# La singularidad tecnológica y el desafío posthumano

**Antonio Diéguez**

## ¿QUÉ ES EL TRANSHUMANISMO?

El transhumanismo es una filosofía de moda, la utopía del momento. Algunos la consideran incluso la cosmovisión propia de la época postmoderna, dominada por la técnica; el único gran relato posible tras el descrédito en el que han caído todos los demás. Internet y las redes sociales bullen con comentarios sobre el asunto. Los medios de comunicación sacan con periodicidad noticias sobre científicos o ingenieros que hablan de la inmortalidad, de la superinteligencia o del volcado de la mente en un ordenador; y nos describen con entusiasmo las inmensas posibilidades vitales que las nuevas tecnologías nos abrirán, entre ellas, exploraciones y colonizaciones de lugares lejanos del universo, o experiencias mentales y sensoriales que somos incapaces siquiera de imaginar.

Es ciertamente un producto que, una vez vencida cierta reticencia inicial, se vende bien entre el público. De una forma más rotunda y convincente que en ocasiones anteriores, el discurso transhumanista nos dice que la ciencia puede ya poner en nuestras manos lo que hasta ahora parecía el producto de la imaginación desbordada de los artistas. El alargamiento indefinido de la vida, la victoria final sobre la muerte, la promesa definitiva de la inmortalidad es toda la justificación que el transhumanismo necesita para afianzarse. Es lo que han prometido siempre, de una forma u otra, las grandes religiones.

Uno de los primeros escritos que dio base a este movimiento fue el artículo del filósofo Max More titulado «Transhumanismo: hacia una filosofía futurista», que se publicó en 1990. En dicho artículo se definía al transhumanismo de la siguiente manera:

> El transhumanismo es un conjunto de filosofías que buscan guiarnos hacia una condición poshumana. El transhumanismo comparte muchos elementos con el humanismo, incluyendo un respeto por la razón y la ciencia, un compromiso con el progreso y una apreciación de la existencia humana (o transhumana) en esta vida en lugar de en alguna «vida» sobrenatural después de la muerte. El transhumanismo difiere, en cambio, del humanismo al reconocer y anticipar las alteraciones radicales en la naturaleza y en las posibilidades vitales que resulta-

> rán del desarrollo de diversas ciencias y tecnologías, como la neurociencia y la farmacología, las investigaciones sobre la extensión de la vida, la nanotecnología, la ultrainteligencia artificial, la exploración del espacio, combinado todo ello con una filosofía y un sistema de valores racionales. (More 1990, p. 6).

El transhumanismo es, por lo tanto, el intento de transformar radicalmente nuestra especie mediante la tecnología. Y esto puede hacerse de varias maneras, no necesariamente excluyentes. Podemos buscar la integración del ser humano con la máquina creando cíborgs o alojando directamente nuestra mente en las máquinas. Pero podemos también intentar mejorar nuestras capacidades biológicas mediante medicamentos o, finalmente, modificar nuestros genes en la línea germinal, de modo que realizando los cambios necesarios, al cabo de un tiempo tengamos una especie nueva, distinta a la nuestra pero mucho mejor, una especie poshumana. En todo caso, hemos de abandonar la pasividad a la que nos hemos visto sometidos en el proceso evolutivo darwiniano que nos ha hecho como somos. Para los transhumanistas, ha llegado la hora de que el ser humano tome el control de su propia evolución.

En otro lugar me he ocupado de analizar algunos de los problemas implicados en el trasnhumanismo biologicista –al que suele designarse como biomejoramiento humano (cf. Diéguez 2016). Quiero centrarme aquí en las tesis del transhumanismo que ve en la creación de una superinteligencia artificial y en la posterior integración con ella el modo de dar el salto a esa nueva fase evolutiva poshumana, y en particular expondré y analizaré las tesis del ingeniero, inventor, empresario y activista cultural Raymond Kurzweil acerca de la singularidad.

## LA SINGULARIDAD SIEMPRE ESTÁ CERCA

La singularidad es algo tan extrañamente singular que, pese a la popularidad que ha alcanzado el término, apareciendo en películas como «Ex machina», o dando nombre a una universidad, nadie sabe con seguridad lo que es, entre otras razones porque tiene sentidos muy diferentes dependiendo de quién lo use. Lo que sí parece saber todo el mundo es que está cerca, como señala el título del libro de su máximo defensor y difusor, Raymond Kurzweil.[1] Este autor lo usa para designar el advenimiento, en algún momento futuro, del primer sistema superinteligente capaz de perfeccionarse a sí mismo, o capaz de fabricar otros sistemas más inteligentes que él, los cuales a su vez puedan hacer lo mismo, y así sucesivamente en un crecimiento exponencial de la inteligencia alcanzada en cada fase, que terminará por hacer de todo el universo una entidad global inteligente. Kurzweil sostiene que en el año 2029 una máquina pasará el test de Turing, es

decir, mostrará una inteligencia igual a la humana, pero su estimación acerca del advenimiento de la singularidad sitúa a ésta más bien en torno al año 2045. En otras palabras, Kurzweil cree que en el plazo de quince años, a contar desde 2030, las máquinas se perfeccionarán tanto a sí mismas, que todo quedará bajo su control, incluyendo los recursos materiales y energéticos necesarios para mantener su crecimiento, e iniciarán su expansión cósmica. La civilización humana habrá llegado entonces a su fin y comenzará una civilización «post-biológica» bajo el dominio de las máquinas.

El término fue usado ya en 1993, en un ensayo titulado «El advenimiento de la singularidad tecnológica», del escritor y matemático Vernor Vinge, para referirse a esa hipotética explosión futura de la inteligencia de las máquinas. Vinge afirma literalmente que significaría «el final de la era humana». Y unos años antes, en 1965, Irving J. Good, uno de los primeros en hacer notar la posibilidad de una «explosión» de la Inteligencia Artificial, había dejado escrito: «la primera máquina ultrainteligente será la última invención que el ser humano necesite hacer, siempre y cuando la máquina sea lo suficientemente dócil para que nos diga cómo mantenerla bajo control». Pero esto último no pasa de ser un deseo piadoso.

Sorprendentemente, Kurzweil no comparte los escenarios pesimistas en los que el triunfo de las máquinas superinteligentes será el fin de nuestra existencia. Todo lo contrario, él cree que será el comienzo de una nueva era de ilimitados horizontes para nuestro desarrollo:

> La singularidad –escribe– nos permitirá trascender [las] limitaciones de nuestros cerebros y cuerpos biológicos. Aumentaremos el control sobre nuestros destinos, nuestra mortalidad estará en nuestras propias manos, podremos vivir tanto como queramos (que es un poco diferente a decir que viviremos para siempre), comprenderemos enteramente el pensamiento humano y expandiremos y aumentaremos enormemente su alcance. Como consecuencia, al final de este siglo la parte no biológica de nuestra inteligencia será billones de billones de veces más poderosa que la débil inteligencia humana producto de la biología. […]
>
> La singularidad constituirá la culminación de la fusión entre nuestra existencia y pensamiento biológico con nuestra tecnología, dando lugar a un mundo que seguirá siendo humano pero que trascenderá nuestras raíces biológicas. En la post-singularidad, no habrá distinción entre humano y máquina o entre realidad física y virtual. (Kurzweil 2012, pp. 9-10).

La propuesta de Kurzweil para evitar nuestra desaparición o nuestra anulación a expensas de las máquinas superinteligentes es, pues, la integración con la máquina, y en particular el volcado o copia de nuestra mente en una de ellas. El futuro será de las máquinas superinteligentes, pero éstas tendrán una inteligencia cuyo origen será humano, lo que lleva a Kurzweil a la discutible conclusión de que «las máquinas futuras serán humanas, aunque no sean

biológicas» y la civilización que creen será por ello una civilización humana (Kurzweil 2012, p. 33). Habría, pues, que distinguir aquí entre dos propuestas diferentes, aunque no incompatibles: la singularidad como resultado de la creación de una superinteligencia artificial (AI) y la singularidad como resultado de la potenciación o aumento de la inteligencia humana (IA) hasta niveles poshumanos debido en buena medida a la integración con la máquina.

## LA «LEY» DE MOORE

Una de las bases argumentativas en las que Kurzweil cimienta su seguridad en que la singularidad está cerca, y que además será un proceso rápido, es la conocida ley de Moore, que lleva el nombre de quien la formuló en 1965, el ingeniero que más tarde sería cofundador de Intel, Gordon E. Moore. Generalizando sobre ella Kurzweil establece su Ley de los Rendimientos Acelerados (*Law of Accelerating Returns*). La «ley» de Moore establece que el número de transistores que pueden colocarse en un microprocesador (o en resumidas cuentas, el poder computacional de los ordenadores) se duplica en periodos que van de dieciocho meses a dos años, es decir, tiene un crecimiento exponencial. Aunque esa es su formulación inicial, lo cierto es que, tras pasar décadas desapercibida (su notoriedad comienza en los 90), ha sido posteriormente extrapolada –con fundamentos más que discutibles– a diversas tecnologías. Esta tesis, sin embargo, está lejos de poder sustentar por sí sola las pretensiones de Kurzweil, como lo pone de evidencia el hecho mismo de que el propio Gordon Moore ha declarado que en su opinión la singularidad no se producirá nunca.

Aun si se siguiera cumpliendo en los próximos años, la ley en realidad expresa una regularidad contingente, y no hay base científica para suponer que pudiera expresar alguna regularidad más fuerte, al modo de las que recogen las leyes científicas genuinas. De hecho, Moore no la presenta como una ley, sino como una aspiración a mantener en la industria de la microelectrónica a lo largo de los años posteriores. Se ha hecho popular entre los informáticos el chascarrillo de que el número de investigadores que predicen la muerte de la ley de Moore se duplica cada dos años. El nombre de «ley» que lleva no le corresponde legítimamente. Nada, en efecto, ninguna causa física o mecanismo subyacente, obliga a que dicha regularidad se siga cumpliendo. Y obviamente existen límites físicos para el crecimiento postulado. No se conoce ningún crecimiento exponencial que no alcance tarde o temprano su fase de meseta, y muchos terminan decayendo. Esto es algo que Kurzweil admite en el caso concreto de la ley de Moore, pero que no considera aplicable igualmente a su generalización de la ley que él propone, la de «los Rendimientos Acelerados».

Para salvar la idea de que el crecimiento exponencial continuará más allá del punto en el que se alcance la fase de meseta en crecimientos exponen-

ciales anteriores, Kurzweil recurre a la idea kuhniana de los cambios de paradigma, solo que él la aplica a los paradigmas tecnológicos, en particular a cambios en los paradigmas de las tecnologías de computación. Cuando una de estas tecnologías se estanca, un cambio permite tomar el relevo a un nuevo paradigma tecnológico y éste actúa como nuevo motor del crecimiento.

Pero incluso este crecimiento exponencial reimpulsado periódicamente por cambios de paradigma tecnológicos habría de cesar en algún momento. Lo que habría que determinar entonces es si ese estancamiento se produciría antes o después de que la superinteligencia generada hasta entonces tuviera un verdadero impacto sobre la vida de los seres humanos, y en particular si esa superinteligencia artificial estaría en condiciones de tomar el control sobre dichas vidas. Pero no está claro que esa fase de meseta en la curva logística que estabiliza lo alcanzado durante la fase de crecimiento haya de producirse en un nivel en el que las máquinas superinteligentes sean tan sofisticadas como para adquirir la voluntad de sustituir al ser humano en el control de nuestro planeta. Kurzweil, sin embargo, lo da por sentado. Afirma con total convencimiento –pero sin más argumento que su confianza en la Ley de Rendimientos Acelerados– que «este límite [al crecimiento de las tendencias exponenciales de las tecnologías de la información] no se alcanzará antes de que las grandes transformaciones descritas en este libro [es decir, las acarreadas por la singularidad] se hayan producido» (Kurzweil 2012, p. 499).

En realidad, lo que está sucediendo es que se acumulan las evidencias de que la validez de la ley de Moore ya no podrá ser sostenida por mucho tiempo. En febrero de 2016 la revista *Nature* publicaba un artículo que comenzaba con la siguiente afirmación: «El mes próximo la industria mundial de semiconductores reconocerá formalmente lo que se ha venido haciendo cada vez más obvio para toda persona implicada: la ley de Moore, el principio que ha potenciado la revolución de las tecnologías de la información desde la década de los 60, está cercana a su fin» (Waldrop 2016, p. 145). El motivo principal de esa ralentización, según explicaba el artículo, era el económico, pero no el único. Quizás como anticipo de ese reconocimiento, a comienzos de febrero de 2016, William Holt, directivo de Intel, declaraba que su empresa iba a comenzar a sacrificar el crecimiento de la velocidad de los microprocesadores –el crecimiento según la ley de Moore– por una mayor eficiencia energética de los mismos. Así que los chips que comenzarán a fabricar no serán mucho más veloces pero, a cambio, consumirán menos batería. De hecho, desde el año 2004, para evitar sobrecalentamientos de los aparatos, la industria ha estancado la velocidad de ejecución de instrucciones de los microprocesadores. El fin de la ley de Moore no significa, sin embargo, el fin de los progresos en capacidad de procesamiento. Hay nuevos tipos de ordenadores y de procesadores en el horizonte que nos traerán a buen seguro grandes sorpresas. Aunque es posible que el objetivo próximo de las empre-

sas que los fabrican no sea tanto el aumento cuantitativo de potencia, como la diversificación y adaptación a las necesidades complejas de los usuarios.

Y si la ley de Moore comienza a ser seriamente cuestionada en su propio terreno –el de la potencia de computación–, las razones para mantenerla fuera de él son aún menos convincentes. Ningún desarrollo tecnológico tiene por qué seguir necesariamente en los próximos decenios un crecimiento exponencial. De hecho hay quien piensa que vivimos en una época en que, comparativamente con otras anteriores, el crecimiento en la innovación se está frenando. Pensemos en los avances tecnológicos que contemplaron los años que cubren la segunda mitad del siglo XIX y el modo en que cambiaron de forma radical la vida cotidiana de una buena parte de la humanidad, o en las grandes innovaciones en ciencias de la computación, en biotecnología, en electrónica, en el uso de la energía nuclear, en ingeniería aeroespacial, que se produjeron entre 1945 y el comienzo de la década de los 70. Si comparamos esto con lo logrado en las dos o tres últimas décadas que hemos vivido, no parece que salgan muy favorecidas, y eso incluso contando en su haber el comienzo de la difusión de Internet. Pero no hace falta pensar que estamos en plena decadencia de inventiva tecnológica para ser escépticos ante la idea de un crecimiento exponencial prolongado en el desarrollo tecnológico. Basta con considerar que las tecnologías son muy diversas, y mientras unas avanzan rápidamente otras pueden estancarse o ralentizarse después de haber crecido, y que hay que contar también con los condicionantes no sólo materiales, sino muy especialmente culturales, que ese desarrollo puede encontrar.

Los singularistas no se arredran, sin embargo, ante estas objeciones. Sus esperanzas están puestas en un cambio radical de tecnología –un cambio de paradigma tecnológico de los que habla Kurzweil– que significará un verdadero salto cualitativo con respecto a todo lo que puedan hacer los ordenadores actuales. Y la propuesta que reclama más atención es la de la computación cuántica. Ésta se basa en el aprovechamiento de que los sistemas cuánticos no se limitan a presentar solo dos estados posibles, digamos 1 y 0, que son los que constituyen un bit de información en la computación al uso, sino que pueden presentar también una superposición de dichos estados. Dicho de otro modo, un bit de información en un ordenador cuántico, lo que se denomina qubit, puede estar en los estados 1 o 0, como un ordenador tradicional, pero puede estar simultáneamente en una superposición de ambos estados a la vez. Esto, junto con el fenómeno del entrelazamiento cuántico, que permite que los estados de dos sistemas estén correlacionados en sus valores incluso aunque ya no estén en contacto físico, faculta al ordenador cuántico para realizar una gigantesca cantidad de operaciones simultáneamente. Tantas, que superaría con mucho todo lo que puede hacer un ordenador convencional. La primera prueba realizada con éxito daba una velocidad de cómputo cien millones de veces mayor que la de cualquier procesador actual.

No obstante, lo destacable aquí es que se necesita que estos qubits funcionen conjuntamente en la citada superposición de estados antes de que colapsen debido a lo que se denomina «decoherencia cuántica», y que consiste en que cualquier mínima perturbación exterior hace que el sistema tome solo uno de los valores posibles, es decir, 0 o 1, perdiéndose de este modo el entrelazamiento cuántico. Por el momento sólo se ha conseguido que funcionen conjuntamente doce qubits en una situación de extremo aislamiento de partículas subatómicas, y el tiempo que dura la superposición de estados y el entrelazamiento es de décimas de microsegundos, pero los investigadores en este campo esperan grandes progresos al respecto en los próximos años.

Todo ello implicará, si se cumplen estas expectativas, un aumento gigantesco en la potencia de cálculo de los ordenadores. Por el momento, sin embargo, solo puede afirmarse que la computación cuántica está en sus inicios y no es posible saber cuál será su futura viabilidad o la velocidad a la que se realizarán los progresos.

## ¿CONSTITUYE TODO ESTO UN PROGRESO HACIA UNA SUPERINTELIGENCIA?

Ésta es una cuestión fundamental que es necesario plantearse. Aun dejando de lado el espinoso problema de la consciencia, o las muy difundidas objeciones a la posibilidad de creación de una inteligencia artificial en sentido estricto planteadas por los filósofos Hubert Dreyfus y John Searle o por el físico Roger Penrose, sobre las que no podemos entrar aquí, el problema de fondo sigue siendo si un aumento en la potencia de cálculo de las máquinas, por grande que sea, es suficiente para generar inteligencia en el sentido en el que afirman los singularistas. Una cosa es que podamos aumentar (incluso enormemente) esta potencia de cálculo y la velocidad de procesamiento, y otra es que eso vaya a dar lugar a algo así como un cerebro mecánico con una inteligencia comparable o superior a la humana. Por añadidura, no está nada claro que todos los problemas a los que debe enfrentarse un ente inteligente sean computables –más bien hay buenas razones para pensar que no es así– y, por tanto, que una máquina inteligente pueda resolverlos operando sólo con algoritmos. Hasta el momento, la Inteligencia Artificial ha proporcionado logros más que notables, como los sistemas expertos, o el sistema Watson para el análisis de datos no estructurados en orden a localizar patrones en los mismos, desarrollado por IBM y ganador del concurso televisivo *Jeopardy!*, o como el programa AlphaGo, que ha ganado a campeones humanos del juego Go, o como los algoritmos de búsqueda de Google. Sin embargo, no ha conseguido (ni se prevé que lo vaya a conseguir en un plazo determinado) la creación de máquinas con inteligencia general y versátil, sensible a los cambios de contexto y aplicable a campos diversos. Sus resultados han estado limitados

a ámbitos muy específicos de aplicación. Un sistema experto, que es uno de sus mejores productos, es mucho más inteligente que un ser humano para el desarrollo de una cierta tarea, al igual que lo es una simple y modesta calculadora, pero no lo es en nada más. Fuera de su ámbito de aplicación es perfectamente inútil.

Uno de los grandes avances de la IA actual, lo que se conoce como «aprendizaje profundo», que implica una mejora enorme en el aprendizaje en redes neuronales artificiales y que es, de hecho, el procedimiento que emplea el sistema Watson, está basado en algo tan antiguo como el teorema de Bayes, en la mayor potencia de cálculo de los ordenadores actuales, y en el acceso a datos masivos (*big data*). No hay ningún gran salto teórico cualitativo detrás. En definitiva, no existe por el momento un sistema de inteligencia artificial que tenga una capacidad general, aplicable a cualquier tipo de problemas, como es el caso de la inteligencia humana. Los ingenieros y científicos que se dedican a la IA suelen adoptar una posición de prudente espera. Son muchos los que parecen pensar que tendremos una máquina con inteligencia comparable a la humana antes de que acabe el siglo, pero en todo caso, cuando han de manifestarse al respecto en publicaciones académicas, sus afirmaciones suelen estar muy lejos de las visiones radicales de los singularistas. Como muestra, un botón. En 2006 la revista *AI Magazine* publicó un artículo titulado «Inteligencia artificial: los próximos veinticinco años» (Stone y Hirsh 2006). En él, veinticinco investigadores relevantes en el campo de la IA expresan su opinión sobre los avances que se han producido y se van a producir en dicho campo. Ninguno de ellos menciona la creación de una superinteligencia general ni hace referencia alguna a la singularidad o algo parecido. Varios se limitan a mencionar avances técnicos parciales en sus propias áreas de investigación, y echan de menos una mayor integración entre dichas áreas. Y los pocos que mencionan la producción de robots con inteligencia comparable a la de los seres humanos, lo hacen en casi todos los casos para decir que estamos aún muy lejos de conseguir tal cosa.

Si merece la pena o no llevar la investigación conducente a una inteligencia artificial general más allá de cierto punto dependerá en buena medida de que seamos capaces de desarrollar los instrumentos para evitar los escenarios distópicos que Kurzweil y otros nos presentan como inevitables. Porque lo que parece claro es que, si la creación de una superinteligencia autónoma es posible, no es muy probable que esto constituya algo positivo para los seres humanos. Aún en el caso de que no intentara acabar directamente con nosotros, su propio crecimiento y despliegue, o la consecución ineludible de sus objetivos –como en el ejemplo socorrido de la máquina dedicada a fabricar a toda costa el mayor número posible de clips de oficina– podría presentar muchos peligros para nuestra especie. No nos destruiría su malevolencia, sino su indiferencia. Convendría, en tal caso, mantener la implementación de los avances en Inteligencia Artificial limitados al desarrollo de sistemas

capaces de realizar tareas concretas y abandonar para siempre el proyecto de crear una IA general, que además de ser el de más incierta factura, sería el más amenazador para la seguridad de los seres humanos. No es extraño que ya vaya circulando por las redes un manifiesto firmado por muchos grandes científicos e ingenieros en pro de un control más estricto de las investigaciones en IA, de modo que puedan tomarse medidas a tiempo contra los riesgos que encierra su desarrollo.[2]

Las tesis de Kurzweil han despertado reacciones muy encontradas.[3] Por un lado, tiene devotos seguidores que le consideran casi un gurú espiritual (papel que aparentemente asume con gusto, puesto que ha escrito también sobre nutrición y estilos de vida sana) y que van difundiendo sus ideas con un entusiasmo y convicción un tanto ingenuos.[4] No es infrecuente que se acojan bajo el paraguas de la *Singularity University*, creada en 2009 con el patrocinio de Google y de la NASA, y que es una institución que tiene mucho más de eficaz empresa de marketing que de universidad propiamente dicha. Pero tiene también fuertes críticos y opositores, algunos de ellos dentro del campo de la Inteligencia Artificial, como el propio Gordon Moore, o incluso, entre los defensores del transhumanismo, como Donna Haraway. El teórico de la computación y científico cognitivo Douglas Hofstadter no pudo ser más duro en sus palabras en una entrevista para la revista *American Scientist*, calificando el libro de una mezcla de buena comida y excrementos de perro (cf. Ross 2007).

Una defensa más reciente y mejor justificada de la posibilidad de que el ser humano pueda crear en un futuro próximo una superinteligencia artificial es la que ha realizado el filósofo de la Universidad de Oxford Nick Bostrom en un libro titulado precisamente así, *Superintelligence: Paths, Dangers, Strategies* (Bostrom 2014). Bostrom no llama a esto «singularidad» ni le atribuye las características que le atribuye Kurzweil, y reconoce además que las dificultades en este proyecto son mucho mayores de lo que se había pensado con anterioridad. Por ello, entre otras razones, su retrato de la situación es más realista y menos entusiasta.

Particularmente reseñable es el esfuerzo que el libro por tratar la cuestión de qué mecanismos podríamos utilizar para impedir que una superinteligencia fuera hostil a los intereses, fines y valores humanos. El resultado de la lectura deja, sin embargo, la desoladora impresión de que no habría ninguno realmente efectivo, ni siquiera aquellos que Bostrom contempla como más dignos de exploración. Los mecanismos de control serían todos inútiles a menos que esa superinteligencia hubiera internalizado de algún modo que nuestros intereses humanos deben ser respetados y promovidos, pero sus reflexiones acerca de los modos posibles de infundir un comportamiento ético en ella son en la práctica un reconocimiento de la futilidad del intento. Y ciertamente sería mucho esperar que una superinteligencia cuya situación en el mundo, cuyos objetivos, cuya «corporización», cuya experiencia, cuya

historia, cuya (carencia de) socialización, cuyas emociones (si las tuviera) estuvieran tan sumamente alejados de todas las atribuciones que podamos hacer acerca de estos rasgos en el caso de nuestra especie, que terminara pese a todo desarrollando un comportamiento ético similar al nuestro. Bostrom tiene la honradez intelectual de admitir que la creación de una superinteligencia sería un riesgo existencial para el ser humano (algo que no palía en absoluto el hecho de que también podría contribuir a disminuir otros riesgos existenciales) y que éste estaría prácticamente desarmado frente a su poder. Él no toma falsas vías de escape como el volcado de nuestra mente en un ordenador o nuestra unión con las máquinas. Lo que su libro pretende es avisarnos del peligro, no dibujar una tecnoutopía en cuya contemplación podamos solazarnos.

No obstante, sigue habiendo mucho de común entre la posición de Kurzweil y la de Bostrom. Ambos están convencidos de que la superinteligencia artificial llegará en un plazo no muy lejano, que lo hará de forma súbita, y que dominará a los seres humanos si estos no toman medidas radicales (como fundirse con ella, en el caso de Kurzweil, o insertando en ella de algún modo un comportamiento ético sensible a los fines y valores de los seres humanos, en el caso de Bostrom). Los escenarios dibujados por Kurzweil y por Bostrom se basan más en sus visiones futuristas que en los datos reales, por mucho que a ellos les guste adornar sus escritos con todo tipo de datos aparentemente confirmatorios. Es cierto que no hay razones irrefutables para pensar que la creación de una superinteligencia artificial no es ni será jamás posible. Probar eso sería tanto como probar que tal superinteligencia tiene que ser imposible *a priori*, y los argumentos esgrimidos hasta ahora por los críticos no consiguen hacerlo. Pero una cosa es que no se pueda probar que es imposible y otra muy distinta es que esté entonces garantizado que la tendremos. Esa supuesta superinteligencia podría ser posible, pero las dificultades para construirla podrían también ser tan grandes que nunca fuéramos capaces de hacerla. Como muy bien dice el filósofo de la información Luciano Floridi, «una verdadera IA no es lógicamente imposible, pero sí es completamente implausible. No tenemos ni idea de cómo podríamos empezar a construirla» (Floridi 2016).

## AGRADECIMIENTO

El autor manifiesta su agradecimiento a Gonzalo Ramos Jiménez, del Departamento de Lenguajes y Ciencias de la Computación de la UMA, por comentarios útiles que ayudaron a mejorar el texto.

## REFERENCIAS


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

DIÉGUEZ, A. (2016, en prensa): «La biología sintética y el imperativo de mejoramiento», *Isegoría*, 54.

EDEN A. H. *et al*. (eds.) (2012): *Singularity Hypotheses: A Scientific and Philosophical Assessment*, Berlín, Springer.

FLORIDI, L. (2016): «Should we be afraid of AI?», *AEON*, 9 de mayo,< https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible>. Consultado el 10/5/16.

KURZWEIL, R. (2012): *La singularidad está cerca. Cuando los humanos trascendamos la biología*, Berlín, Lola books.

MORE, M. (1990): «Transhumanism: Toward a Futurist Philosophy», *Extropy*, 6, pp. 6-11.

SCARUFFI, P. (2013: *Demystifying Machine Intelligence*, Omniware Publishing, <http://www.scaruffi.com/singular/download.pdf>. Consultado el 28/4/16.

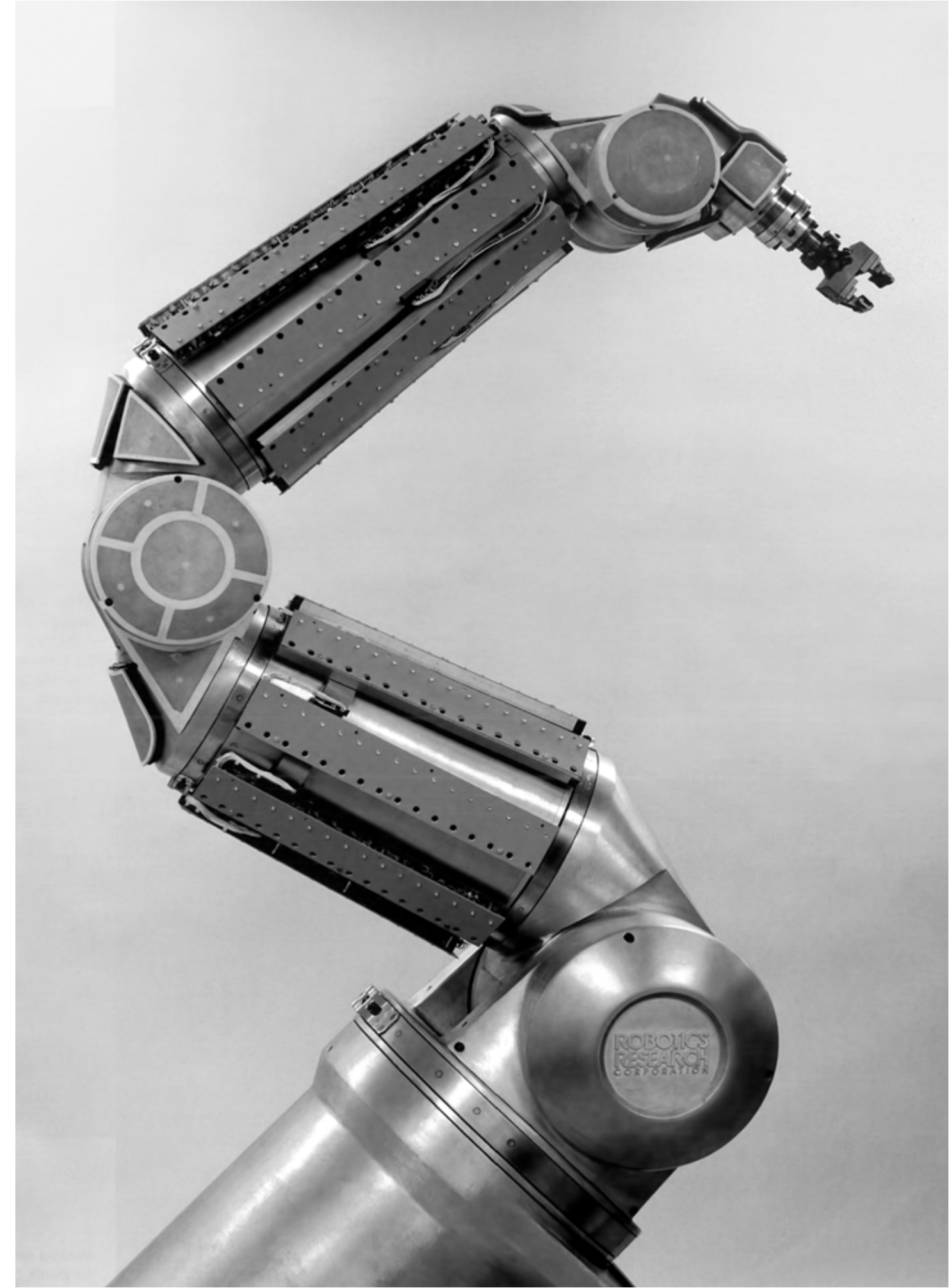
STONE, M. y H. HIRSH (2006): «Artificial Intelligence: The Next Twenty-Five Years», *AI Magazine*, 26 (4), pp. 85-97.

WALDROP, M. (2016): «More than Moore», *Nature*, 530, 11 de febrero, pp. 144-147.


## NOTAS

1. Me refiero a su libro más conocido, titulado *La singularidad está cerca. Cuando los humanos trascendamos la biología* (Berlín, Lola books, 2012). El original en inglés su publicó en 2005.
2. Puede consultarse en la siguiente dirección: <http://futureoflife.org/ai-open-letter/>.
3. Un análisis crítico interesante y bien documentado puede encontrarse en Scaruffi (2013) y un análisis de los pros y contras realizado por autores favorables y otros escépticos puede verse en Eden *et al*. (eds.) (2012).
4. Es muy instructivo al respecto ver la entrevista que Iñaki Gabilondo le hizo a uno de sus discípulos, el ingeniero de origen venezolano José Luis Cordeiro, para el programa «Cuando ya no esté. El mundo dentro de 25 años», retransmitida por el canal #0 el 17 de marzo de 2016.


**ANTONIO DIÉGUEZ es catedrático de Lógica y Filosofía de la Ciencia en la Universidad de Málaga. Ha sido profesor invitado en las universidades de Helsinki y de Harvard. Es presidente electo de la Asociación Iberoamericana de Filosofía de la Biología. Una de sus líneas de investigación principales ha sido el debate sobre el realismo científico. Acerca de esta cuestión publicó el libro *Realismo científico* (Universidad de Málaga, 1998). Ha trabajado asimismo sobre cuestiones de la filosofía de la tecnología, con atención a las tesis del determinismo tecnológico. En los últimos años se dedica especialmente a la filosofía de la biología, indagando sobre cuestiones de epistemología evolucionista. Sobre este tema ha publicado *La evolución del conocimiento. De la mente animal a la mente humana* (Biblioteca Nueva, 2011). Es autor asimismo de *La vida bajo escrutinio. Una introducción a la filosofía de la biología* (Biblioteca Buridán, 2012).**


ROBOTICS
RESEARCH
CORPORATION