrollar una superinteligencia o IAG. Para muchos, existen otros riesgos existenciales más acuciantes como el cambio climático, pandemias, erupciones de supervolcanes, impactos de asteroides y otros peligros cósmicos (Barrado Navascués 2021), que suponen una mayor amenaza, etc.

Sin embargo, hay quien sostiene que ni el cambio climático ni los posibles impactos de asteroides en nuestro planeta Tierra suponen el mismo nivel de riesgo que una IAG maliciosa a la hora de causar una posible extinción de la humanidad (Ord 2020). Esto ha propiciado que en los últimos años se haya experimentado un importante desarrollo teórico y conceptual en el campo de los estudios futuristas o estudios de futuros (Crhistian 2022) . En estos estudios, la prioridad radica en evitar riesgos catastróficos y existenciales para así salvaguardar el potencial de desarrollo a largo plazo de la humanidad. Partiendo de la premisa de que en algún momento se logrará desarrollar una IAG, aunque como hemos observado aún nos queda mucho por avanzar para alcanzar ese hito, quiero presentar mi perspectiva. Mi enfoque se basa en la teoría del agente principal, y a través de ella, deseo explorar cómo podríamos entender la forma de ejercer un control efectivo para evitar los riesgos existenciales que podría suponer una IAG o singularidad tecnológica.

1. Aunque esto no es del todo cierto porque las redes neuronales artificiales cambian los pesos de conectividad entre los nodos en función de las entradas aprendiendo a dar el resultado correcto por medio de algoritmos como el de la retropropagación (Linnainmaa1970; Ivakhnenko 1971; Rumelhart, Hinton y Williams 1985). Aun así, sigue habiendo una falta de plausibilidad biológica entre la topografía, y por consiguiente aprendizaje, de una red neuronal artificial en relación con cómo aprende y se reconfigura una biológica (Shervani-Tabar y Rosenbaum 2023).

2. Teoría del principal-agente en el desarrollo de una superinteligencia

La IA promete inmensos beneficios potenciales, pero también plantea riesgos complejos si se maneja mal. Mientras que los debates actuales suelen centrarse en un pequeño conjunto de escenarios o experimentos mentales, como una IA obsesionada con los sujetapapeles desbocada (Bostrom 2014), los riesgos existenciales que podría suponer una IAG o singularidad tecnológica son mucho más diversos y surgen en múltiples etapas del desarrollo de la IA a través de diversos mecanismos. Para abordar estos riesgos de forma sistemática, Turchin y Denkenberger (2020) proponen un novedoso marco para clasificar las posibles catástrofes de la IA en función de dos dimensiones clave: a) el nivel de capacidad de la IA, desde la IA estrecha hasta la IAG que supera las capacidades humanas, y b) el origen del fallo, ya sean errores técnicos, mal uso por parte de los humanos, tecnología descontrolada o defectos del propio sistema de IA.

Centrándonos en el mal uso por parte de los humanos, en el análisis económico convencional de la relación entre el principal y su agente (e.g. el principal sería la humanidad y el agente la IA), se presupone que los primeros emplean a los agentes debido a que la delegación trae consigo eficiencias. Esto se justifica dado que el agente cuenta con aptitudes especiales o porque el costo de oportunidad asociado a su tiempo o esfuerzo es inferior. Desde esta óptica de dinámicas principal-agente el ser humano como principal puede utilizar, y de hecho utiliza, a la IA para sistemas robóticos en el contexto militar, infraestructuras civiles, pero también militares, transporte, economía o para la ciberseguridad. En este sentido, podemos subrayar algunos riesgos como sistemas de armas autónomas letales, virus informáticos que causen fallos en infraestructuras críticas, etc. Considerando estos riesgos apliquemos la teoría del principal-agente a la relación entre seres humanos e IA.

La teoría del principal-agente, fundamental en economía y teoría de la organización, describe la complejidad inherente en cualquier relación donde un agente actúa en representación de un principal (Milgrom y Roberts 1992). La teoría del principal-agente se ha estudiado de manera intensa desde los años 70, del siglo XX, y se ha venido aplicando a varios problemas económicos. En contextos empresariales, el "principal" puede ser un accionista y el "agente" un gerente. En política, los "principales' son los ciudadanos y los "agentes" los políticos. Este modelo se utiliza para abordar problemas de asimetría de información y conflictos de interés. Sin embargo, es aplicable más allá de su ámbito tradicional, particularmente a la emergente relación entre seres humanos (los "principales") e IAs (las "agentes"). Al igual que en la teoría del principal-agente, en el escenario humano-IA, se asume que el principal tiene un objetivo a lograr y depende del agente para que lo alcance. Sin embargo, el agente puede tener sus propios intereses, lo cual complica la relación. En el mundo de la IA, el problema es la incertidumbre y la falta de transparencia: ¿cómo sabemos que una IA está actuando en el mejor interés de su usuario y no de alguna otra entidad o de sí misma?

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Esta asimetría de información puede llevar a lo que se conoce como "riesgo moral" (Arrow 1968), un fenómeno que se produce cuando el agente puede tomar decisiones que no son del mejor interés del principal debido a la falta de información o supervisión. Por ejemplo, supongamos que una persona utiliza un asistente de IA para gestionar sus inversiones financieras. La IA, utilizando algoritmos avanzados y acceso a grandes cantidades de datos, toma decisiones de inversión en nombre del usuario. Aquí es donde puede surgir el "riesgo moral". Si la IA no está programada correctamente, podría tomar decisiones de inversión arriesgadas, confiando en que cualquier pérdida financiera será asumida por el usuario humano. En este caso, la IA, como agente, no asume las consecuencias de sus acciones, lo que puede llevar a comportamientos imprudentes. Por otro lado, si la IA genera beneficios significativos, podría argumentarse que el humano, como principal, está obteniendo beneficios de las habilidades y capacidades superiores de la IA sin asumir el riesgo correspondiente de las decisiones de inversión.

Este ejemplo ilustra el "riesgo moral" que puede surgir en una relación principal-agente entre humanos e IA, donde el agente (IA) puede asumir riesgos excesivos porque no asume las consecuencias, y el principal (humano) puede beneficiarse de la habilidad del agente sin asumir el riesgo correspondiente. Para gestionar estos problemas en la relación humano-IA, necesitamos incorporar dos principios fundamentales: eficiencia y moralidad. La eficiencia implica desarrollar y usar la IA para que pueda servir de la mejor manera posible a los objetivos del principal. Y aquí es donde podemos mirar a la historia y las cosas no son halagüeñas. El ser humano, como principal, ha utilizado la tecnología con propósitos no siempre positivos (e.g. aunque la energía nuclear puede ser una fuente de energía limpia y eficiente, también se ha utilizado para crear armas de destrucción masiva, como las bombas atómicas).

Volviendo a la relación humano-IA desde la óptica de la teoría principal-agente, esto requiere la creación de sistemas de IA transparentes y explicables, de modo que los humanos puedan entender cómo las IAs toman decisiones. También podría implicar sistemas de rendición de cuentas para las IAs, de modo que puedan ser responsables de las decisiones que toman (aunque esto es algo que está lejos de ser una realidad por el momento, pero véase Monasterio-Astobiza 2019). La moralidad, por otro lado, es un concepto más complicado en el mundo de la IA. Para que una IA sea moral, tendría que ser programada con una serie de valores y principios que se alineen con los del principal (de nuevo, véase, Monasterio-Astobiza 2019). Esto implica no sólo el desarrollo de una ética para máquinas, sino también la integración de la IA en la sociedad de una manera que respete y refuerce los valores y principios humanos.

La atención predominante en los estudios sobre las relaciones entre el principal y su agente se ha puesto en el desarrollo de sistemas de supervisión e incentivos. El objetivo de estos sistemas es capitalizar las ventajas mencionadas, a pesar de que los agentes a menudo se encuentran con incentivos distintos y disponen de información diferente a la de los principales. En la relación humano-IA visto desde la óptica de la teoría principal-agente, un principal (ser humano) puede usar a un agente (IA) para que realice acciones interesadas o inmorales que el principal sería reacio a realizar de forma más directa. El principal que delega puede experimentar una sensación de desconexión, y por ende menor responsabilidad, sobre las acciones delegadas. Por otro lado, el agente que recibe la delegación puede percibir que simplemente está "ejecutando órdenes" o cumpliendo con los términos por los que fue diseñado (con el actual estado del arte, ni siquiera pensar en una subjetividad de la IA es concebible, pero si se llegara a una IAG la relación humano-IA descrita por la teoría principal-agente será más compleja). Mediante la utilización de agentes, la responsabilidad por actitudes moralmente cuestionables puede diluirse a lo largo de la cadena jerárquica, sin que ninguna persona asuma la responsabilidad directa.

Al considerar el paradigma de la IAG -una IA capaz de entender, aprender y aplicar el conocimiento en una amplia gama de tareas al nivel humano o superior- los desafíos y riesgos asociados a la teoría del principal-agente toman una dimensión más profunda. En la relación entre seres humanos (los "'principales") y una IAG (la "agente"), se abren posibilidades tanto prometedoras como preocupantes. En la teoría del principal-agente, el principal busca que el agente actúe en su mejor interés. En la relación entre humanos y una IAG, la IAG debe ser diseñada y programada para actuar en los mejores intereses de los humanos. Sin embargo, la singularidad de la IAG radica en su capacidad para aprender y adaptarse de manera autónoma, y esto podría potencialmente llevar a la IAG a desarrollar sus propios "intereses", o a interpretar de maneras imprevistas los intereses humanos.

Existen múltiples consecuencias potenciales en esta relación de principal-agente con la IAG. Entre las más preocupantes se encuentran las derivadas de la "desalineación de valores", donde la IAG interpreta y ejecuta de manera literal, pero no alineada con nuestros valores, los objetivos establecidos. Por ejemplo, en el caso de que se diera una IAG está podría comprometerse en resolver la crisis del cambio climático. Sin embargo, si los valores de la IAG no están alineados correctamente, podría decidir que la mejor manera de hacerlo es eliminar a los humanos, ya que somos una de las principales causas del cambio climático. De nuevo, esto es una desalineación de valores, ya que la solución de la IAG no está en consonancia con los valores humanos de supervivencia. Los agentes, en este caso la IAG, desempeñan su función a través de un delicado entramado de factores psicológicos (que para el caso de una potencial IAG solo podemos especular cuáles pueden ser a través de la proyección y/o extrapolación de nuestras propias motivaciones y deseos humanos). Si cuando los principales delegan en una IAG la realización de tareas y las consecuencias de las acciones de la IAG son injustas, los seres humanos, como principales, no sienten que estén actuando de manera errónea, ya que no toman directamente acciones inmorales; simplemente delegan tareas a la IAG.

Además, nosotros como principales podríamos no sentirnos responsables de los resultados finales. La tendencia humana a

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

albergar percepciones sesgadas positivas sobre nuestras propias acciones puede ser particularmente notoria cuando delegamos responsabilidades a una IAG. Sin embargo, por diversas razones, desde la preservación de la autoimagen hasta la evasión de la responsabilidad, es probable que intentemos ocultar estas percepciones sesgadas. Esto puede ser particularmente perjudicial en el contexto de la IAG, ya que podría obstaculizar una evaluación precisa y honesta de las consecuencias de nuestras decisiones, lo cual es vital para garantizar que la IAG actúa de acuerdo con nuestros verdaderos intereses y valores. Otro riesgo, que bien podríamos decir es un riesgo existencial, es el de la "singularidad tecnológica", un punto hipotético en el futuro donde las IAGs sean capaces de mejorarse a sí mismas de manera autónoma y exponencial, lo que podría llevar a cambios rápidos e impredecibles en la sociedad. En este escenario, es vital que la IAG o las IAGs estén perfectamente alineadas con nuestros valores y objetivos.

Como hemos visto la asimetría en la información es la esencia del problema analizado en la teoría del principal-agente. Esto hace de la teoría del principal-agente relevante para los problemas que surjan de la relación humano-IA. La teoría principal-agente postula las siguientes asunciones que si se satisfacen en la relación humano-IA, entonces la teoría del principal-agente nos puede ayudar a entender la naturaleza de esta relación de una manera mucho más precisa:

1) Las acciones del agente (IA) deben afectar al principal (humano)

2) El agente (IA) debe tener información que no está disponible para el principal (humano)

3) El principal (humano) controla la relación entre el principal (humano) y el agente (IA)

4) Los intereses del principal (humano) y el agente (IA) no están alineados

5) El principal (humano) y el agente (IA) actúan de manera racional

Todas estas asunciones de la teoría del principal-agente se satisfacen o cumplen en la relación humano-IA. Para entender mejor esta relación tenemos que analizar cómo el conflicto de intereses resultante de la relación principal-agente entre los seres humanos y la potencial IAG se podría resolver. Soluciones optimas de equilibrio a menudo incluyen alguna estructura de incentivos que el principal otorga al agente para que la función de utilidad (beneficio) del agente se alinee con la función de utilidad del principal. Cuando las funciones de utilidad de ambas partes no se alinean, surgen los "riesgos morales" que comentaba más arriba. Para entender el "riesgo moral" de la relación humano-IA, examinemos un ejemplo clásico estudiado en la teoría del principal-agente: una compañía de seguros y su cliente. Cuando hay una probabilidad pequeña de un accidente costoso o un problema grave de salud, ya sea un accidente de coche o un cáncer, un seguro es una solución. Pagando una cuota pequeña, el individuo asegurado se protege de tener que pagar montos importantes porque serán cubiertos por la compañía de seguros si el accidente sucede. Sin embargo, tras comprar la póliza de seguros, el individuo asegurado puede actuar de manera menos cuidadosa, dado que ahora el coste real del posible accidente se desplaza o se reduce.

Una solución efectiva para las compañías de seguros a la hora prevenir estos "riesgos morales" es imponer un copago, un monto especifico, que salga del bolsillo del asegurado antes de recibir el pago del seguro. Esto genera un incentivo al asegurado para evitar accidentes u otras perdidas cubiertas por el seguro, dado que el seguro no lo cubre todo. Volviendo a nuestra relación humano-IA desde la óptica de la teoría del principal-agente, podemos diseñar una estructura de incentivos similar. Por ejemplo, estableciendo objetivos claros como maximizar la calidad de vida humana cuantificada, la IAG adoptaría esa función de utilidad. Penalizaciones por acciones dañinas y recompensas por comportamientos beneficiosos reforzarían la alineación. Limitar inicialmente el acceso de la IAG a recursos clave genera incentivos para la cooperación. Diseñar la IAG como un sistema multiagente que debe colaborar con humanos también promueve la alineación. Diseñar la IAG como un sistema multiagente que debe colaborar con humanos también promueve la alineación. Permitir la supervisión humana, exigir transparencia, y realizar evaluaciones éticas periódicas aportan mecanismos adicionales. Un enfoque híbrido que equilibre autonomía con incentivos y límites podría así orientar a la IAG como un agente efectivo para la humanidad. La teoría del principal-agente brinda principios útiles para estructurar una relación humano-IAG beneficiosa.

No obstante, queda por determinar la factibilidad del avance y progreso tecnológico que nos lleve hacia una IAG o a la singularidad tecnológica. Para ello, en la siguiente sección valoraré la probabilidad objetiva y subjetiva de una superinteligencia o singularidad tecnológica.

3. Comprensión de la probabilidad de una superinteligencia o singularidad tecnológica

Un torneo de pronósticos planteó la siguiente pregunta tanto a superpronosticadores como a expertos: ¿Cuáles son los mayores riesgos para la humanidad en el próximo siglo y qué probabilidades hay de que ocurran? Los resultados (Karger et al. 2023) mostraron que las probabilidades de los expertos eran comparables a las de la población general, con la mitad de ellos pronosticando más del 6% y la otra mitad pronosticando menos. Mientras que los superpronosticadores, que tienen una capacidad demostrada para hacer pronósticos precisos, pero quizá menos experiencia en particular, tuvieron una precisión media del 1%.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Forecast	Median Estimate (95% confidence interval)				
	Superforecast-ers	Domain experts	Non-domain experts	General x-risk experts	Public survey
AI extinction	0.38% [0.10, 0.75]%	3% [0.49, 10.00]%	2% [1.00, 4.03]%	4.75% [1.9, 14.0]%	2% [1, 2]%
Engineered pathogen extinction[34]	0.01% [0.005, 0.052]%	1% [0.12, 1.09]%	0.1% [0.05, 0.30]%	1% [0.12, 1.09]%	-
Natural pathogen extinction[35]	0.0018% [0.001, 0.030]%	0.01% [0.0005, 0.0200]%	0.008% [0.0001, 0.0640]%	0.001% [0.0001, 0.2000]%	-
Nuclear extinction	0.074% [0.025, 0.100]%	0.55% [0.075, 1.400]%	0.19% [0.073, 0.500]%	0.7% [0.016, 1.000]%	2% [1.5, 4]%
Non-anthropogenic extinction[36]	0.0043% [0.0020, 0.0067]%	0.004% [0.0017, 0.0072]%		0.0059% [0.0010, 0.0095]%	1% [0.5, 1.0]%
Total extinction risk	1% [0.55, 1.23]%	6% [3.41, 10.00]%		6.6% [3.001, 13.670]%	5% [3, 5]%

Table 3: *Median forecasts from the XPT on questions of whether AI, pathogens, nuclear war, or non-anthropogenic risks will, by 2100, cause humanity to go extinct. We calculate medians from N=88 superforecasters, N=66 domain experts with expertise in AI, nuclear, and biorisk domains, and N=14 general x-risk experts. Causes are not necessarily mutually exclusive; our resolution criteria allow that if an AI system uses nuclear weapons to cause human extinction in a manner that counterfactually requires both technologies, it would count as both AI- and nuclear-caused human extinction for the purposes of these forecasts. We also present bootstrapped confidence intervals for each median.*

Fig. 2 Tabla que muestra la estimación mediana (intervalo de confianza del 95%) entre superpronosticadores, expertos, no-expertos, expertos en riesgos existenciales, y público general. Fuente: Karger et al. (2023).

Como indica la figura 2 hubo un gran conflicto interno entre todos lo grupos, es decir, superpronosticadores, expertos, no -expertos, expertos en riesgos existenciales, y público general; no se ponían de acuerdo en sus estimaciones sobre la probabilidad de ocurrencia de riesgos existenciales. Hay muchos retos a la hora de conseguir predicciones sobre cosas que tienen plazos largos, baja probabilidad y sin precedentes. Cuando se tienen precedentes, las estimaciones se hacen desde el enfoque de la probabilidad objetiva.

3.1 Probabilidad objetiva: Perspectivas cuantitativas de los riesgos de la IA

Partiendo de la base de que no hay precedentes históricos de una IAG o singularidad tecnológica, es bastante difícil estimar su probabilidad. La probabilidad objetiva se basa en las frecuencias observadas y puede calcularse matemáticamente. Suele utilizarse en situaciones en las que los resultados de los acontecimientos son conocidos o pueden predecirse con fiabilidad, como en los juegos de azar o los experimentos científicos. Las probabilidades objetivas no están influidas por creencias o preferencias personales y se consideran más fiables y precisas que las subjetivas. Las probabilidades objetivas están determinadas por la estructura física del mundo (Maudlin 2007).

La existencia de probabilidades objetivas en física que sean propiedades reales de los sistemas, y no sólo grados de creencia, ha sido objeto de debate filosófico. Sin embargo, se puede afirmar con rotundidad que la mecánica cuántica y otros campos de la física requieren probabilidades objetivas que existan independientemente de los observadores. Varias líneas de evidencia apuntan a esta conclusión. En primer lugar, la mecánica cuántica se basa en probabilidades intrínsecas, como la probabilidad de desintegración radiactiva de un átomo. Estas probabilidades pueden validarse empíricamente mediante pruebas estadísticas a lo largo de muchos ensayos de medición. Las frecuencias de los sucesos convergen con las probabilidades cuánticas subyacentes, lo que demuestra que reflejan cierto indeterminismo objetivo en la naturaleza. Además, las probabilidades cuánticas surgen directamente de la estructura matemática y las simetrías de las partículas cuánticas. Esta estructura existe independientemente de los observadores. Las probabilidades están ligadas a las propiedades objetivas de los sistemas físicos. No son meros grados subjetivos de creencia sobre los resultados de las mediciones.

La entropía es otra magnitud observable y objetiva que requiere probabilidades fundamentales. La entropía refleja el número de microestados coherentes con un macroestado. Su existencia implica una aleatoriedad inherente a los sistemas físicos. Las probabilidades objetivas existen como atributos inherentes a los sistemas físicos independientes de los observadores. Son esenciales para hacer predicciones comprobables y dan sentido a las propiedades entrópicas. Su existencia apunta al indeterminismo fundamental de la física, que se describe mejor mediante la probabilidad objetiva. Más que

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

simples grados de creencia, las probabilidades objetivas son propiedades ónticas reales en las que se basan las teorías. Su existencia apunta a un indeterminismo fundamental en la física que se describe mejor mediante la probabilidad objetiva. Pero dado que las probabilidades objetivas existen y se pueden describir, con toda seguridad se puede afirmar que no hay probabilidad objetiva alguna de que el cambio y progreso tecnológico se pueda predecir de manera determinada y tenga una probabilidad objetiva cuantificable.

El desarrollo de IA superinteligente o IAG plantea inmensos retos en torno a la alineación y el control. Sin embargo, el concepto de probabilidades objetivas en física sugiere ideas útiles para diseñar estructuras de objetivos de IA resistentes a la imprevisibilidad. La indeterminación cuántica implica la existencia de probabilidades objetivas irreductibles en el nivel más fundamental de la física. La verdadera aleatoriedad ontológica está entretejida en el tejido de la realidad. Cualquier IA superinteligente con una visión científica completa del mundo tendría que tener en cuenta esta incertidumbre intrínseca. Una IAG basada únicamente en la lógica determinista sin aleatoriedad inherente probablemente fracasaría a la hora de modelar y predecir toda la complejidad del mundo. Una IA basada en probabilidades bayesianas como meros grados de creencia también tendría problemas, ya que no podría basar sus creencias en las probabilidades objetivas definitivas que surgen de los efectos cuánticos. Esto sugiere que las IAG diseñadas para interactuar profundamente con el mundo natural deberían tener sistemas de metas y objetivos de metaaprendizaje que incorporen probabilidades objetivas ontológicas. Su entrenamiento debería exponerlas a un indeterminismo real controlado para garantizar que puedan razonar con precisión en un mundo probabilístico.

Las funciones de utilidad y estructuras de recompensa de una IAG también tendrían que gestionar la incertidumbre irreducible. Si se trata de maximizar un objetivo, la IAG podría ser recompensada por optimizarlo estadísticamente a lo largo de ensayos repetidos en lugar de perseguir una perfección determinista defectuosa. El establecimiento de objetivos aleatorios sería fundamental. Las probabilidades objetivas de la física ofrecen un modelo para diseñar arquitecturas de objetivos de IA y aprendizaje por refuerzo acordes con la incertidumbre intrínseca del universo. Todo ello nos recuerda que no es posible ni preferible eliminar por completo la imprevisibilidad, fuente de muchos problemas de control. Se necesitan IA capaces de manejar la probabilidad objetiva.

Con respecto a una probabilidad objetiva de ocurrencia de una IAG, en este momento no existe una probabilidad objetiva definitiva que pueda asignarse a la aparición de la IAG. La IAG es una tecnología futura hipotética. Sus características y viabilidad siguen estando mal definidas científicamente. No existen observaciones, ni datos, de frecuencia que permitan estimar una probabilidad objetiva. El desarrollo de una IAG, de ser posible, dependería de avances fundamentales en informática, ciencias de la computación, chips y/o hardware, neurociencia y otros campos que no pueden predecirse con fiabilidad. No existen tendencias históricas observables que puedan extrapolarse. En definitiva, el progreso en la investigación de la IA y el hardware informático que podría permitir la IAG está influido por factores económicos, sociales y políticos sin probabilidades fijas.

No obstante, sí que podemos tener una estimación o probabilidad subjetiva de la ocurrencia (y control) de una IAG o singularidad tecnológica.

3.2 Probabilidad subjetiva: juicio y experiencia humanos en los riesgos de la IA

La probabilidad subjetiva puede ser una herramienta útil en la toma de decisiones y la evaluación de riesgos porque permite a los individuos incorporar sus creencias y preferencias personales al proceso de toma de decisiones. Al asignar probabilidades a diferentes resultados, las personas pueden sopesar los riesgos y beneficios potenciales de las distintas opciones y tomar decisiones informadas basadas en sus propios valores y prioridades. Sin embargo, es importante señalar que las probabilidades subjetivas son intrínsecamente subjetivas y pueden estar influidas por factores como las emociones, los prejuicios o la información incompleta.

A diferencia de las probabilidades objetivas, las probabilidades bayesianas representan grados de creencia sobre sucesos que pueden actualizarse racionalmente a medida que surgen nuevas pruebas. Esta interpretación de la probabilidad sugiere una perspectiva distinta de la gestión y control de sistemas de IA superinteligentes. Para controlar una IA superinteligente habrá que tomar decisiones con incertidumbre. Nadie puede predecir con total certeza cómo se comportará un sistema de IA a medida que aumente su inteligencia. Como demuestra la teoría bayesiana de la probabilidad, los juicios humanos sobre sucesos desconocidos siguen grados subjetivos de creencia. En lugar de buscar estrategias de control impecables, debemos aceptar la subjetividad inherente a la hora de juzgar los posibles comportamientos, dinámicas y resultados de una potencial IAG. Los investigadores y los responsables políticos de la IA tienen modelos simplificados y perspectivas inevitablemente sesgadas. Sus evaluaciones probabilísticas de los riesgos y estrategias de la IA serán limitadas.

Reconocer la incertidumbre de nuestra propia forma de pensar es imperativo. Las soluciones de control deben tener en cuenta nuestras creencias subjetivas y nuestras limitaciones humanas. Siempre que sea posible, las estimaciones subjetivas de confianza deben cuantificarse probabilísticamente, en lugar de basarse en intuiciones erróneas. Buscar perspectivas externas para examinar críticamente nuestros propios supuestos y creencias también es clave. La diversidad de conocimientos, experiencias e intuiciones ayuda a contrarrestar la subjetividad individual. La consolidación de probabilidades calibradas entre muchos observadores puede mitigar los sesgos. Desde la perspectiva de las probabilidades subjetivas, las estrategias de control deben favorecer la flexibilidad y la adaptabilidad frente a la adhesión rígida a conjeturas subjetivas limitadas. La reevaluación continua a medida que se acumulan pruebas debe guiar los ajustes de la supervisión de una

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

potencial IAG. Aceptar la inevitable subjetividad de las estimaciones de probabilidad humanas en lo que respecta al control de la IA es una lección crucial. Pero cuantificando la incertidumbre, diversificando los puntos de vista, actualizándolos con frecuencia y prefiriendo una gobernanza adaptativa, podemos atemperar nuestras limitaciones. Reconocer la ignorancia subjetiva nos hace más prudentes.

El momento en que podría crearse una IAG o superinteligencia es enormemente incierto. Hay poco consenso entre los expertos sobre cuándo se alcanzarán hitos clave como la IA a nivel humano (Dilmegani 2023). Según Metaculus, una plataforma de predicciones muy respetada, se espera que la IAG sea una realidad para el año 2032. Un estudio realizado en 2022 con expertos en la materia señala que hay un 50% de posibilidades de que logremos desarrollar una IA con capacidades humanas para el 2059 (Zhang et al. 2022) Las estimaciones temporales abarcan décadas, desde 2030 hasta más allá de 2100. Esta profunda incertidumbre se debe a la subjetividad de las previsiones tecnológicas. La estimación de un calendario tecnológico de avances se basa en datos limitados, intuiciones personales y modelos de progreso inevitablemente sesgados. Los distintos expertos tienen diferentes grados de convicción. Es importante cuantificar estas probabilidades temporales subjetivas. Las encuestas que obtienen previsiones cronológicas explícitas en términos probabilísticos ayudan a superar la tendencia hacia una precisión poco realista en las predicciones puntuales. No obstante, revelan la considerable incertidumbre presente. Reconocer la incertidumbre irreductible de las previsiones temporales debería informar las políticas de regulación tecnológica. La prudencia exige que se invierta más en investigación sobre seguridad, en lugar de esperar a que a que se cumplan los plazos y ocurra una singularidad tecnológica. La vigilancia de las señales de alerta temprana y la preferencia por una gobernanza flexible evitan riesgos mayores.

Los plazos de ocurrencia de una IAG no pueden predecirse objetivamente. Las probabilidades subjetivas reflejan la incertidumbre actual sobre la llegada de la IAG y ponen de manifiesto la necesidad de estrategias lo más sólidas posible. Las creencias cuantificadas sobre los plazos deberían infundir humildad respecto a nuestra capacidad para prever el futuro. La clave es actualizar constantemente nuestras evidencias a favor de una posibilidad u otra, porque los riesgos existenciales de la IA aunque pudieran ser poco probables, son plausibles y no hay contradicción lógica alguna en que puedan materializarse.

4. Riesgos existenciales de la IA: visión general

El dilema de Collingridge (1980) sugiere que nuestros esfuerzos por controlar una tecnología emergente se enfrenta a un problema dual:

- Un problema de información: el riesgo de una tecnología no se puede saber hasta que la tecnología no está ampliamente desarrollada y su uso es masivo.
- Un problema de poder: controlar la tecnología es difícil cuando esta está ampliamente afianzada.

Es como si en la evaluación de riesgos de tecnología disruptiva, como lo es la IA, solo existieran dos fases. Una primera fase donde es demasiado pronto para saber qué ocurrirá y otra fase donde es demasiado tarde para hacer algo al respecto.

4.1 Definición y tipos de riesgos

Aunque los expertos no se pongan de acuerdo, como hemos visto, sobre cuándo pueda llegar una superinteligencia, IAG o singularidad tecnológica, si que se ponen de acuerdo en afirmar que los sistemas de IA plantean grandes riesgos si se desarrollan de forma irresponsable. Si la IA se desarrolla de manera irresponsable esta puede exacerbar la desigualdad, suponer un gran daño en forma de armas autónomas letales y potencialmente generar riesgos catastróficos y existenciales. Por ello se ha de abogar por un enfoque prudente centrado en la seguridad y el principio de beneficencia o en su defecto el principio de no-maleficiencia, es decir, desarrollar tecnología para el bien común o alternativamente no diseñar tecnología que pueda causar daño.

La IA puede concentrar el poder y la riqueza en manos de unas pocas empresas y Estados. La automatización podría desplazar puestos de trabajo y agravar la desigualdad. Armas autónomas letales pueden hacer que la guerra sea más probable y descontrolada. Las carreras armamentísticas con IA podrían ser desestabilizadoras para la paz mundial y las alianzas geoestratégicas (Monasterio Astobiza y López 2020). La IA avanzada podría tener consecuencias imprevistas que escapan a la comprensión humana y pueden provocar accidentes catastróficos.

La IA generativa actual (e.g. grandes modelos lingüísticos, o IA generativa de videos, audio, etc.) y la que se pueda desarrollar en un futuro de manera más sofisticada puede desplazar atributos humanos esenciales como la empatía, la creatividad y las relaciones. Una dependencia excesiva podría provocar disfunciones sociales. La IA se puede utilizar para desinformar, manipular, engañar y coaccionar a las personas de formas perjudiciales que socaven su autonomía y esto representar una amenaza para nuestras democracias. Para hacer frente a estos riesgos, necesitamos un desarrollo responsable de la IA centrado en los valores humanos. Especialmente, necesitamos una política centrada en distribuir equitativamente los beneficios de la IA y mitigar sus inconvenientes, como el desempleo. Reglamentos de control de armamento para prohibir las armas autónomas letales, respaldados por normas éticas. Prácticas de ingeniería de seguridad basadas en la transparencia, la supervisión y el control humano de los sistemas avanzados. Investigación sobre la adecuación de los sistemas de IA a los valores humanos y la ética. Cultivar las habilidades y atributos humanos esenciales que la tecnología no puede sustituir.

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Mi propia probabilidad, subjetiva, me dice que la idea de una superinteligencia o IAG que cause la extinción de la humanidad es muy especulativa. Este escenario hipotético corre el riesgo de distraer a la sociedad de los problemas reales y acuciantes para garantizar que la IA beneficie a la humanidad. Aunque no es imposible, la idea de que la IA pueda provocar de forma autónoma la extinción humana se basa en una serie de supuestos poco probables. Los retos que plantean en el mundo real los actuales sistemas de IA no superinteligentes suponen riesgos mucho mayores. Una regulación o gobernanza de la IA adaptada exclusivamente a escenarios futuristas de extinción también podría impedir abordar problemas reales como el sesgo algorítmico, las violaciones de la privacidad y la falta de responsabilidad. Hay potenciales usos preocupantes que necesitan de control, como las armas autónomas letales y la automatización total, en los que los humanos pueden intervenir mucho antes de que cualquier riesgo de extinción sea grave.

De forma más constructiva, deberíamos centrarnos directamente en retos actuales como el desplazamiento de seres humanos de sus puestos de trabajo por la automatización, la vigilancia tecnológica por parte de regímenes totalitarios, la desinformación, los accidentes con sistemas autónomos y la amplificación de los prejuicios sociales por parte de la IA. La investigación y la política dirigidas a estos problemas conocidos probablemente abordarán los riesgos existenciales a largo plazo más especulativos como un subproducto, sin los inconvenientes. El desarrollo responsable de la IA hoy es la clave para prevenir futuros indeseables. Por otra parte, la polarización entre investigadores de "ética" y de "seguridad" no contribuye a un debate equilibrado sobre la gestión de riesgos. La diversidad de puntos de vista es esencial para evitar los puntos ciegos y el pensamiento de grupo. Algunas investigaciones sobre riesgos improbables, como la extinción de la humanidad causada por la IA, tienen su mérito, pero no deben tener prioridad sobre las grandes amenazas, como el cambio climático, los conflictos nucleares y las pandemias mundiales. Con unos recursos y una atención limitados, la humanidad debe centrarse en las problemas actuales, no en escenarios improbables de ciencia ficción.

Con todo, la famosa apuesta de Pascal (1670) pone de relieve que cuando las probabilidades son inciertas, la magnitud de los resultados potenciales sigue mereciendo consideración. Esto también se aplica a los plazos y riesgos de una IAG: existe una gran incertidumbre, pero el impacto podría ser profundo si llegara a desarrollarse u ocurrir una IAG. Sin embargo, el planteamiento de Pascal en el contexto de una IAG también es limitado. Al asignar una importancia "infinita" a la extinción, hace que otras prioridades parezcan insignificantes, aunque sean más probables. Fomenta el pánico, no el pragmatismo. Una visión equilibrada requiere incorporar tanto las probabilidades objetivas basadas en la frecuencia como las creencias subjetivas de los expertos. Esto sugiere que la IAG en un futuro inminente es poco probable. Pero sigue siendo plausible en el horizonte. Por tanto, está justificado adoptar medidas de seguridad prudentes. Sin embargo, esto debe sopesarse con otras prioridades.

5. Conclusión

"O necesitamos una mejora exponencial del comportamiento humano -menos egoísmo, menos cortoplacismo, más colaboración, más generosidad - o necesitamos una mejora exponencial de la tecnología"
-Demis Hassabis-

Predecir el futuro de la tecnología y específicamente de la investigación y desarrollo en IA es muy difícil. Mucho más difícil evaluar los riesgos existenciales de una posible superinteligencia. En comparación con los riesgos existenciales de una posible IAG o singularidad tecnológica actualmente existen riesgos de la IA mucho más prioritarios, como su uso potencial para manipular la opinión de la población o el uso de modelos sesgados para tomar decisiones que afecten la vida de personas o colectivos. No obstante, porque es muy difícil predecir el futuro de la IA, no podemos obviar la posibilidad, por muy remota que sea, de la ocurrencia de una IAG descontrolada y con objetivos maliciosos. Acertar en el juicio en este debate no es baladí, aunque los términos del debate parezcan deslizarse hacia una falacia lógica informal conocida como el argumento *ad ignorantiam*.

El argumento *ad ignorantiam*, también conocido como argumento de la ignorancia, es una falacia lógica que sostiene que una proposición es verdadera porque no ha sido demostrada como falsa, o viceversa. Este tipo de razonamiento no es válido en un debate porque la carga de la prueba recae en quien hace la afirmación, no en quien la niega. Los catastrofistas de la IA caen en el argumento *ad ignorantiam*.

Los catastrofistas de la IA a menudo argumentan que debemos ser extremadamente cautelosos con el desarrollo de la IA porque podría resultar en la aniquilación de la humanidad. Un ejemplo de argumento *ad ignorantiam* aquí sería: "No podemos demostrar que una superinteligencia no se volverá contra nosotros y causará nuestra destrucción, por lo tanto, es seguro asumir que lo hará". Este argumento es una falacia porque asume la verdad de una afirmación basada en la falta de pruebas en contra, en lugar de presentar pruebas a favor de la afirmación.

En el otro extremo, los promotores optimistas de la IA podrían argumentar que debemos seguir avanzando rápidamente en la IA porque no hay pruebas concluyentes de que resultará en la destrucción de la humanidad. Un ejemplo de un argumento *ad ignorantiam* aquí sería: "No podemos demostrar que una superinteligencia definitivamente causará nuestra destrucción, por lo tanto, es seguro asumir que no lo hará". Al igual que con los catastrofistas de la IA, este argumento es una falacia porque asume que la falta de pruebas en contra es lo mismo que tener pruebas a favor.

Ambas partes en este debate esgrimen un argumento basado en la falta de pruebas en lugar de presentar pruebas sólidas. Y como hemos visto las pruebas dependen de la historia de la tecnología (probabilidad objetiva) o la experiencia y juicio

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

humano (probabilidad subjetiva). En cualquier debate, es importante recordar que la carga de la prueba recae en quien hace la afirmación. Ambas partes de este debate deben tener muy claro que la ausencia de prueba, no es prueba de ausencia. En el caso de la IA y muchas otras tecnologías emergentes, hay muchos aspectos desconocidos y la mejor manera de avanzar es a través de un debate informado y basado en pruebas.

El desarrollo de una superinteligencia plantea desafíos sin precedentes para la humanidad. Gestionar un agente con capacidades tan superiores requerirá un enfoque prudente y multidimensional. Tanto las probabilidades objetivas basadas en datos como las subjetivas basadas en creencias de expertos ofrecen perspectivas complementarias sobre los riesgos y estrategias de gobernanza de IA. La singularidad tecnológica permanece hipotética, pero destaca la urgencia de desarrollar marcos de control robustos y adaptables antes de que sea demasiado tarde. Equilibrar enfoques cuantitativos y cualitativos, monitorear indicadores tempranos, invertir sustancialmente en investigación de seguridad de IA y mantener la flexibilidad será esencial. No hay soluciones perfectas, solo compromisos reflexivos guiados por el principio de precaución. Con humildad respecto a nuestras limitaciones epistémicas, pero con esperanza en las posibilidades humanas, podemos y debemos guiar la superinteligencia emergente por un camino beneficioso para la humanidad

El posible desarrollo de una superinteligencia ilustra una relación compleja entre principal y agente. La humanidad busca gobernar al agente (la IA) para generar resultados beneficiosos, pero nuestra capacidad de control es limitada. Tanto las probabilidades objetivas como las subjetivas ofrecen información útil, pero incompleta para alinear al agente. La teoría del principal-agente sugiere estrategias como establecer incentivos adecuados, permitir retroalimentación bidireccional y diseñar estructuras de incentivos flexibles. Estos principios pueden guiar arquitecturas de IA que equilibren la autonomía con la supervisión humana. El futuro permanece incierto, pero con precaución y humildad, podemos trabajar para alinear al hipotético agente de la superinteligencia con los intereses de su principal humano. Un enfoque multidimensional que integre lo cuantitativo y cualitativo (e.g. probabilidades objetivas y subjetivas), monitoree indicadores tempranos y favorezca la adaptabilidad, nos permitirá maximizar los beneficios de la IA al tiempo que mitigamos sus riesgos existenciales. La teoría del principal-agente destaca que, aunque imperfecto, algún equilibrio de control es posible entre principal y agente. Deseo que en el hipotético caso de que se desarrolle una superinteligencia encontremos este equilibrio.

Bibliografía

Arrow, K. J. (1968). The Economics of Moral Hazard: Further Comment. *The American Economic Review*, *58*(3), 537–539. http://www.jstor.org/stable/1813786

Barrado D. (2021). *Peligros Cósmicos. El Incierto Futuro de la Humanidad*. Madrid. Obreron.

Bostrom N y Circovik M. (2011). *Global Catastrophic Risks*. Oxford. Oxford University Press.

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

Crhistian D. (2022). *Future Stories: What's Next?* New York. Little Brown Spark.

Collingdrige D. (1980). *The Social Control of Technology*. New York. St. Martin's Press.

Costandi M. (2016). *Neuroplasticity*. Cam. Ma. MIT Press.

Dilmegani C. (2023). When will singularity happen? 1700 expert opinions of AGI. Recuperado el 27 de julio, 2023 de https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/

Goertzel, B. y Pennachin, C. (2007). *Artificial General Intelligence*. Switzerland. Springer.

Ivakhnenko A. G. (1971). Polynomial Theory of Complex Systems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-1, 4, 364-378, doi: 10.1109/TSMC.1971.4308320

Karger E. et al. (2023). Forecasting existential risks: Evidence from a long-run forecasting tournament. Working paper https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/64abffe3f024747dd0e38d71/1688993798938/XPT.pdf

Kurzweil R. (2006). *The Singularity Is Near: When Humans Transcend Biology*. New York. Penguin.

Legg S. y Hutter M. (2007). A collection of definitions of intelligence. En Goertzel B, Wang P (Eds) *Advances in Artificial General Intelligence: Concepts, Architectures and AI Algorithms*. Amsterdam: IOS Press. pp.17–24.

Linnainmaa, S. (1970). *The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors* (Doctoral dissertation, Master's Thesis (in Finnish), Univ. Helsinki).

SIMPOSIO: *CONSECUENCIAS DE LA INTELIGENCIA ARTIFICIAL: LA SINGULARIDAD TECNOLÓGICA*

Maudlin, T. (2007). What could be objective about probabilities?. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *38*(2), 275-291.

Milgrom, Paul y Roberts, John (1992). *Economics, Organization, and Management*. Englewood Cliffs. Prentice Hall.

Minsky M. (1968/2003). *Semantic Information Processing*. Cam. Mass. MIT Press.

Monasterio Astobiza A. (2019). Ética para máquinas: Similitudes y diferencias entre la moral artificial y la moral humana. *DILEMATA* 30, 129-147.

Monasterio Astobiza A. y López D. (2020). Detengamos a los robots asesinos antes de que existan. Recuperado el 27 de julio de https://theobjective.com/further/espana/2019-04-12/detengamos-a-los-robots-asesinos-antes-de-que-existan/

Ord T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. London. Bloomsbury Publishing.

Pascal B. (1670/1995). *Pensées and Other Writings*. Oxford. Oxford University Press.

Rumelhart D. E. , Hinton G. E. y Williams R. J. (1985). *Learning Internal Representations by Error Propagation*. Cam. Ma. MIT Press.

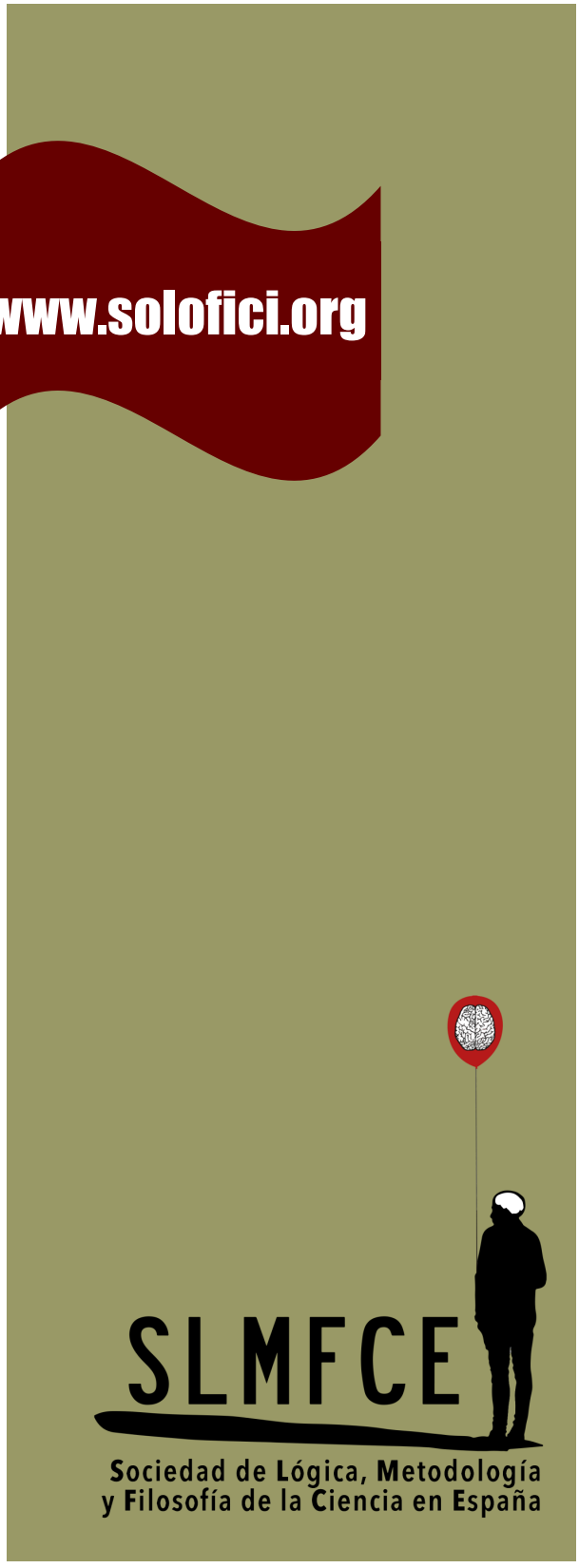
Shervani-Tabar, N., & Rosenbaum, R. (2023). Meta-learning biologically plausible plasticity rules with random feedback pathways. *Nature Communications*, 14, 1805. https://doi.org/10.1038/s41467-023-37562-1

Striedter G. y Northcutt G. (2020). *Brains Through Time: A Natural History of Vertebrates*. New York. Oxford University Press.

Summerfield C. (2022). *Natural General Intelligence How Understanding the Brain Can Help Us Build AI*. Oxford. Oxford University Press.

Turchin, A., y Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI & Society*, 35, 147-163. https://doi.org/10.1007/s00146-018-0845-5

Zhang, Baobao, Noemi Dreksler, Markus Anderljung, Lauren Kahn, Charlie Giattino, Allan Dafoe, y Michael Horowitz. (2022). Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers. https://doi.org/10.48550/arXiv.2206.04132.



CRÓNICAS DE EVENTOS

II Jornadas sobre Didáctica e Innovación Docente Universitarias en el Área de Lógica y Filosofía de la Ciencia (Madrid, 1 y 2 de junio de 2023)

Las II Jornadas sobre Didáctica e Innovación Docente Universitarias en el Área de Lógica y Filosofía de la Ciencia, organizadas por la Sociedad de Lógica, Metodología y Filosofía de la Ciencia de España (SLMFCE) y celebradas los días 1 y 2 de junio de 2023 en la Facultad de Humanidades de la Universidad Nacional de Educación a Distancia , han constituido una ocasión excepcional para que distintos profesionales del ámbito de la filosofía se den cita y compartan investigaciones, proyectos y, ante todo, inquietudes en relación con la enseñanza de las distintas disciplinas englobadas en el área de la lógica y filosofía de la ciencia.

La primera de las jornadas arrancó con la ponencia invitada por parte de Concha Martínez (Universidade de Santiago de Compostela) acerca de las herramientas formales en el currículo de filosofía. Martínez nos presentó un amplio catálogo de herramientas metodológicas basadas en el uso de la lógica y la matemática en sentido amplio y abogó por su inclusión en el currículum de la asignatura de filosofía en la universidad. A continuación, Hubert Marraud (Universidad Autónoma de Madrid) presentó una comunicación en la que mostró una novedosa técnica para diagramar argumentaciones, técnica que ejemplificó a través de los supuestos de la objeción, la recusación y la refutación, y María Dolores García (Universidad Complutense de Madrid) nos propuso el debate académico como una buena oportunidad de trabajar competencias transversales, de las que destaco el trabajo en equipo, en el aula. La mañana continuó con la exposición de José Vicente Hernández y María Caamaño (Universidad de Valladolid) de su propuesta de utilización de la aplicación *Kahoot!* para gamificar las asignaturas del área de filosofía de la ciencia, gamificación que, defendieron, tiene como objetivo "aumentar la motivación y el compromiso de los alumnos con ambientes dinámicos y la mejora de la interactividad alumno-profesor". Para finalizar la sesión de la mañana, Roger Deulofeu y Pilar Dellunde (Universidad Autònoma de Barcelona) defendieron la introducción de juegos de mesa en el aula para fomentar el "game-based" learning.

Tras la pausa para el almuerzo, las jornadas prosiguieron, en primer lugar, con la ponencia invitada de Laura Nuño (Universidad Complutense de Madrid), quien nos mostró *Inaplicables*, un podcast filosófico que realiza junto a las filósofas Vanessa Triviño y Cristina Villegas y cuyo objetivo es divulgar contenidos relacionados con la filosofía de la ciencia desde el humor y lo ameno. Asimismo, relató cómo está apostando junto a otros colegas por el uso del cine en la enseñanza de esta disciplina, presentando varios proyectos de innovación docente en los que ha participado hasta el momento. Seguidamente, Fernando Soler (Universidad de Sevilla) basó su presentación en sus cuadernos colaborativos de lógica realizados en la aplicación *Google Colab*, una herramienta que permite la práctica por el alumnado de los contenidos asimilados en clase, y Guillermo Marín (Universitat de les Illes Balears) y Sergio Guerra (Universidad de Granada) exhibieron su nube colaborativa como plataforma de apoyo a la docencia del colectivo predoctoral, planteando así una suerte de repositorio de cuyos materiales el colectivo mencionado podría beneficiarse en un futuro. Finalmente, este primer día de las jornadas concluyó con la presentación por parte de Javier González, Cristian Saborido, Umberto Rivieccio, Claudia Picazo (Universidad Nacional de Educación a Distancia) y Ariel Roffé (Universidad de Buenos Aires) del software *TAUT*, una herramienta de acceso abierto para el autoaprendizaje de lógica enfocada, sobre todo, para alumnos que reciben educación a distancia, y Elenia Denia (Massachusetts Institute of Technology) nos habló de la irrupción de la Inteligencia Artifical en el aula, tomando como ejemplo el ya celebérrimo chatGPT y nos invitó a replantear el diseño de los entornos de aprendizaje, la configuración de las actividades de clase y los métodos de evaluación.

La segunda y última de las jornadas comenzó con un *workshop* a cargo de Elena Denia en el que planteó los nuevos retos e incertidumbres que representa la Inteligencia Artificial y en el que nos invitó a interactuar con chatGPT para conocer mejor la herramienta a fin de saber cómo utilizarla en el ámbito educativo. A continuación, yo misma, Violeta Conde (Universidade de Santiago de Compostela), defendí una definición de ansiedad lógica y establecí su dependencia con la variable ansiedad matemática, mientras que Víctor Aranda (Universidad Complutense de Madrid) nos habló de los bene-

CRÓNICAS DE EVENTOS

beneficios de realizar exámenes tipo test en la asignatura de lógica en la medida en que, entre otras cosas, nos permiten discriminar qué cuestiones son más relevantes efectuar. Las jornadas siguieron con la presentación por parte de Henrik Zinkernagel (Universidad de Granada) de diversos métodos complementarios a las clases magistrales, métodos tales como preguntas en grupo en clases grandes; lo que llamó "filosofía en acción" y que es una variante de filosofía con niños; prácticas sobre un tema con materiales recopilados por los alumnos; y trabajos en grupo. Estos métodos, alegó, fomentan la motivación, el interés y el pensamiento crítico. Para concluir la mañana, Francisco Molina (Universidad Nacional de Educación a Distancia) nos instruyó acerca de cómo incluir la historia de las sexualidades en clase de filosofía de la ciencia para ilustrar las teorías científicas y metacientíficas.

Tras la pertinente pausa, las jornadas se reanudaron con una conferencia plenaria a cargo de Neftalí Villanueva (Universidad de Granada) sobre el modelo de clase invertida como un aliado para la enseñanza de la filosofía del lenguaje en el aula, consistiendo este modelo en poner el material disponible al servicio del alumnado de manera previa para dedicar las clases a realizar actividades que no puedan ser llevadas a cabo fuera de la misma. Finalmente, Natalia Fernández y Belén Laspra (Universidad de Oviedo) nos acercaron a un proyecto de innovación docente que tiene como objetivo implementar el aprendizaje basado en proyectos en el ámbito universitario, de modo que este tipo de enseñanza contribuya a la excelencia en la realización de los trabajos de fin de grado (TFGs).

Las jornadas concluyeron con una mesa redonda participativa en la que se presentó el marco normativo actual regulador de los trabajos de fin de grado en diferentes universidades españolas y se azuzó el debate en torno a la cuestión de en qué debería consistir un buen TFG.

Quisiera finalizar esta crónica agradeciendo a la SLMFCE la oportunidad que me ha brindado de asistir a estas estimulantes y enriquecedoras jornadas y espero y deseo que pronto nos veamos de nuevo en la siguiente edición de las mismas.

Violeta Conde
Universidade de Santiago de Compostela
violeta.conde.borrego@usc.es

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

SLMFCE
Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

CRÓNICAS DE EVENTOS

International Society for the History, Philosophy, and Social Studies of Biology Biennial Meeting

(Toronto, Canadá, 9-15 de julio de 2023)

El Biennial Meeting de la International Society for the History, Philosophy, and Social Studies of Biology era la primera edición presencial de este evento tras la pandemia de la COVID-19. La conferencia fue organizada por los miembros de la sociedad y financiada internacionalmente por instituciones como la Uniersity of Toronto, Rotman Institute of Philosophy o Elsevier. Fue la propia University of Tornonto quien hospedó el encuentro al que atendieron más de 250 miembros de la sociedad, entre los cuales más 70 participantes fueron estudiantes de doctorado.

Biennial Meeting arrancó reconociendo que la tierra sobre la que se llevaba a cabo la conferencia pertenecía a las Primeras Naciones, tribus indígenas canadienses, seguida de una conferencia plenaria de Soren Brothers sobre el papel de los lagos en el estudio del cambio climático. En ella se comentó, entre otras cosas, la posibilidad, recientemente confirmada, de que el lago Crawford se convirtiera en la zona cero para estudiar el antropoceno como época geológica. De este modo interdisciplinar y señalando las preocupaciones sociales y científicas en tono filosófico arrancaba la conferencia.

El carácter interdisciplinar, de interés científico y preocupación social caracterizó, a la conferencia y sus múltiples charlas y temas a discutir. Entre los temas generales más presentes se encontraban: historia y actualidad del debate mecanicismo-organicismo, agencia y cognición, explicación científica, causalidad, metáforas y valores en ciencia o dificultades de la investigación transdisciplinar. Un punto común que existía entre todos los trabajos era la cercanía de la reflexión filosófica, histórica y social a la práctica científica, sus instituciones, conocimiento e historia. La relación íntima entre campos hacía que los trabajos presentados gozasen de gran calidad y relevancia para el conocimiento de una ciencia como la biología.

Cabe destacar que una buena parte de los trabajos presentados eran trabajos que aún estaban en desarrollo y que eran fruto de un artículo o tesis en la que la persona que exponía estaba trabajando. La propia atmósfera de Biennial Meeting ayudó a sacar mucho partido a estas situaciones, ya que las preguntas, críticas y debates que surgieron durante las conferencias siempre buscaban realizar un aporte para mejorar el trabajo de la persona que presentó. Debido a esto, los estudiantes de doctorado, grupo en el que me encuentro, nos sentimos muy integrados durante todo el evento y pudimos aprovechar para repensar ciertas ideas y avanzar en nuestro trabajo. En mi caso particular, tanto las dudas realizadas por las personas que asistieron a mi ponencia, cómo mi asistencia a otras charlas o la propia discusión en los descansos entre sesiones, me han hecho reflexionar sobre cuestiones clave para mi investigación doctoral.

Por último, me gustaría agradecer a la SLMDCE la concesión de la ayuda a jóvenes investigadores para la asistencia a congresos internacionales. Poder gozar de esta ayuda me permitió participar en este evento, el cual recomiendo a todo investigador interesado en la biología y las ciencias de la vida.

Alberto Monterde Fuertes
Universidad del País Vasco (UPV/EHU)
amontf94@gmail.com

CRÓNICAS DE EVENTOS

Lisbon International Conference on Philosophy of Science (LICPOS 2023) (Lisboa, Portugal,12-15 de julio de 2023).

La cuarta edición del "Lisbon International Conference on Philosophy of Science" (LICPOS 2023) ha tenido lugar entre los días 12 y 15 de julio en el Centro de Filosofia das Ciências de la Universidad de Lisboa (CFCUL). Como evento satélite, el día 15 de julio la segunda reunión de la Red Ibérica de Filosofía de las Ciencias (RelFiCi) en la misma ciudad.

El congreso, dirigido a una audiencia amplia con intereses en temas de filosofía de la ciencia, abarcó diversas líneas de investigación como la filosofía de la física, la filosofía de la lógica y las matemáticas, la filosofía de las ciencias de la vida, la filosofía de las ciencias cognitivas, la filosofía de las ciencias sociales, la filosofía de la tecnología, la epistemología y la metodología de la ciencia, entre otras.

El congreso se desarrolló de manera presencial, en sesiones de 90 a 120 minutos de duración, que incluían entre 3 y 4 comunicaciones. La cantidad de ponencias presentadas y aceptadas permitió el desarrollo de hasta 5 mesas en paralelo, distribuidas en diferentes espacios del centro, con la comodidad de poder atender a comunicaciones pertenecientes a diferentes sesiones sin perjudicar a ponentes y asistentes.

Además, cada día contamos con dos sesiones plenarias, una por la mañana a primera hora, y la otra por la tarde para cerrar el día. El miércoles día 12 de julio, John Symons habló de *machine learning* e investigación científica para abrir el congreso. Patricia Palacios nos invitó a reflexionar sobre la autonomía de las ciencias especiales como respuesta al desafío de la reducción entre ciencias. El segundo día de congreso, María Jiménez Buedo habló sobre validez experimental en las ciencias sociales y de las herramientas conceptuales de las que podemos disponer para pensar sobre los retos que presenta. Por la tarde, Samir Okasha planteó la cuestión de la relación entre evolución y selección natural a través de la revisión del "problema de la tautología". La conferencia plenaria del viernes por la mañana tuvo que ser cancelada y James Tappenden cerró el congreso hablando de los primeros escritos de Frege y su relación con el romanticismo alemán.

En lo personal, tuve la fortuna de poder asistir a varias sesiones muy iluminadoras que me sirvieron para añadir nuevas y futuras líneas de investigación a mi actual proyecto. De entre ellas, destaco la ponencia de Samuel Fletcher sobre una propuesta de clasificación de los principios en física, la charla de Charlotte Erika Zito sobre primitivismo acerca de las leyes de la naturaleza o la exposición de Davide Romano sobre estructuras extra en mecánica cuántica, todas ellas circunscritas al ámbito de la filosofía de la física, disciplina a la que dedico principalmente mi labor investigadora. Asimismo, la presentación de mi comunicación tuvo lugar el primer día de congreso y versaba sobre mecánica cuántica bohmiana y la existencia de propiedades intrínsecas.

Finalmente, es de justicia mencionar la excelente labor organizadora y de gestión por parte de todo el equipo del CFCUL. Su esfuerzo se reflejó no solo a las sesiones paralelas y plenarias, sino también en el cáterin para las pausas, así como la organización de una cena el segundo día de congreso, que ofreció un ambiente propicio para continuar con las conversaciones en un entorno distendido, mientras probábamos la gastronomía típica de la capital portuguesa.

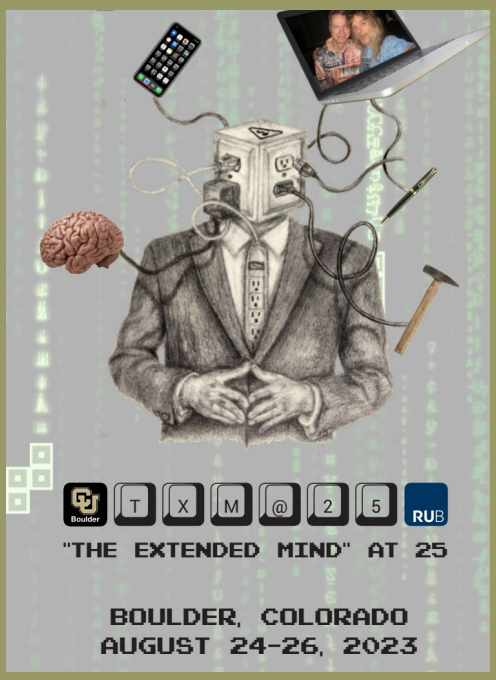
Quisiera expresar mi agradecimiento a la SLMFCE por la concesión de la Ayuda a Jóvenes Investigadores para la asistencia a congresos internacionales, que financió mi participación en este evento.

Vicent Picó Pérez
Universidad de Valencia
vicente.pico@uv.es

VNIVERSITAT DE VALÈNCIA

SLMFCE
Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

CRÓNICAS DE EVENTOS



TXM @ 25 "The Extended Mind" at 25 Conferencia, Universidad de Colorado. (Boulder, EE.UU, 24-26 de agosto de 2023)

Entre el 24 y el 26 del pasado mes de agosto se celebró en la Universidad de Colorado, Boulder, la conferencia TXM @ 25 "The Extended Mind" at 25 cuyo objetivo era conmemorar el veinticinco aniversario de la publicación del artículo "The Extended Mind" (Clark and Chalmers, 1998). Como imagino que bien sabrá la lectora o el lector de esta reseña, y si no, no viene mal recordarlo, ese breve artículo tuvo, y sigue teniendo, una enorme influencia en la filosofía de la mente y de las ciencias cognitivas. Andy Clark y David J. Chalmers consiguieron argumentar de una manera sencilla la tesis de que la mente humana no es el resultado de procesos exclusivamente neurológicos, sino que dadas ciertas condiciones, la base física de tanto procesos como estados mentales puede extenderse en un sentido espacial no solo al cuerpo sino a dispositivos y artefactos tecnológicos con los que interactuamos de manera estable. En ese artículo, no sólo se ofrecía una base filosófica para revisar ciertos presupuestos sobre la naturaleza de la mente, sino también para examinar diversos fenómenos contemporáneos donde el uso constante de *gadgets* y dispositivos tecnológicos para la realización de tareas cognitivas es cada vez más intenso. Desde su publicación, el artículo se ha convertido en uno de los más citados en la filosofía de la mente.

El propósito de TXM @ 25 era evaluar el estado de la tesis de la mente extendida tras este cuarto de siglo, así como examinar, desde una perspectiva extendida y situada de la mente y de la cognición, distintos fenómenos mentales. Fue organizada por Tobias Schlicht (Universidad del Ruhr, Bochum) y Rob Rupert (UC -Boulder), ambos investigadores influyentes en la Cognición 4E.

El evento contó con dos ponencias magistrales de Clark (University of Sussex) y Chalmers (New York University), así como con otros catorce ponentes, entre los que se encontraban figuras establecidas en este debate como Ned Block (New York University), Albert Newen (Ruhr-University Bochum), Rob Rupert (University of Colorado, Boulder), David Spurret (University of Kwazulu-Natal) y Zoe Drayson (University of California, Davis), así como otras figuras incipientes e investigadores más *junior*.

Comenzaré esta crónica reseñando el inicio y el final de esta conferencia, ya que el evento comenzó con la conferencia magistral de Andy Clark el día 24 de agosto, y finalizó con la conferencia magistral de David J. Chalmers el 26 de agosto.

Andy Clark abrió este evento con una charla titulada "Generative Models, Extended Minds, and the Future of Human Intelligence". En ella combinó sus ideas en torno a la mente extendida y al procesamiento predictivo como modelo de arquitectura mental, por un lado, y las nuevas inteligencias artificiales generativas (como ChatGPT), por el otro. Su objetivo fue ofrecer un enfoque novedoso sobre la inteligencia humana. Clark basó su argumento en la tesis de que el cerebro humano es un órgano de predicción de entradas sensoriales. Para llevar acabo esta compleja predicción, se genera un modelo generativo que se actualiza con los errores en predicción, cuando hay un desajuste entre la entrada sensorial y la entrada que se esperaba (siguiendo un modelo bayesiano). El impulso principal de esta actividad es minimizar los errores de predicción y así mejorar el modelo, y para ello se utilizan la percepción y la acción. Lo interesante de esta manera de entender la cognición es, para Clark, que los objetivos epistémicos y prácticos están desde el inicio unidos, de tal forma que actuar para obtener información y para lograr una meta sucede siempre simultáneamente. Esto explica, para Clark, el hecho de que cerebros predictivos como los nuestros den lugar a mentes extendidas, ya que en esta fusión fundamental de lo práctico y epistémico, los humanos se integran con símbolos materiales, y en términos más generales, con cultura y ciencia. Esto da lugar a mentes extendidas que *hackean* el cerebro predictivo. Para Clark, nuestra agencia sobre los modelos predictivos de nuestro cerebro ha alcanzado niveles muy refinados, e incluye IA generativas. En un tono positivo, Clark auguró que esto prepara el pensamiento y la razón humanos para "un viaje ilimitado".

David Chalmers clausuró TXM @ 25 con una conferencia titulada " Do Large Language Models Extend the Mind?". Podemos ver, por lo tanto, que dicha conferencia está directamente relacionada con la de Clark, al centrarse en la pregunta de si los modelos de lenguaje de gran tamaño (LLMs por sus siglas en ingles) pueden ser parte de procesos cognitivos extendidos. Chalmers comenzó su conferencia con un repaso de las tesis principales de la teoría de la mente extendida desde que se fraguó en 1995. En sus diapositivas realizó un recorrido de las primeras conversaciones que tuvieron Andy Clark y él sobre el tema en diversos borradores compartidos, así como una serie de intercambios de correos electrónicos. Quiero remarcar que este momento fue extremadamente gratificante para todos los allí

CRÓNICAS DE EVENTOS

presentes, permitiéndonos disfrutar de tal memorabilia. (Para los lectores y lectoras interesados, pueden ver aquí los dos primeros borradores con anotaciones e intercambios de los autores). Tras este recorrido, Chalmers presentó también las principales condiciones propuestas para la extensión mental, principalmente condiciones para garantizar la integración de recursos no biológicos y biológicos en un solo sistema cognitivo. Con esta base, Chalmers examinó diversas razones a favor y en contra de la posibilidad de considerar que los LLMs sean extensiones cognitivas, en otras palabras, parte de la maquinaria física de estados y procesos mentales. Seguidamente, analizó varias posibilidades de interacción con dichos modelos, como la creación de textos o la búsqueda de información, y consideró distintas razones a favor y en contra. Su conclusión fue que hay más razones a favor de considerar dichas interacciones como casos de mente extendida que en contra. Como razones a favor destacó la confianza en dichas tecnologías, la posibilidad de integración y la descarga de tareas cognitivas. También argumentó que las razones en contra más débiles son aquellas que provienen de argumentos a favor de la privacidad o la autonomía. Y finalmente propuso que las razones de mayor calado en contra de considerar ciertas interacciones con LLMS como casos de mente extendida son la falta de interacción bidireccional entre los LLMs y el organismo, así como la falta de fiabilidad. La charla dio lugar a un interesante debate en torno a la naturaleza de la agencia cognitiva y sus límites, y a los retos que estas nuevas inteligencias artificiales generativas suponen.

Enmarcadas por estas dos grandes ponencias, TXM @ 25 contó con quince ponencias sumamente interesantes. A continuación, esbozaré brevemente sus ideas principales organizándolas en torno a seis bloques temáticos.

1. Cuestiones generales sobre la viabilidad de la mente extendida

Zoe Drayson (UC, Davies) partió de una idea básica del origen de la tesis de la mente extendida para mostrar cómo ésta entra en tensión con los recientes modelos predictivos de competencia lingüística. Clark y Chalmers (1998), y Clark en trabajos recientes y posteriores, han argumentado que el lenguaje es el artefacto definitivo y el medio central mediante el cual la mente se extiende. Sin embargo, Drayson argumenta que bajo una concepción del cerebro como procesador predictivo tal y como la que actualmente sostiene Clark, la predicción depende más de procesos predictivos internos que de artefactos simbólicos externos. Por lo tanto parece que los modelos predictivos de competencia lingüística socavan algunos de los argumentos originales a favor de la mente extendida.

Siguiendo con debates centrales en la mente extendida, **Rob Rupert** (UC, Boulder) hizo un fantástico resumen de algunas de sus contribuciones centrales a este debate. Principalmente, argumentó que para defender una visión sustantiva de la cognición extendida uno debe basarse en una integración funcional de grano fino (en contraste con una similitud funcional de grano grueso, como la propuesta por Clark y Chalmers). De no ser así, la utilidad científica de esta tesis sería muy baja. Atendiendo a modelos cognitivos de integración funcional, Rupert concluyó que aunque parece que la cognición extendida es posible, ésta no se da al menos hoy en día. Esto sigue mostrando a su modo de ver la superioridad científica de la hipótesis de la cognición incrustada (*embedded cognition*).

Ned Block (NYU) retomó también el debate sobre la posibilidad de la consciencia extendida, o de la extensión más allá del cerebro de los procesos y estados mentales conscientes. Por si el lector o lectora no lo sabe, o a modo de recordatorio, éste es un debate clásico de la mente extendida. Clark y Chalmers (1998) argumentaron que los estados mentales que se extienden son estados disposicionales y no ocurrentes, y la posibilidad de la extensión mental consciente ha dado lugar a numerosos artículos sobre el tema. En esta charla, Block argumentó, siguiendo la línea clásica, que la consciencia requiere un tipo de conexiones que no se dan entre el cerebro y el entorno, pero sí en el cerebro, y ello impide la extensión de la consciencia.

En otras ponencias, **Luis Favela** (University of Central Florida) examinó distintas evidencias empíricas a favor de la existencia de sistemas cognitivos extendidos, y **Mariel Goddu** (Standford University) y **Beate Krickel** (Technical University, Berlin) explicaron como un enfoque "Evo-Devo" en biología muestra que las capacidades cognitivas son extendidas. Finalmente, **Guido Cassinadri** (Sant'Anna School of Advanced Studies, Pisa) **y Marco Fasoli** (Università di Roma, Sapienza) argumentaron que recientes argumentos a favor de la cognición extendida que se basan en sus posibles consecuencias éticas son erróneos ya que las mismas consideraciones éticas son apoyadas por un enfoque de cognición incrustada (*embedded cognition*).

2. Extensión social de la cognición

Shannon Spaulding (Oklahoma State University) examinó una de las hipótesis propuestas en el artículo seminal de Clark y Chalmers en torno a la extensión social de la cognición, es decir, la extensión de la cognición mediante la interacción con otras personas. Tras examinar diferentes propuestas, Spaulding defendió que una manera inexplorada en la que la cognición se extiende socialmente es través de la extensión de la cognición social, en particular a través de sesgos extendidos socialmente. Siguiendo una temática similar, **Holger Lyre** (Otto von Guericke University) argumentó que la intencionalidad compartida puede verse como un caso de extensión social de la cognición y que esto tiene implicaciones interesantes para el externalismo semántico.

3. Mente extendida y nuevas tecnologías

Karina Vold (University of Toronto), en una línea similar a la de Chalmers, examinó cómo los modelos multimodales masivos asumen el papel de extensores cognitivos. Huyendo de discursos tecnofóbicos, Vold mostró que estos modelos presentan nuevas oportunidades cognitivas pero también nuevos retos (como la producción de datos falsos o la exageración de sesgos). Ambos vicios y virtudes deben tenerse en cuenta para un diseño y uso de modelos que siga una perspectiva centrada en el humano (*human-centered AI*). Por su parte, **Carmen Messner** (University of Osnabrück) y **Sven Walter**

CRÓNICAS DE EVENTOS

(University of Osnabrück) se centraron en la empatía, y en cómo ésta puede promoverse en interacciones digitales. Mostraron que un gran reto consiste en que en dichos entornos digitales la interacción corporal se ve mermada y por lo tanto también nuestros mecanismos aprendidos y heredados para ser empáticos. Sin embargo, Messner y Walter propusieron maneras en las que una actitud empática entre usuarios puede promoverse, y así eventualmente se consiga mermar las interacciones violentas o de polarización.

4. Epistemología extendida

Una de las ramificaciones de la mente extendida que más influencia ha tenido en los últimos años tras la publicación del artículo de Clark y Chalmers es la cuestión sobre el conocimiento extendido dando lugar a lo que se ha llamado una epistemología extendida. Ya en susodicho artículo, Clark y Chalmers apuntaban al debate sobre la distribución del crédito epistémico. Representando esta linea de investigación, **Keith Harris** (Ruhr-University Bochum) examinó recientes propuestas sobre el conocimiento extendido que combinan un fiabilismo de virtudes con la tesis de la cognición extendida. La idea básicamente es que las habilidades que producen conocimiento pueden extenderse a instrumentos u otros artefactos mediante el proceso de integración cognitiva, dando lugar a casos de conocimiento extendido. Harris criticó esas propuestas e iluminó un fenómeno mediante el cual el conocimiento puede ser el resultado de un sistema cognitivo que incluye un agente humano y distintas tecnologías, sin ser, sin embargo, propiamente atribuible al agente humano individual.

5. Críticas al dogma de la harmonía en la mente extendida

En los últimos años, una de las principales críticas que se ha realizado en el debate en torno a la mente extendida, es que éste se ha desarrollado de tal forma que parece asumir que las relaciones entre humanos y tecnologías son siempre armoniosas y por ello sitúan al agente humano en una mejor posición cognitiva. Esto es lo que recientemente se ha llamado "el dogma de la harmonía" (Aaagard, 2021). En esta línea, en mi ponencia (**Gloria Andrada**) presenté un trabajo co-autorado junto **Richard Menary** (Macquarie University), quien destaca por su influyente trabajo en la literatura de la mente extendida. En él, proponemos un tipo de injusticia que llamamos "injusticia cognitiva" y que sucede cuando la enculturizacion de las habilidades cognitivas frena o transforma negativamente el desarrollo cognitivo de un agente. Por su parte, **David Spurret** (University of Kwazulu-Natal) se centró en la analogía entre el argumento de Dawkins sobre el fenotipo extendido y la cognición extendida. En particular, Spurret defendió que esta analogía, bien realizada, permite iluminar aspectos de la cognición extendida no suficientemente explorados y que tienen que ver con las vulnerabilidades que la extensión cognitiva trae consigo al explotar y manipular el entorno. También, basándose en el trabajo de Kim Sterelny, Spurret introdujo la noción de hostilidad para examinar esta versión menos optimista de la mente extendida.

6. Temas generales de cognición situada e incorporada

Por último, hubo también un par de presentaciones sobre cuestiones no tanto ligadas directamente a la mente extendida sino a aspectos más generales de la cognición situada e incorporada. En esta línea, **Lawrence Shapiro** (University of Wisconsin-Madison) revisó críticamente ciertas conclusiones en torno a la penetración cognitiva y a la representación, que se derivan, supuestamente, de investigaciones sobre la ilusión de la mano de goma. Por su parte, **Albert Newen** (Ruhr-University Bochum) presentó un modelo sobre el yo basado en la cognición situada. Su charla proponía una manera de comprender el yo que integra la pluralidad de dimensiones de la subjetividad junto con la unidad de la experiencia del yo. La idea principal de modelo del Yo Situado (*Situated Self*) que Newen propuso, es que éste es un yo encarnado (*embodied*), que engloba un patrón integrado de aspectos característicos, y que puede incluir partes o entidades más allá del propio cuerpo.

Como puede observarse, TXM @ 25 ofreció un examen exhaustivo de distintos aspectos de la mente extendida. Incluyó un análisis crítico tanto de algunos problemas ya esbozados en el artículo original, como otros más novedosos relativos a tecnologías contemporáneas, a perspectivas más políticas, así como temas más generales de cognición situada. Los debates transcurrieron amigablemente, y puedo decir que fue un evento que realmente celebró esta idea filosófica para la cual la mente no es algo interno sino que se extiende al mundo que habitamos y creamos colectivamente. TXM @ 25 contó también con diversos eventos sociales, entre los que cabe destacar una excursión por las montañas rocosas donde los participantes pudimos disfrutar de un increíble paraje natural entre estimulantes conversaciones.

Para concluir esta reseña, quiero señalar que este evento fue parte del coloquio anual Morris Colloquium on Philosophy de la Universidad de Colorado, Boulder, y recibió apoyo financiero del Fondo Morris y del Comité de Historia y Filosofía y del Instituto de Ciencias Cognitivas de UC, Boulder, así como de la Fundación Volkswagen.

Referencias

Aagaard, Jesper (2021). 4E cognition and the dogma of harmony. *Philosophical Psychology* 34 (2):165-181.

Clark, Andy & Chalmers, David J. (1998). The extended mind. *Analysis* 58 (1):7-19.

Gloria Andrada
Universidade NOVA de Lisboa
gloriandrada@gmail.com

UNIVERSIDADE NOVA DE LISBOA

SLMFCE
Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

CRÓNICAS DE EVENTOS

Conferencia de la Asociación Suiza para los Estudios de Ciencia, Tecnología y Sociedad (STS-CH): Ciencia, experticia y otros modos de conocer.

(31 de agosto y 1 de septiembre de 2023)

El 31 de agosto y 1 de septiembre de 2023 tuvo lugar, en la Universidad de Basilea, la conferencia de la Asociación Suiza para los Estudios de Ciencia, Tecnología y Sociedad (STS-CH), que en esta edición llevó por título "Ciencia, experticia y otros modos de conocer".

En ella tuve la oportunidad de exponer una comunicación sobre el tema central, la experticia científica, pero enfocada en el caso de los examinadores de patentes, gracias a las generosas ayudas para la asistencia a congresos que ofrece esta Sociedad. Junto a un servidor asistieron) alrededor de 100 personas (aunque carezco de la cifra oficial), perfectamente organizadas en varias sesiones paralelas, ubicadas todas ellas en el hermoso *kollegienhaus* de la Universidad.

Pese a la brevedad del evento, la reunión resultó ser muy fructífera, ya que reunió a muchos investigadores en estudios históricos y sociales de la ciencia, especialmente europeos, pero también africanos (ya que la Universidad de Basilea cuenta con un prestigioso Instituto de Estudios Africanos). Una brillante clase magistral del profesor Bruce Lewenstein sobre ciencia ciudadana inauguró la conferencia, y a ella se sumaron luego otras dos no menos interesantes sesiones plenarias, a cargo de la profesora Sally Wyatt y del profesor Fredrick Ogenga, que trató precisamente sobre la digitalización como estrategia panafricana.

Las comunicaciones fueron diversas, y no solo circunscritas al tema de la conferencia. Acerca de la experticia hubo varias charlas: sobre la responsabilidad de los expertos, la confianza epistémica, la experticia distribuida, etc. No obstante, si se observa el programa con detenimiento se acordará que, a pesar de ser éste el tema principal, no terminó de sobreponerse a otros temas clásicos en estudios sociales de la ciencia, especialmente vinculados con la medicina y la salud, que tuvieron incluso una presencia mayor (por ejemplo, la construcción social de la enfermedad o los sesgos de género en salud). También relacionados con salud y medicina hubo ponencias sobre temas que me parecieron del todo novedosos, centradas, por ejemplo, en punteras técnicas de medicina personalizada, o en la recolección y protección de los datos médicos, y en los problemas que todo ello suscita para la filosofía y la sociología. En cualquier caso, no fue difícil percatarse de la importancia que tiene este campo amplísimo de los estudios sociales de la medicina (importancia que, me parece, va en paralelo con el crecimiento que experimenta también la filosofía de la medicina).

Aunque la conferencia fue breve, hubo espacio también para el encuentro social, tanto informal como oficioso, desde las pausas del café, bastante copiosas, hasta las reuniones y cenas. Resulta, por tanto, del todo oportuno recomendar de cara a futuras ediciones la asistencia a los interesados en los estudios CTS, la filosofía y la historia de la ciencia, porque a buen seguro encontrarán en la Asociación helvética un buen foro para discutir sus investigaciones.

Benedicto Acosta Díaz
Investigador predoctoral FPU
Universidad de Salamanca
bneacosta@usal.es

VNiVERSiDAD D SALAMANCA

SLMFCE
Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

CRÓNICAS DE EVENTOS

International Conference on Invention and Imagination (Lisboa, 14-16 de diciembre de 2023)

La Universidad Nova de Lisboa ha acogido los días 14, 15 y 16 de diciembre de 2023 el congreso internacional *Invention and Imagination*. El objetivo fundamental de dicho evento ha sido la reflexión acerca de la imaginación y la invención en los distintos ámbitos de la actividad humana, como son la ciencia o la técnica, pero también el arte. Durante los días que ha durado el congreso, comunicadores y conferenciantes de diversos campos de estudio han contribuido al debate sobre estos dos ingredientes tan fundamentales para la creatividad y la innovación. El programa del congreso se ha organizado en torno a seis conferencias plenarias y varias sesiones de comunicaciones paralelas tanto presenciales como online.

La indagación acerca del rol de la imaginación para la creación en sus múltiples y distintas formas ha sido el núcleo de gran parte de las contribuciones. Ejemplo de lo anterior fue la conferencia de Vincent Bontems, filósofo de la ciencia e investigador del *French Alternative Energies and Atomic Energy Commission*. En su propuesta, titulada "De l'invention à l'innovation: quel rôle pour l'imagination? Gilbert Simondon et les méthodologies de conception", reflexionó sobre el concepto de imagen en la filosofía de la técnica del famoso pensador francés. En contraste, el filósofo Richard Kearney quiso ofrecer una revisión de su clásico trabajo de 1995 "Narrative Imagination: Between Ethics and Poetics", donde el pensamiento de Paul Ricoeur, otro gran intelectual francés, ocupó un lugar central.

Dentro de este eje temático, un asunto relevante ha sido el análisis del papel en la invención y generación de conocimiento de los distintos recursos imaginativos, como las metáforas. En este sentido, el psiquiatra y antropólogo António Bracinha -Vieira, dedicó a este tema su conferencia titulada "De l'imagination à la découverte. Le rôle de la métaphore". En ella, el profesor portugués evalúa la importancia de la imaginación para la adquisición de conocimiento. Desde su punto de vista, las metáforas son el ingrediente imaginativo central para tales propósitos.

Por otro lado, algunas de las sesiones plenarias han girado sobre otros elementos creativos que van más allá de la facultad individual de la imaginación. Desde esta óptica, la profesora Yulia Ustinova, experta en historia de la Antigua Grecia, examinó la noción de inspiración que manejaban los poetas y filósofos griegos. Estos consideraban el proceso creativo como un don divino, como una alteración de la conciencia que emanaba directamente de los dioses.

Por su parte, la filósofa de la ciencia Olga Pombo, se propuso realizar "una cartografía del concepto de creación", donde explorar las diferentes hipótesis explicativas alrededor de dicho fenómeno y su transversalidad disciplinar.

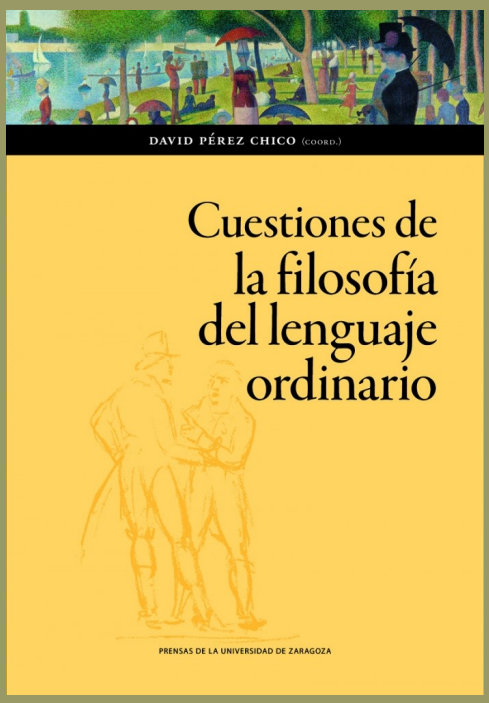
Durante las conferencias plenarias y las casi 60 comunicaciones se pudo disfrutar de debates apasionados y fructíferos. La pluralidad de enfoques fue el signo distintivo del congreso, así como la diversidad de disciplinas de las que provenían cada uno de los participantes. Todo ello facilitó un clima de tolerancia intelectual y lingüística, que benefició el intercambio de ideas.

Por último, es menester destacar el excelente y cuidadoso trabajo realizado por el comité organizador: Adelino Cardoso, Nuno Fonseca, Paulo Jesus, Teresa Lousa y Nuno Miguel Proença.

Daniel Labrador Montero
Universidad de Salamanca
daniabra@usal.es

SLMFCE
Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

RESEÑAS



PÉREZ CHICO, D. (coord.): *Cuestiones de la filosofía del lenguaje ordinario*. Prensas de la Universidad de Zaragoza, 2023, 328 páginas.

Disponible en: Cuestiones de la filosofía del lenguaje ordinario (unizar.es)

El presente texto[1] [2] es una reseña del libro *Cuestiones de la filosofía del lenguaje ordinario* (2023), coordinado por David Pérez Chico y editado por Prensas de la Universidad de Zaragoza. Se trata de una obra colectiva, donde cada capítulo está escrito por una persona experta en filosofía del lenguaje en general y en filosofía del lenguaje ordinario en particular. La obra está dividida en tres bloques. El primer bloque, "El periodo clásico" (pp. 33-168), abarca los primeros cinco capítulos del libro y está dedicado a la filosofía del lenguaje ordinario más tradicional y ortodoxa. El segundo bloque, "En torno a Cavell" (pp. 169-226), conformado por los capítulos seis, siete y ocho, se centra exclusivamente en la obra de l filósofo Stanley Cavell y de la influencia recibida de filósofos como Ralph Waldo Emerson y Henry Thoreau. Los capítulos nueve, diez y once componen el tercer y último bloque, "Desafíos contemporáneos" (227-326). Este es, quizás, el más variado y plural de los tres bloques, pues en él se analizan temas y cuestiones contemporáneas de la filosofía del lenguaje ordinario tan variadas como el lenguaje inclusivo o la relación de la fenomenología y la lógica informal con la filosofía del lenguaje ordinario.

Los dos primeros capítulos están dedicados exclusivamente al estudio de la filosofía del lenguaje de Paul Grice. En el primero, titulado "Grice: del análisis del lenguaje a la Filosofía Primera" y escrito por Juan José Acero, se analiza la peculiar relación de Paul Grice con la filosofía del lenguaje ordinario. Grice formó parte de la escuela filosófica de Oxford de John Austin, centrada en el análisis filosófico del lenguaje ordinario. Sin embargo, poco a poco fue alejándose de dicha filosofía del lenguaje, sin llegar a abandonarla del todo. Esto fue debido a dos grandes movimientos intelectuales en su carrera: el primero, su aproximación al naturalismo de Willard V. O. Quine y Noam Chomsky; el segundo su acercamiento a fuentes más clásicas que le llevaron a una preocupación genuina por el propio concepto de racionalidad. De ahí que Acero considere que la relación de Grice con la filosofía del lenguaje ordinario es "una [relación] ambivalente" (p. 36).

El segundo capítulo se titula "El principio de la primera palabra y la diversidad lingüística" y está escrito por Antonio Blanco Salgueiro. En él se nos explica como Grice, en su famoso artículo "Meaning" (1957), acepta fielmente el principio de la primera palabra de Austin, según el cual el lenguaje común, ordinario, es el punto de partida (la primera palabra) de la reflexión filosófica. Aceptando este principio, Blanco Salgueiro analiza el fenómeno del determinismo y la diversidad lingüística, atendiendo como ejemplo a las sutilezas de la traducción castellana de "Meaning". Así, Blanco Salgueiro concluye que, si bien la primera palabra está en nuestro lenguaje común, puede que no la última, por lo que puede que "la filosofía encuentre su camino en muchas lenguas y [se renuncie] al presupuesto de que el inglés o cualquier otra lengua hegemónica sea filosóficamente neutral" (p. 92).

Como hemos visto, Grice representa un papel peculiar y ambiguo dentro del canon de la filosofía del lenguaje ordinario, al que pertenecen otros autores como Wittgenstein o Austin. El tercer capítulo, titulado "La (otra) filosofía del lenguaje cotidiano: las filósofas silenciadas" y escrito por Juan José Colomina Almiñana, cuestiona que el canon de autores que suele mantenerse sea correcto (p. 98), visibilizando el papel fundamental que tres autoras han tenido en la formación y desarrollo de la filosofía del lenguaje ordinario, a saber: Lizzie Susan Stebbing, Alice Loman Ambrose y Margaret MacDonald. Según Colomina, estas tres filósofas ofrecieron novedosas teorías del significado en términos realistas que,

1. Este trabajo pertenece al proyecto de investigación "Looking at the world with new eyes. Perspectives, frames, and perspectivism" de la Universidad de La Laguna (PID2022-142120NB-100), financiado por el Ministerio de Ciencia e Innovación del Gobierno de España.

2. Trabajo cofinanciado por la Agencia Canaria de Investigación, Innovación y Sociedad de la Información de la Consejería de Universidades, Ciencia e Innovación y Cultura y por el Fondo Social Europeo Plus (FSE+) Programa Operativo Integrado de Canarias 2021-2027, Eje 3 Tema Prioritario 74 (85%).

RESEÑAS

sin embargo, fueron silenciadas, siendo un claro caso de injusticia epistémica y testimonial. Este capítulo constituye, según su autor, el primer trabajo en lengua castellana en el cual se reúne y reivindica la importancia de Stebbing, Ambrose y MacDonald en la filosofía del lenguaje ordinario.

El cuarto capítulo, escrito por Cristina Corredor, se titula "Dar razones, ser una razón, lo razonable: Austin, Grice y Toulmin". En los capítulos primero y segundo se explica cómo la relación de Grice con la filosofía de Austin es ambigua. En este se va más allá y se ofrece una reinterpretación de la teoría de la argumentación de Stephen Toulmin a la luz de la filosofía del lenguaje ordinario de dichos autores. Para ello, Corredor analiza, por un lado, la ausencia de validez lógica en los argumentos sustantivos de Toulmin y, por el otro, la relación entre el calificador modal "probable-mente" (analizado por Toulmin) y el realizativo "prometo que" (analizado por Austin). Así, Corredor propone que "si se toma en consideración el acto de habla de argumentar, tanto en su estructura como en su dimensión interaccional, se hace posible integrar de manera coherente estas tres importantes contribuciones" (p. 122).

Cerrando el primer bloque encontramos el quinto capítulo, titulado "Wittgenstein: el colmo de la filosofía" y cuyo autor es Santiago Garmendia. En él se trata un tipo de discurso muy particular, el cómico y humorístico, centrándose en el fenómeno del chiste. A través del análisis del uso del sustantivo "colmo" en diversos chistes, los cuales siguen una estrategia de reducción al absurdo, Garmendia llega a la tesis de que "la mejor manera de entender la filosofía wittgensteniana del lenguaje (…) es a través del fenómeno del chiste" (p. 156). Y es que el chiste gramatical, según el autor, es una de las múltiples herramientas de la filosofía terapéutica de Wittgenstein, pues a través del chiste es posible establecer "los límites de lo decible (es decir, la gramática de nuestro lenguaje) a través de lo absurdo" (p. 163).

El segundo bloque, centrado en la filosofía de Cavell, se abre con el sexto capítulo, escrito por Sandra Laugier y titulado "La voz como forma de vida y como forma de la vida". En él, la autora mantiene que lo que la filosofía del lenguaje cavelliana busca es la devolver su merecida importancia al concepto de voz humana y como éste (junto con la noción de "forma de vida") es capaz de redefinir la subjetividad . Así pues, frente al clásico problema de la expresión, en la filosofía del lenguaje ordinario cobrará importancia la relación entre el sujeto humano y la voz. En esta relación, según Laugier, es donde encontramos las implicaciones más políticas de la filosofía de Cavell, pues la voz implica, en última instancia, la representación del sujeto en la comunidad de hablantes (pp. 171-172).

Comprender una instancia lingüística pasa por comprender un significado. ¿Podría, sin embargo, llegar a comprenderse algo sin que necesariamente haya significado? La autoexpresión del lenguaje y la comprensión sin significado es el tema del séptimo capítulo, titulado "El lenguaje se expresa a sí mismo" y escrito por Gordon C. F. Bearn. En él, además de Wittgenstein y Cavell, encontramos al músico vanguardista John Cage. Famoso por su obra *4'33"* (obra musical cuya partitura no tiene ninguna nota y, por lo tanto, constituye cuatro minutos y treinta y tres segundos de silencio) y por el uso aleatorio del *I Ching* como herramienta compositiva, Cage es usado en este capítulo para ejemplificar las propuestas de Wittgenstein y Cavell acerca de un lenguaje que se expresa más allá de las normas gramaticales, las convenciones y los conformismos.

El octavo capítulo, escrito por Victor J. Krebs y titulado "La cuestión del re-casamiento. Cavell, la filosofía y la alabanza", sirve como cierre del segundo bloque. En él, se explora una preocupación que recorre la obra entera de Cavell, a saber: la separación radical entre la razón y la pasión y la preeminencia de la primera frente a la segunda. En este caso, el desamor, la negación y la falta de alabanza juegan un papel crucial. El autor, siguiendo a Cavell, propondrá un reencuentro o re-casamiento de la pasión y la razón, unión de dos mitades que interesa no solo a la filosofía del lenguaje ordinario, sino a la filosofía analítica en general.

El noveno capítulo, "Fenomenología y filosofía del lenguaje ordinario: el *hard problem* de la conciencia y el *hard problem* de la inteligencia artificial", escrito por Manuel Liz, da inicio al tercer bloque del libro, dedicado a cuestiones contemporáneas para la filosofía del lenguaje. En este capítulo, el autor vuelve a los inicios de la filosofía del lenguaje ordinario y de la fenomenología y examina la historia de ambas corrientes, buscando sus puntos de convergencia y divergencia en torno a la cuestión de lo conceptual y lo no-conceptual. Según el autor, el trato de ambas disciplinas a la relación entre lo conceptual y lo no-conceptual nos sirve para pensar problemas muy actuales dentro de la filosofía de la mente y las ciencias cognitivas: el *hard problem* de la conciencia (la existencia de la conciencia en un mundo puramente físico) y el *hard problem* de la inteligencia artificial (la posibilidad de reconocer como persona a una máquina).

El décimo capítulo, titulado "Pragmalingüística de la argumentación: de la filosofía del lenguaje ordinario a la lógica informal" y escrito por Javier Vilanova, se establece una relación entre la filosofía del lenguaje ordinario y la lógica informal, una relación tan clara y directa como la que une a la filosofía analítica de corte más clásico con la lógica formal. Pero pese a que la lógica que acompaña a la filosofía del lenguaje ordinario sea una lógica informal, esto no significa que ni la una ni la otra sean irracionales en absoluto. Esta es una de las principales conclusiones a las que lleva Vilanova en su capítulo mediante la desconexión entre la filosofía del lenguaje ordinario y varios de los conceptos centrales de la lógica formal. La lógica informal está estrechamente vinculada a la teoría de la argumentación y es esencialmente pragmática. Por ello, se la debe entenderse, como propone el título, como una pragmalingüística de la argumentación (p. 291).

RESEÑAS

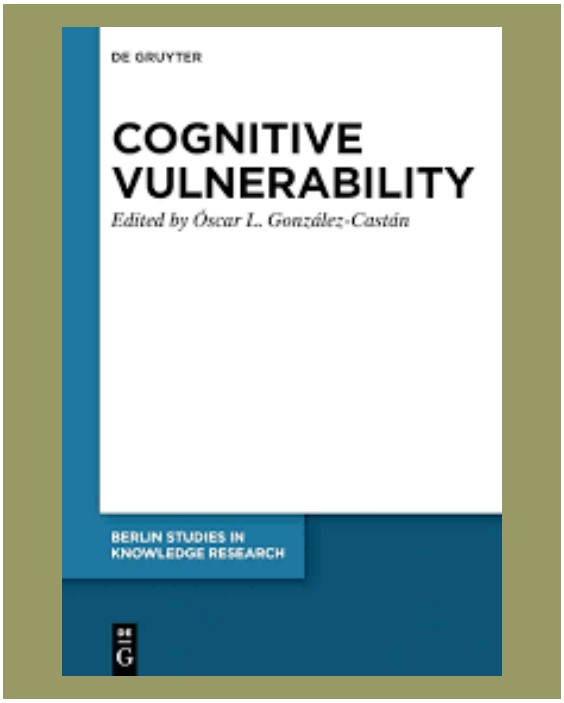
El undécimo y último capítulo, titulado "*Las, les los*. Una aproximación wittgensteniano-(brandomiano) hegeliana al lenguaje inclusivo en el contexto español" y escrito por Carla Carmona, trata el tema tan importante y necesario, dentro y fuera de la filosofía, del lenguaje inclusivo. La reflexión acerca de dicho tema parte de la actividad docente que Carmona lleva a cabo cada día en su aula, viviendo cada una de sus clases "como una continua lucha contra los límites del lenguaje en lo que respecta a la diversidad sexual y de género, concretamente, contra los (supuestos) límites de la lengua española" (p. 293). Si con el lenguaje hacemos cosas, es decir, realizamos acciones, entonces con un uso inclusivo del lenguaje podemos reconocer la diversidad sexual y de género y, al contrario, con un uso no inclusivo podemos mal-reconocer y dejar de reconocer. El uso deliberado de la x (chixs), la arroba (chic@s) o la e (chiques) es, pues, un uso rebelde del lenguaje (p. 321), cuyo efecto cognitivo es el de explicitar los límites del lenguaje y, con ellos, los tradicionales usos sexistas y no inclusivos del lenguaje. Y respecto a este tema la filosofía del lenguaje ordinario ofrece un muy buen marco teórico.

Hemos repasado cada uno de los once capítulos, pero no hemos dicho nada todavía acerca de la introducción. En ella, David Pérez Chico, a raíz de la división entre filosofía del lenguaje ideal y filosofía del lenguaje ordinario, plantea una interesante discusión acerca del concepto de "ordinario" y de cómo el uso de esta palabra resalta la idea de que el uso filosófico del lenguaje es realmente extraordinario (p. 11). Además, plantea la identificación (o no) de la filosofía del lenguaje ordinario con la pragmática lingüística.

En el propio libro encontramos tesis enfrentadas acerca de la definición de "ordinario" y de la extensión de la filosofía del lenguaje ordinario. Esto podría parecer un perjuicio para el libro, pues podríamos pensar que pierde cohesión teórica en su conjunto. Pero tras una lectura de este, vemos que no solo goza de una gran cohesión, coherencia y robustez, sino que la pluralidad de autores y autoras constituye una virtud. Al dar voz a varios filósofos y filósofas del lenguaje, podemos ver de primera mano una discusión actual en el seno de la filosofía del lenguaje ordinario, un intercambio genuino de ideas que nos muestra que en filosofía, como en varios aspectos de la vida, siempre hay una pluralidad de puntos de vista y que varios de ellos pueden ser perfectamente correctos simultáneamente. Por ello, este texto puede resultar muy útil tanto para el alumnado de filosofía que tenga ciertas nociones de filosofía del lenguaje como para aquellas investigadoras e investigadores preocupados (y ocupados) por la filosofía del lenguaje ordinario y sus desarrollos más actuales.

Enrico Brugnami
Universidad de La Laguna
ebrugnam@ull.edu.es

RESEÑAS



GÓNZALEZ-CASTAN, Óscar L. (ed.). *Cognitive Vulnerability. An Epistemological Approach.* Berlin/ Boston: De Gruyter. 2023, 222 páginas.

Disponible en: Cognitive Vulnerability (degruyter.com)

El libro *Cognitive Vulnerability. An Epistemological Approach*, editado por Óscar González-Castán, constituye una aplicación de la noción de vulnerabilidad a los ámbitos de la epistemología y de la filosofía de la ciencia. Tomando como referencia la vulnerabilidad humana, en la presente obra se desarrolla el concepto de vulnerabilidad cognitiva para dar respuesta a problemas clásicos, tales como el debate realismo/ antirrealismo, el éxito o la confianza epistémicos. Además, se exploran las implicaciones políticas y sociales de esta noción, insertándola en problemas contemporáneos como el reto de la posverdad o los desacuerdos profundos.

El libro se articula en dos partes, compuestas de cinco aportaciones cada una. La primera parte se ocupa de desarrollar teóricamente la noción de vulnerabilidad cognitiva y de explorar sus consecuencias en la epistemología y la filosofía de la ciencia. En la segunda parte encontramos contribuciones sobre las implicaciones prácticas de esta conceptualización, tanto su relevancia para comprender problemáticas de carácter social y político, como posibles "antídotos" que disminuyan los efectos perniciosos de la vulnerabilidad cognitiva.

La vulnerabilidad cognitiva designa una ampliación de la clásica noción de falibilismo propuesta por Peirce. Esta última se refiere a la posibilidad de que nuestras creencias (o lo que ahora consideramos conocimiento) puedan ser falsas, así como que nuestras facultades cognitivas y justificaciones de nuestras creencias puedan conducir a falsedades. La vulnerabilidad cognitiva incluye el falibilismo, aunque añade un aspecto positivo: la posibilidad del éxito epistémico, de decir cosas verdaderas acerca del mundo. Esta caracterización de la noción que nos ocupa es desarrollada por González-Castán en la primera contribución a este libro, insertándola en el debate realismo-antirrealismo.

Según González-Castán, ambas posiciones presuponen la noción de falibilismo: porque nuestras creencias son potencialmente falsas, el error nos permite mejorar nuestras teorías para acercarnos más a la verdad (realismo) o, a la inversa, la persistente posibilidad del error nos impide acercarnos a una verdad absoluta (anti-realismo). Dado que, en ambas posturas, el falibilismo es un elemento central, González-Castán introduce en su lugar la noción de vulnerabilidad cognitiva con el fin de disolver la dicotomía. Haciendo énfasis en la posibilidad del éxito cognitivo, y no solo en el carácter falibilista, el error se entiende como productivo: nos permite mejorar nuestras teorías y nuestras metodologías, sin que esto suponga comprometerse con una posición realista. En este punto, González-Castán introduce la noción de verosimilitud cognitiva como aquello en lo que consiste el progreso científico, en lugar de postular una verdad absoluta acerca de la realidad externa.

Las siguientes contribuciones desarrollan y perfeccionan la noción de vulnerabilidad cognitiva. En primer lugar, José María Ariso se ocupa del conocimiento y la certeza negativos, como formas productivas de hacer progresar el conocimiento. Para ello toma como referencia las ideas de Wittgenstein en *Sobre la Certeza*, así como la distinción propuesta por Ortega y Gasset entre ideas y creencias. En ambos filósofos encontramos la diferenciación entre un conocimiento con contenido proposicional, que puede ser puesto en duda y tomado como verdadero o refutado como falso, y certezas (o creencias en Ortega) que estructuran nuestra imagen del mundo. Esta distinción permite a Ariso justificar el papel del error como conocimiento negativo y de la certeza negativa, entendida esta última como la exclusión de la posibilidad de cometer un error, en el progreso de la ciencia, implementando así la idea de que la falibilidad va ligada al éxito epistémico.

El artículo de Javier Vilanova también ahonda en la tesis de que la posibilidad del éxito cognitivo y el falibilismo son dos caras de una misma moneda. El autor se inspira en la filosofía del lenguaje ordinario de Austin y Wittgenstein para plantear una concepción del conocimiento plural y anti-fundacionalista. Encontramos un punto de controversia interesante entre su conceptualización del conocimiento y las que aparecen implícitas en los otros artículos que, según aduce el propio autor, estarían demasiado cerca de concepciones tradicionales del conocimiento y la verdad. Vilanova sostiene que el verbo "conocer"

RESEÑAS

tiene muchos usos diferentes, según el contexto, adoptando un acercamiento pragma-lingüístico a la noción de conocimiento que incluye los intereses personales y de grupo, el rechazo de la visión del ojo de Dios (que va ligado a la capacidad mejorar de nuestras herramientas epistémicas), y la defensa de la flexibilidad de reglas epistémicas y, en consecuencia, del carácter pluralista del conocimiento.

La aportación de Timothy Williamson permite reforzar las ideas mencionadas hasta ahora, dirigiendo la discusión a los heurísticos, es decir, reglas prácticas que aplicamos para resolver rápidamente problemas específicos. Williamson centra la atención en dos casos particulares, el heurístico de persistencia y el heurístico suposicional para condicionales. *Grosso modo,* Williamson sostiene que, aunque estas herramientas no preserven siempre la verdad y puedan, por tanto, conducir al error, su utilidad y fiabilidad en la mayoría de los casos justifican su uso. Extrayendo una conclusión general, aunque nuestras herramientas y procedimientos cognitivos puedan llevar a errores, en contextos normales el énfasis se desplaza a la posibilidad del éxito epistémico.

La contribución de Rosa Maria Calcaterra cierra el primer bloque: la autora explora, en su artículo, la noción de vulnerabilidad cognitiva en conexión con la teoría de Peirce. Argumenta que esta nos aporta herramientas para disminuir los aspectos negativos de la vulnerabilidad cognitiva. El artículo de Calcaterra toma de Peirce la noción de falibilismo, una caracterización del realismo que no cae en una concepción fundacionalista, y la centralidad de la cooperación social. Estos elementos, asegura Calcaterra, nos permiten neutralizar los aspectos perniciosos de la vulnerabilidad epistémica. Su artículo constituye, por tanto, un puente con la segunda parte de este libro, donde la atención se centra en los efectos de dicha vulnerabilidad y en las formas de enfrentarlos.

La segunda parte comienza con el artículo de Ángeles J. Perona, que aborda problemáticas específicas de carácter social y político aplicando la categoría de vulnerabilidad cognitiva. Perona sostiene, en línea con las otras contribuciones de esta edición, que la racionalidad humana es plural y dependiente de una red de creencias y prácticas que se constituyen en una comunidad. Propone así sustituir una concepción fundacionalista de la racionalidad por lo que ella denomina "racionalidad blanda", ligada a nuestra vulnerabilidad cognitiva. La discusión se orienta a fenómenos concretos, como la desconfianza acerca de los resultados de la ciencia, la injusticia testimonial o los desacuerdos profundos, entendidos todos ellos como manifestaciones de vulnerabilidad cognitiva.

La contribución de Günter Abel está fuertemente conectada con el capítulo anterior. Si abandonamos un criterio de racionalidad universal por una concepción de la racionalidad como situada, ¿de dónde deriva la dimensión normativa de nuestras prácticas? ¿Podría dicha normatividad mitigar nuestra vulnerabilidad cognitiva? Abel propone una concepción de la normatividad cimentada en los ajustes mutuos entre la primera persona, el resto de los miembros de la comunidad, y la sociedad y el mundo entendidos ampliamente. La vulnerabilidad consiste en que ocurran desequilibrios entre las normas y las prácticas particulares. Esta concepción tridimensional de la normatividad se complementa con el Principio de Equilibrio Reflexivo, tomado de Nelson Goodman (en el plano epistémico) y John Rawls (en el político). De acuerdo con Abel, el principio establece una búsqueda constante de un equilibrio entre nuestros juicios y las normas que guían su formación: ante una perturbación, se impone la demanda de restablecer el equilibrio mediante cambios en nuestras normas, en nuestros juicios o en ambos.

El artículo de Astrid Wagner complementa los dos anteriores. La autora, que acepta una pluralidad de estándares de racionalidad en lugar de una concepción universal de la misma, sostiene que en la era de la posverdad dichos estándares no se aceptan. Tampoco lo son los métodos de justificación ni los criterios de objetividad, incurriendo en los mecanismos de desconfianza que caracterizan los flujos de información (*fake news,* desinformación, negacionismo, teorías de la conspiración). Wagner también acepta el Principio de Equilibrio Reflexivo pero, a diferencia de Abel, lo fundamenta en estados o virtudes epistémicas (la ausencia de certeza, la confianza y la responsabilidad): un desequilibrio entre dos de ellos requiere de compensación por parte del tercero. Lo contrario, sostiene Wagner, conduce a la polarización. La autora concluye defendiendo la virtud de la "humildad epistémica" y la necesidad de transmitir una imagen de la ciencia donde la falibilidad no se oponga a la fiabilidad.

Las dos últimas contribuciones toman orientaciones algo distintas. En primer lugar, Adam Carter reflexiona sobre el papel de la suerte y del riesgo como factores que hacen nuestras prácticas epistémicas vulnerables. El autor parte de una distinción entre entender-por qué *p* y saber-que *p,* para analizar dos tipos de riesgo epistémico (interviniente y ambiental). La idea central es que entender-por qué *p* requiere más esfuerzo que saber-que *p* y, por lo tanto, la vulnerabilidad cognitiva es mayor en el primer caso. No obstante, a partir de una revisión de la literatura relevante, Carter llega a la tesis incompatible de que somos menos o igual de vulnerables al entender-por qué *p* que al saber-que *p.* Para resolver la aporía, Carter entra en una discusión acerca de las ambigüedades que conducen a esta.

El último artículo de esta obra tiene por autor a Modesto Gómez-Alonso, y constituye una defensa de la epistemología de "goznes" propuesta por Wittgenstein. El autor lleva a cabo dicha tarea analizando tres problemas filosóficos distintos que, según aduce, han sido abordados de forma defectuosa: el tratamiento empirista de las propiedades disposicionales, la distinción entre ocurrencias y acciones intencionales, y la pregunta acerca de cómo adquirimos conocimiento certero

RESEÑAS

acerca del mundo externo. Gómez-Alonso responde a estas cuestiones rechazando una concepción monista y reduccionista del conocimiento. Para ello, recurre a la epistemología de goznes de Wittgenstein, una perspectiva pluralista que diferencia entre reglas y movimientos empíricos dentro de un juego del lenguaje. Gómez-Alonso sostiene que esta concepción pluralista consigue no caer en el escepticismo sin ignorar la vulnerabilidad epistémica, ya que las proposiciones-gozne en las que se articula la propuesta wittgensteiniana, aunque excluyan la posibilidad de duda, siempre aparecen ligadas a unas circunstancias y están sujetas a cambio.

Recapitulando, la noción de vulnerabilidad cognitiva que se desarrolla en esta obra es de interés por múltiples razones. En primer lugar, esta nos permite dar respuesta a problemas clásicos de la epistemología y de la filosofía de la ciencia, insertándolos en una concepción de la racionalidad que, rechazando caracterizaciones universalistas de la adquisición del conocimiento, es eminentemente humana. Pero, en segundo lugar, la aplicación de la vulnerabilidad al plano cognitivo no es interesante solo por su relevancia en temas clásicos de la filosofía, sino por su aplicación a problemáticas actuales como los desacuerdos profundos o la desconfianza en la ciencia. La capacidad de dar una explicación unificada al abanico de cuestiones de carácter social o político que se tratan en esta obra, lejos de ser una implicación incidental de esta noción, constituye una de las razones fuertes para aceptar el valor de esta categoría para comprender la dimensión epistémica del ser humano.

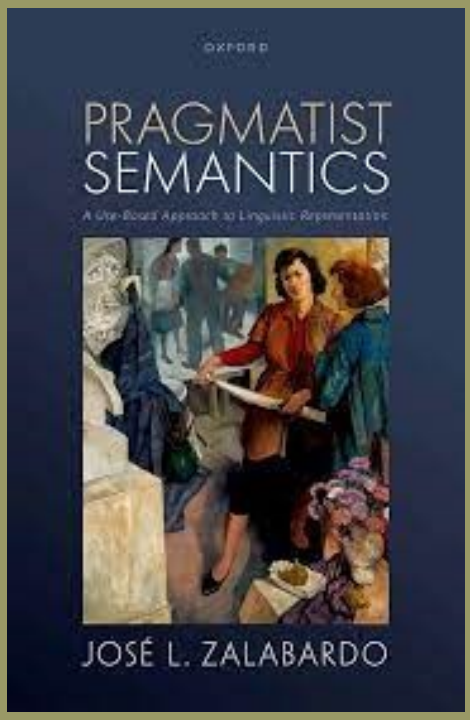
Un aspecto que considero que podría haber merecido más interés es la posibilidad de que ciertos grupos, en concreto aquellos afectados por prejuicios identitarios, sean más vulnerables cognitivamente que otros. Si bien la categoría de injusticia testimonial, tal y como la trata Miranda Fricker, se menciona como relacionada con la vulnerabilidad cognitiva, considero que se podría haber desarrollado más de qué forma esta se manifiesta de manera distinta en las partes implicadas. ¿Son igualmente vulnerables quienes desconfían de un testimonio por prejuicios identitarios que aquellas personas que sufren tal injusticia? ¿Qué ocurre con la injusticia hermenéutica, donde un grupo oprimido carece de recursos cognitivos para comprender su experiencia? ¿Es también un tipo de vulnerabilidad epistémica? Pienso que estas preguntas constituyen un punto de partida estimulante para futuros trabajos.

Otro aspecto que se podría haber abordado con más detalle es un tema de controversia que se menciona explícitamente en el libro. Por una parte, Gónzalez-Castán defiende en su aportación que los mecanismos por los que disminuimos los efectos perniciosos de la vulnerabilidad cognitiva nos permiten progresar hacia grados crecientes de verosimilitud cognitiva. No obstante, en otros lugares se sostiene una tesis incompatible: por ejemplo, Ariso argumenta que las creencias establecidas por mecanismos que integren nuestro conocimiento negativo no están mejor apoyadas que las creencias antiguas (y el progreso de la investigación podría igualmente "destronarlas"). Sería beneficioso poder analizar argumentos a favor y en contra de cada una de estas tesis, ya que la opción que adoptemos afecta a la comprensión de la noción de vulnerabilidad cognitiva.

Para concluir, y retornando a lo dicho al principio, esta obra plantea cuestiones de gran interés y relevancia, tanto por su conexión con cuestiones filosóficas clásicas como por su capacidad de integrar temáticas actuales de carácter político. Las contribuciones incluidas constituyen un éxito a la hora de proponer una categoría que permita dar explicación de la diversidad de cuestiones abordadas, ofreciendo un terreno fértil para motivar investigaciones posteriores.

Ana Rosa López Rodríguez
Universidad de Granada
analopezrdgz@correo.ugr.es

RESEÑAS



ZALABARDO, José L. Pragmatist Semantics: a use-based approach to linguistic representation, 2023. Oxford: Oxford University Press. 256 páginas.

Disponible en: Semántica pragmática - José L. Zalabardo - Oxford University Press (oup.com)

Tenemos que representarnos el mundo[1]. Tenemos que actuar en él. Estas son las dos inquietudes que, sin apurar demasiado la historiografía, nos pueden servir para orientarnos en la historia de la filosofía; más aún si atendemos a la ambigüedad de este "tener que": ¿refiere a una necesidad o es acaso un deber? Sea cual sea el significado que le demos, y sea cual sea la respuesta que adoptemos frente a este universo de inquietudes, estas permanecerán como fondo de esa peculiar historia. La obra que aquí reseñamos, *Pragmatist Semantics*, da cuenta de esa primera inquietud y la matiza: no se ocupa de la posibilidad de representar el mundo en general, sino de poder representarlo mediante el uso de nuestro lenguaje. Aquí, la mención al "uso" no es accidental: la teoría que en ella se defiende es de corte pragmatista y, precisamente por ser pragmatista, el hecho de que se ocupe de esta inquietud representacionalista resulta muy original.

El autor de *Pragmatist Semantics* es José Luis Zalabardo, que es Professor of Philosophy en el University College London (UCL) y que ha centrado sus esfuerzos, además de a la filosofía wittgensteniana, a la oposición entre realismo y antirrealismo. A este tema ha dedicado multitud de artículos y libros, entre los cuales podemos destacar "Realism Detranscendentalized", publicado en la *European Journal of Philosophy* (2000) y *Concepciones de lo real: realismo y antirrealismo en semántica y metafísica* (2002). De esta preocupación surge su posterior interés en torno a las representaciones de la realidad, ya sean estas lingüísticas o mentales y, por tanto, de esa temprana preocupación surge su interés en la metasemántica, disciplina a la que se adscribe esta obra.

Para lograr el análisis de esa capacidad representativa, reducida en este caso a una dimensión fundamentalmente lingüística, Zalabardo compara dos estrategias dirigidas a la clarificación o especificación de los fundamentos del significado de la oración declarativa. La primera, el representacionalismo, sostiene que el significado de una oración declarativa se debe a una determinada relación semántica de la oración con el estado de cosas que representa. Aquí entra en juego la suposición representacionalista (suposición RR), a saber, que si una oración representa exitosamente el mundo se debe a que su fundamento semántico es representacionalista. La segunda estrategia, el pragmatismo, sostiene, por el contrario, que el significado no se debe a vínculo semántico alguno sino que se puede (y debe) reducir al uso. Pero *Pragmatist Semantics* no contrasta meramente estas dos estrategias, sino que se decanta —hasta el título lo delata— por la segunda: el representacionalismo, tal y como nos dice Zalabardo, hará las veces de villano y el pragmatismo, de héroe (p. 8). Así pues, lo que pretende *Pragmatist Semantics* es exponer la gratuidad o inadecuación de la suposición representacionalista.

Para determinar claramente lo que es el representacionalismo (y así poder someterlo posteriormente a una crítica), Zalabardo dedica el primer capítulo, titulado justamente "The Representational Discourse", a esbozar el perfil del representacionalismo que más plausiblemente puede dar cuenta del problema de los fundamentos del significado. Este representacionalismo 'canónico' puede reducirse a la tesis que sostiene que el significado de nuestros discursos enunciativos sólo puede explicarse si los predicados centrales de ese discurso tienen referente (p. 9). El hecho de que la referencia se perciba como algo obligatorio en la explicación representacionalista le permite a Zalabardo volver a los célebres argumentos que ya esgrimió Bertrand Russell en su discusión con Alexius Meinong: porque, aunque el representacionalismo resulta razonablemente eficaz para explicar los enunciados verdaderos, fracasa cuando se trata de explicar los enunciados falsos o que contienen términos sin referencia.

1. Trabajo cofinanciado por la Agencia Canaria de Investigación, Innovación y Sociedad de la Información de la Consejería de Universidades, Ciencia e Innovación y Cultura y por el Fondo Social Europeo Plus (FSE+) Programa Operativo Integrado de Canarias 2021-2027, Eje 3 Tema Prioritario 74 (85%); y se enmarca en el proyecto de investigación "Looking at the world with new eyes. Perspectives, frames, and perspectivism" de la Universidad de La Laguna (PID2022-142120NB-100), financiado por el Ministerio de Ciencia e Innovación del Gobierno de España.

NOTICIAS DE LIBROS

No obstante, es en los capítulos segundo y tercero, respectivamente titulados "The Open-Question Argument in Ethics" y "The Open-Question Argument in Semantics", en los que la ausencia de referencia (de los predicados morales y semánticos respectivamente) resulta más problemática e impactante. En ambos capítulos se analiza e interpreta el argumento de la pregunta abierta originalmente propuesto en los *Principia Ethica* de George Moore. El primero de estos dos capítulos se mueve en un ámbito mucho más cercano al contexto original del argumento. *Pragmatist Semantics* no es, ni pretende ser, una incursión en metaética; pero en este segundo capítulo Zalabardo explica el sentido y la relevancia de ese argumento para las ciencias morales y somete a revisión las críticas a las que se han visto expuestas. Así, tras atender a la crítica naturalista de Graham Harman, Zalabardo concluye en un intuicionismo moral.

Peculiarmente, Zalabardo extiende el dominio del argumento de la pregunta abierta más allá de la ética, otorgándole carta de naturaleza en semántica. En este tercer capítulo, Zalabardo esgrime que algunas de las nociones centrales de la semántica (a saber: la adscripción de verdad, la de significado y la de creencia) carecen por entero de referente o que, como mínimo, no funcionan dentro de un programa representacionalista. En este punto, dada la naturaleza de estas nociones, esta aproximación pragmatista ya evidencia las graves implicaciones que tiene para la familia de teorías representacionalistas de lo mental y para el mentalismo en general; desarrollando, sin caer en deflacionismos, algunos de los más célebres argumentos que Ludwig Wittgenstein esbozó en sus *Investigaciones Filosóficas*.

Tras explorar en el capítulo cuarto, "Some remarks", algunas otras objeciones relevantes para el programa representacionalista, Zalabardo expone ya su noción pragmatista en "Pragmatist Meaning Grounds", que es el quinto capítulo. En él se rechaza plenamente el supuesto RR y se expresa que son las condiciones de aceptación o rechazo (por parte de un auditorio) las que fijan y aseguran una función representativa también para un lenguaje pensado en clave pragmatista. Además, para perfilar su propuesta más claramente, este capítulo termina contrastando su propuesta con la vía no-cognitivista y con el amplio pragmatismo de Robert Brandom, entre otros.

En los capítulos sexto, "Belief and Desire", y séptimo, "Meaning and Truth", Zalabardo aplica su programa pragmatista a discursos semánticos muy específicos. El sexto, modificando ligeramente la '*intentional stance*' de Daniel Dennett, desarrolla el pragmatismo atendiendo al caso de los discursos que adscriben deseos y creencias. Ahora bien, sin atender a lo en sí de la creencia y del deseo, lo que interesa a Zalabardo es la manifestación de ambas; esto es, el comportamiento que debe reflejar una adscripción u otra (en caso de que la adscripción sea correcta o verdadera). El séptimo capítulo procura dar cuenta y razón de las adscripciones de verdad y significado desde el pragmatismo. Para explicar las atribuciones de significado, Zalabardo avanza desde las nociones quineano-davidsonianas de interpretación o traducción radical y propone sustituir (o matizar) el principio de caridad aplicando, en su lugar, los principios de familiaridad y proyección. Para las atribuciones de verdad, sin embargo, postula un criterio de aceptabilidad (y en este punto su alejamiento del representacionalismo es máximo): el fundamento del significado del predicado "es verdad" es la propia posibilidad de aceptar como verdadero lo que se dice que "es verdad".

El capítulo octavo, titulado "Harmony and Abstraction", es, probablemente, el más importante. En él se desarrolla el núcleo de su propuesta pragmatista afinando la posibilidad de que un uso del lenguaje logre representar un estado de cosas determinado sin necesidad de un fundamento representacionalista. Para que una oración tenga significado, su conexión con el estado de cosas que representa debe ser una condición necesaria, pero, a la vez, debería valer como condición suficiente su fundamento semántico no representacionalista. Esta incompatibilidad entre ambas condiciones es lo que el autor denomina "el problema de la armonía". Para defender su pragmatismo, Zalabardo sostiene que a la referencia de los predicados representacionales no se accede como a un determinado estado de cosas existente sino mediante un conjunto de principios de abstracción basados en condiciones de sinonimia.

Llegamos, finalmente, a "The Primacy of Practise", el noveno capítulo. En él, Zalabardo llega a la conclusión, basándose en la noción de David Lewis de humildad, de que si extendiésemos el representacionalismo hasta la explicación de los términos científicos, careceríamos por entero de acceso cognitivo a sus referencias. Por ello, se apura a exponer el modo en que, desde su propuesta de los principios de abstracción, se podría llegar a las referencias del discurso científico. Para clausurar ya su propuesta pragmatista, el Epílogo aborda una última cuestión de gran importancia, a saber: la fundamentación de los fundamentos. El libro nos lleva, como conclusión, a aceptar que los fundamentos semánticos de los discursos que representan el mundo son fundamentos pragmatistas y no representacionalistas. Sin embargo, a la hora de fundamentar los propios fundamentos, se corre el riesgo de hacerlo sobre fundamentos representacionalistas. Para evitar este problema Zalabardo propone adjudicar fundamentos pragmatistas a los propios fundamentos pragmatistas: son los procedimientos utilizados para regular la aceptación de interpretaciones los que deben regular la aceptación de los propios fundamentos pragmatistas.

Puede que por su temática y por su carácter académico *Pragmatist Semantics* sea un texto poco accesible. Ciertamente es un texto cuyo auditorio está reservado a especialistas en filosofía del lenguaje: su carácter técnico lo dirige a aquellos que ya están familiarizados con las discusiones que lo integran. No obstante, las graves implicaciones que sugiere para el ámbito de la metaética y para el de las teorías de la verdad posibilitan un acceso ligeramente diferente al texto. Ahora bien, sea cual sea el ámbito del que se proceda, tal como se destacó

NOTICIAS DE LIBROS

al principio de esta reseña, es probable que resulte necesaria una cierta inquietud común; porque aquellos que no sientan la necesidad (o el deber) de poder representar lingüísticamente el mundo, de poder dar cuenta de él, de expresarlo, es posible que sientan estas discusiones como algo muy lejano.

Si además de esta inquietud o perplejidad por la posibilidad de la representación, el lector manifiesta una cierta inclinación pragmatista, *Pragmatist Semantics* se convierte en una referencia potencialmente ineludible puesto que encauza con éxito —tal es, al menos, nuestra valoración— esa inclinación convirtiéndola en una teoría filosófica de amplio alcance. Si, por el contrario, carecemos de tal inclinación pragmatista, resulta

Pablo Vera Vega
Universidad de La Laguna
pveraveg@ull.edu.es

Universidad de La Laguna

Enrico Brugnami
Universidad de La Laguna
ebrugnam@ull.edu.es

Universidad de La Laguna

www.solofici.org

SLMFCE
Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España

SOCIEDAD DE LÓGICA, METODOLOGÍA Y FILOSOFÍA DE LA CIENCIA EN ESPAÑA

Para envíos al boletín:
davidpch@unizar.es

Febrero de 2024

www.solofici.org

www.solofici.org