

MILENARISMO TECNOLÓGICO: LA COMPETENCIA ENTRE SERES HUMANOS Y ROBOTS INTELIGENTES

ANTONIO DIÉGUEZ
Universidad de Málaga

Precisamente los rasgos más característicos de la condición humana –por ejemplo el miedo a la muerte, la aversión hacia el propio cuerpo, el deseo de moralidad o de liberarse de los errores– unidos a un deseo de dominio evolutivo sobre el mundo natural constituyen las fuerzas fundamentales que, de forma creciente, empujan a los humanos para crear máquinas.

B. Mazlish, *La cuarta discontinuidad*, pp. 308-9.

Resumen. Este artículo se dedica, en primer lugar, a exponer y analizar las opiniones de algunos investigadores en IA (inteligencia artificial) sobre las consecuencias de las máquinas inteligentes (futuro posthumano); en segundo lugar se critican estas posiciones desde el punto de vista de la posibilidad de controlar el desarrollo de la tecnología.

Abstract. Firstly, this article is devoted to expose the ideas of some distinguished scholars in AI (Artificial Intelligence) about the social consequences of intelligent machines (post-human future). Secondly, these positions are criticised from the view-point of the possibility of controlling the development of technology.

La posibilidad de crear máquinas con una inteligencia igual o superior a la humana ha dejado de ser desde hace unos años un tema exclusivo de la ciencia-ficción para convertirse en un asunto bajo el es-

crutinio de la ciencia. El campo de la Inteligencia Artificial (IA) se basa en la aspiración razonada de obtener en un plazo no demasiado lejano tales máquinas inteligentes. O mejor sería decir máquinas con procesos mentales inteligentes, evitando de este modo asumir que la inteligencia sea una propiedad única y homogénea.

Hay ciertamente quienes piensan que esta posibilidad debe ser descartada de antemano, al menos si por inteligencia (o por procesos mentales inteligentes) entendemos algo que no está sometido a reglas pre-determinadas, algo que faculta a los seres humanos para reconocer rápidamente lo relevante y lo accesorio en un entorno cambiante, que les permite ser intuitivos y creativos tanto en el terreno de la teoría como en el del arte, algo que incluye la capacidad para comprender significados (contenidos semánticos) y para usar el lenguaje haciendo referencia con él al mundo real; algo, en fin, que tiene como manifestaciones singulares la consciencia y lo que habitualmente llamamos 'sentido común'. Los críticos de la IA insisten en que éstas son características sin las cuales no cabe hablar de inteligencia, y ponen en cuestión que las máquinas puedan desplegar alguna vez tales características (cf. Dreyfus 1993, Weizenbaum 1984 y Searle 1980; para un análisis, véase Martínez Freire 1995, cap. 8).

Estas voces discrepantes –hay que decirlo– no son tenidas muy en cuenta por los científicos e ingenieros implicados directamente en proyectos de IA, y por el momento sus argumentos no son definitivos contra los que presentan los defensores de las máquinas inteligentes, aunque desde luego no carezcan de plausibilidad inicial en muchos aspectos¹. Es cierto que los avances realizados en el campo de la IA –de los que los sistemas expertos y las redes neuronales artificiales son la mejor muestra– no han sido tan rápidos ni tan espectaculares como todavía se esperaba a principios de los 80, pero han sido lo suficientemente importantes como para que las expectativas creadas en torno a dicho campo se mantengan en alza (cf. Martínez Freire 1996).

¹ Algunas de estas críticas, especialmente la de Dreyfus, han sido recogidas sin embargo en propuestas que se presentan como alternativas a la visión dominante en IA, tanto en su versión simbólica como en su versión conexionista. Estas propuestas intentan superar la concepción de la mente como un centro desencarnado de razonamiento lógico y de computación para considerarla más bien como un sistema encarnado (*embodied*) de control de las actividades de un cuerpo inmerso en un entorno. (Cf. Varela et al. 1991, Brooks 1997, van Gelder 1997 y Clark 1997).

No es mi intención, sin embargo, entrar aquí en el debate sobre la posibilidad real de crear máquinas superinteligentes, ni en el de la diferencia entre simular la inteligencia y tener inteligencia. Para seguir hasta el final la línea de la argumentación que me interesa desarrollar en estas páginas, supondré que las máquinas inteligentes con capacidades superiores a las humanas estarán a nuestro lado en un futuro más o menos lejano, si bien soy consciente de lo problemática y controvertida que es una suposición como ésta. Lo que haré a continuación será exponer en primer lugar algunas de las implicaciones sociales más radicales que destacados investigadores en IA han extraído de esa posibilidad, e intentaré mostrar después que, incluso aceptando tal suposición, las previsiones que efectúan sobre el mundo que se avecina están muy deficientemente fundadas y deben ser tenidas como posibilidades muy remotas, cuando no como meras fantasías milenaristas.

LOS ROBOTS DEL ANOCHECER

Una cuestión que, con toda su crudeza, ha atraído de modo especial a algunos científicos y figuras relevantes de la IA es la de las relaciones que podrían establecerse entre el ser humano y las máquinas en un futuro en el cual éstas fueran superiores en inteligencia. ¿Qué sucederá con el ser humano cuando estas máquinas superinteligentes sean robots o estén integradas en robots capaces de dotarlas de movimiento y puedan construirse a sí mismas y proliferar de forma rápida? Uno de los primeros en buscar una respuesta fue Edward Fredkin, gerente del Laboratorio de Inteligencia Artificial del Massachusetts Institute of Technology. En 1979 ya había forjado algunas conclusiones al respecto que expuso en una entrevista televisiva:

Hay tres grandes acontecimientos en la historia. Uno, la creación del universo. Otro, la aparición de la vida. El tercero, que creo de igual importancia, es la aparición de la inteligencia artificial. Ésta es una forma de vida [sic] muy diferente, y tiene posibilidades de crecimiento intelectual difíciles de imaginar. Estas máquinas evolucionarán: algunos computadores inteligentes diseñarán otros, y se harán más listos. La cuestión es dónde quedamos nosotros. Es bastante complicado imaginar una máquina millones de veces más lista que la persona más lista y que, sin embargo, siga siendo nuestra

*esclava y haga lo que queremos. Puede que condesciendan a hablar-nos, puede que jueguen a cosas que nos gusten, puede que nos tengan como mascotas*².

No muy lejos en el tiempo, en 1981, Robert Jastrow, profesor de Astronomía y Geología en la Universidad de Columbia y presidente del Comité de Exploración Lunar de la NASA, escribía lo siguiente en una obra de divulgación sobre el funcionamiento y evolución de cerebro titulada *El telar mágico*:

Hay en acción poderosas fuerzas evolutivas –más culturales que biológicas– que pueden conducir a una forma de vida inteligente más exótica y evolucionada a partir del hombre, pero hija de su cerebro antes que de sus órganos sexuales. [...] Es una vida artificial, hecha de chips de silicio en vez de neuronas [...]. (Jastrow 1985, pp. 145-6).

Jastrow situaba en 1995 el inicio de la competencia entre el hombre y los ordenadores como forma naciente de vida. Es decir, creía que para entonces los ordenadores habrían igualado a los seres humanos en inteligencia. Siguiendo ideas del matemático John Kemeny, Jastrow sostenía que en un principio nuestra relación con ellos sería simbiótica. Los ordenadores satisfacerían necesidades sociales y económicas de los seres humanos y éstos a cambio satisfacerían las necesidades de mantenimiento y reproducción de los ordenadores. El beneficio sería mutuo. Pero la relación simbiótica acabaría cuando los ordenadores fueran mucho más inteligentes que los humanos. A partir de entonces, éstos dejarían de tener utilidad alguna para las máquinas inteligentes. *Ante nosotros surge la visión –añadía en la misma obra– de gigantescos cerebros empapados de la sabiduría de la raza humana y perfeccionándose a partir de ahí. Si esta visión es exacta, el hombre está condenado a un status de subordinación en su propio planeta* (Jastrow 1985, p. 172). Ante la imposibilidad de prescindir de los ordenadores, la única solución que veía Jastrow era la de transferir el contenido de las mentes humanas individuales a ordenadores, de modo que tendríamos mentes humanas con cuerpos de máquina y podríamos alcanzar así la inmortalidad. Una idea bastante socorrida, como vamos a ver.

² Palabras pronunciadas en una entrevista para la BBC TV y recogidas en Copeland 1996, p. 17. Copeland no da la fecha de la entrevista, pero casi las mismas palabras aparecen en la entrevista concedida por Fredkin a Pamela McCorduck y publicada en 1979 (cf. McCorduck 1991, pp. 348 y ss).

El fallo notorio de esta predicción de igualación y superación de la inteligencia humana para 1995 podría imputarse, con algo de buena voluntad, al hecho de que Jastrow no era un investigador en IA, sino un científico interesado en ese campo, que hablaba sin un conocimiento directo del estado de la investigación. Pero los que quizás sean los más entusiastas defensores de un futuro en manos de las máquinas inteligentes sí son conocedores de primera mano de la situación real en dicho campo. Me refiero a Hans Moravec, investigador en el Instituto de Robótica de la Universidad Carnegie-Mellon en Pittsburgh, y a Marvin Minsky, investigador del Laboratorio de Inteligencia Artificial del Massachusetts Institute of Technology y uno de los padres fundadores de la Inteligencia Artificial como disciplina científica.

Hace unos quince años Hans Moravec afirmaba complacido que la potencia de los programas de los robots más avanzados en aquel momento era comparable a la de los sistemas de control de los insectos (cf. Moravec 1986). Consideraba entonces improbable, pero no imposible, la existencia de robots con capacidad humana en un plazo de diez años, es decir, de nuevo para el año 1995. Pero no eran estas sus tesis más llamativas. Sostenía que las máquinas inteligentes serán «habitantes alternativos de nuestro nicho ecológico» (1986, p. 112) y que, por tanto, nuestra existencia estará amenazada incluso aunque dichas máquinas quisieran ser benévolas con nosotros.

Moravec pensaba que sería un error reaccionar ante tal perspectiva suprimiendo la investigación en Inteligencia Artificial y en Robótica. Eso sería ir contra el progreso. *Si los Estados Unidos* —escribía— *detuvieran unilateralmente su desarrollo tecnológico, una idea que ha estado en ocasiones de moda, pronto sucumbirían o bien al poder militar de los soviéticos o a los éxitos económicos de sus socios comerciales.* Y su fuera toda la humanidad en su conjunto la que decidiera no recorrer el camino que abre la IA y que lleva, según sus tesis, a la extinción más que probable de nuestra especie, el resultado sería igualmente la extinción, ya que ese sería el precio a pagar si «por algún milagro maligno e improbable la especie humana decidiera renunciar al progreso» (1986, p. 113). Sin embargo, nada de esto apenaba demasiado a Moravec. Mas bien al contrario, su Apocalipsis particular incluía también la visión de una Nueva Jerusalén constituida por fábricas de robots autorreproductivos diseminadas por los asteroides. Esas fábricas podrían «hacer a alguien inmensamente rico» y, con una tasa de reproducción suficiente, crecerían exponencialmente por el universo (1986, pp. 113-114).

Intentaba convencernos de que la perspectiva del fin para nuestra especie biológica era menos dramática de lo que habíamos pensado siempre, puesto que después de todo dejaríamos descendencia. Una descendencia inesperada, eso sí: los robots inteligentes. Así pues –afirmaba en términos similares a los de Jastrow–, «será nuestra progenie intelectual, no genética, la que heredará el universo», una civilización mecánica que «se llevará consigo todo lo que nosotros consideramos importante, incluyendo la información de nuestras mentes y genes». Esto último les permitirá, si así lo quieren alguna vez, reconstruir de nuevo a los seres humanos (1986, p. 114).

Para competir con alguna posibilidad en esta carrera, Moravec presentaba una alternativa: liberar a nuestra mente del cuerpo mortal que la encierra y trasladarla a un cuerpo mecánico, es decir, hacer de los seres humanos algo radicalmente nuevo, una síntesis de hombre y máquina, capaz de responder en el mismo nivel al desafío de los robots computerizados.

Podríamos, por ejemplo, cuando la tecnología lo permitiera, transferir nuestras mentes a una máquina programada paso a paso para simular perfectamente el comportamiento de todas nuestras neuronas. Mejor aún, podríamos hacer copias mecánicas de nosotros mismos, incluida nuestra mente. Así no moriríamos hasta que se destruyera la última de nuestras copias, porque «una copia fiel es exactamente tan buena como el original» (1986, p. 115). O el modo menos traumático de conseguir la inmortalidad computacional: llevamos toda la vida a nuestro lado un ordenador que aprende a simular todo lo que somos y lo que hacemos, hasta conseguir una copia perfecta de uno mismo. Al morir, el ordenador toma nuestro puesto y –eso al menos asegura Moravec– nadie sufre la pérdida causada por nuestra muerte.

A muchos le podría parecer que una vida eterna así no es vivida por uno mismo sino por sus copias y que, por tanto, no tiene nada de vida eterna. Sólo que en realidad estas copias, según Moravec, son uno mismo. Así que, por extraño que suene el asunto, aunque es uno mismo el que ha muerto, al mismo tiempo es inmortal ya que existe identidad total entre la mente del que muere y sus copias.

Una última posibilidad sugerida es la de integrar en nuestro cerebro, en concreto en el cuerpo calloso, un ordenador que iría sustituyendo las funciones de éste a medida que fuéramos envejeciendo, hasta que finalmente nuestra mente sea la del ordenador.

Estas ideas eran desarrolladas con más detenimiento en su libro de 1988 *Mind Children*. Anuncia allí, desde las primeras páginas, un futuro «postbiológico» y «sobrenatural» en el que el género humano será superado y desplazado, con el orgullo que experimentaría cualquier padre, por su «progenie artificial», por sus «hijos mentales». Sin embargo, la visión del final de los tiempos, que incluye esta vez una resurrección computacional de los muertos (1988, p. 123), asciende ahora a alturas mayores:

Nuestra especulación termina en una supercivilización, síntesis de toda la vida del sistema solar, que constantemente mejora y se expande, que se propaga más allá del sol, que convierte en mente la no-vida. Y posiblemente haya otras burbujas expandiéndose desde algún otro lugar. ¿Qué sucede si nos encontramos con una de ellas? Una posibilidad es la fusión negociada, lo que sólo requeriría un esquema de traducción entre las representaciones de la memoria. Este proceso, que posiblemente está ocurriendo ahora en algún lugar, podría convertir al universo entero en una extensa entidad pensante, un preludio de cosas todavía más grandes. (Moravec 1988, p. 116).

En la actualidad, cuando se han visto ya frustradas bastantes expectativas de los setenta y ochenta en el campo de la IA, Moravec ha moderado levemente su tono, pero no el contenido de su mensaje, ni su optimismo inquietante. En el año 2000 vuelve a afirmar que los robots de hoy en día pueden simular el sistema nervioso de insectos. Si tenemos en cuenta que esto mismo es lo que afirmaba en 1985, cabría suponer que su esperanza inicial de conseguir robots con inteligencia humana en unas pocas décadas debería haberse tambaleado un tanto. Pero no es así. Con una confianza inamovible sitúa en el 2010 la posibilidad de construir robots con la inteligencia de un lagarto, y antes del 2050 los robots nos habrán superado en inteligencia a los humanos. Eso significa, entre otras muchas cosas, que a partir de esa fecha la ciencia la harán los robots. Serán ellos los que investiguen y los que creen cultura. En cuanto a los seres humanos, Moravec no les reserva ya una extinción inevitable. En su lugar se les promete un futuro quizás demasiado apacible para algunos: «Probablemente ocuparán su tiempo en diversas actividades sociales, recreativas y artísticas, no muy distintas de las que hoy llenan el ocio de jubilados o de personas acomodadas» (Moravec 2000, p. 86).

Marvin Minsky comparte desde hace años las ideas de Moravec, incluida la de sustituir nuestros cerebros por máquinas para conseguir la inmortalidad, es decir, según sus propias palabras «convertirnos en ordenadores» para vivir sin límite temporal y superar la muerte. El problema de superpoblación que eso podría crear es solucionado con dos propuestas que, para mantener el tono académico, calificaremos de demasiado audaces: hacer a la gente más pequeña o mandarla a vivir al espacio. Minsky profetiza también con entusiasmo que los robots inteligentes nos sustituirán en un futuro no muy lejano y, como hijos espirituales nuestros, heredarán la Tierra (cf. Minsky 1986 y 1994).

ECOLOGÍA DE LOS ROBOTS

Quizás lo primero que haya que advertir tras la exposición de estas opiniones es que no pueden tomarse como representativas de toda la comunidad de científicos e ingenieros que trabajan en campos relacionados con la computadora electrónica, tanto en el *hardware* como en el *software*. Los especialistas en ciencias de la computación suelen ser bastante realistas acerca de las posibilidades reales de aplicación de los resultados de sus investigaciones. Sus previsiones rara vez van más allá del corto plazo y habitualmente están más interesados en resolver problemas concretos de programación, sujetos en muchos casos a demandas comerciales, que en darse a especulaciones futuristas como las que acabamos de presentar.

Es posible que éstas estén más difundidas entre los investigadores en IA, dado que son ellos los que más implicados están en la tarea de dotar de inteligencia a las máquinas, aunque no todos las compartan. Parece claro, sin embargo, que hay mucho de intención propagandística en ellas y que en gran parte su objetivo ha sido el de llamar la atención de la opinión pública sobre un área de investigación necesitada de grandes recursos de financiación para sus proyectos. Y no se puede negar que se ha logrado un éxito más que notable en ese objetivo propagandístico, en gran parte gracias a la labor personal de Marvin Minsky.

No obstante, con independencia de la intención que haya tras ellas, son tesis que merece la pena considerar en sí mismas, porque son representativas de un modo muy extendido de entender el desarrollo tecnológico y el papel que deben repartirse en él los científicos y los ciu-

dadanos. Su mera formulación pone de manifiesto la insuficiencia de la que todavía adolece el campo de la Inteligencia Artificial en lo que se refiere a la reflexión acerca de qué tipo de máquinas debemos crear y cuánto control estamos dispuestos a delegar en ellas. No hay aún una «tecnológica» que esté a la altura de las reflexiones actuales en bioética, por mencionar el caso más parecido.

No es accidental, en efecto, que los textos que hemos citado estén impregnados de un tono milenarista que está lejos de ser una rareza entre los entusiastas del progreso tecnológico (cf. Noble 1997). Muchos de los grandes científicos e ingenieros que más han contribuido a ese progreso han sido en sus vidas y convicciones una mezcla extraña de apocalípticos e integrados, por utilizar la terminología que hizo famosa Umberto Eco. Desde una visión milenarista pueden justificar con más facilidad cualquier sacrificio capaz de preparar el advenimiento de la nueva era de plenitud. Así es como mejor se puede entender, por ejemplo, la insensibilidad de los autores citados para el sufrimiento humano que prevén que causará su propio trabajo. El problema es que el milenarismo lleva ya más de mil años prediciendo un final inminente de la historia, con el consiguiente advenimiento de un mundo nuevo, sin que por el momento se haya cumplido la predicción, y mientras tanto ha legitimado demasiado sufrimiento en este mundo. Este milenarismo con robots no parece que vaya a ser muy distinto.

Pero lo que quiero destacar ahora es que hay muchos puntos infundados en los argumentos mezclados de profecías que ofrecen Moravec y Minsky. Nos ceñiremos al primero de ambos autores por razones de espacio y por ser el que más atención ha prestado a la fundamentación de sus predicciones.

Para empezar, la cuestión de si los robots heredarán la Tierra junto con el resto de la galaxia ni siquiera se plantearía en el caso de que los robots estuvieran especializados en tareas inteligentes muy específicas y carecieran, por tanto, de una inteligencia diversificada capaz de tratar satisfactoriamente problemas muy distintos, como sucede con la inteligencia humana. Ahora bien, ésta es una situación bastante más probable que la que dibuja Moravec, aunque sólo sea porque desde el punto de vista de la rentabilidad en el mercado interesaría mucho más tener robots del primer tipo que del segundo.

Supongamos, sin embargo, que la inteligencia de los robots no les limita a tareas muy específicas. Las condiciones que deberían cumplir-

se para que el final fuera la extinción de la especie humana por su causa son al menos las tres siguientes:

Los robots deberían estar capacitados desde el principio, o deberían poder capacitarse a sí mismos, para la autoconservación y la reproducción. Y no sólo deberían estar capacitados para ambas cosas, sino que una vez que asumieran el control sobre su propio destino, deberían querer ejercer esa capacidad. Deberían, en suma, tener capacidad y deseo de autoconservación y reproducción. Si no quisieran cuidar de su propio mantenimiento o no quisieran hacer copias de sí mismos, sería su existencia la que estaría condenada de antemano. Esto plantea ya una primera dificultad, pues no es obvio que una máquina, por inteligente que sea, desarrolle por sí sola un deseo por «persistir en el ser» y por multiplicarse.

Deberían, en segundo lugar, tener autonomía para satisfacer sus necesidades. Es decir, deberían ser capaces de subsistir sin los seres humanos. Ello exigiría, entre otras cosas, obtener sus propias fuentes de energía, así como cualquier otro requisito para su funcionamiento. Y dado que han de hacerlo de un modo inteligente, han de poder formar una cierta imagen de sí mismos y de su relación con el mundo exterior. Igualmente, han de poder marcar sus propios fines y determinar su conducta al cumplimiento de los mismos. Es decir, deberían tener un cierto grado de autoconsciencia. Pero tampoco es obvio que puedan (ni deban) tenerla.

En tercer lugar, los recursos utilizados para satisfacer esas necesidades deberían coincidir ampliamente con los que utilizan los seres humanos. Lo cual significa que sus necesidades deberían ser también muy parecidas a las de los seres humanos. Nadie, por lo que yo sé, ha desarrollado todavía una ecología de los robots inteligentes, pero es razonable pensar, dado lo que sabemos de la ecología de las especies biológicas, que si tuvieran necesidades muy distintas, o teniendo necesidades parecidas, el modo de satisfacerlas variara sustancialmente, no tendría por qué producirse una competencia entre hombres y robots por un determinado nicho ecológico.

Moravec da por sentado que las condiciones primera y segunda se cumplirán. Está convencido de que *tarde o temprano nuestras máquinas serán lo suficientemente entendidas como para encargarse sin ninguna ayuda de su propio mantenimiento, reproducción y mejoramiento* (Moravec 1988, p. 4). No obstante, reconoce que podrían faltarles al-

gunas características que suelen considerarse esencialmente unidas a la inteligencia humana: *De hecho, la investigación en robótica es demasiado práctica como para plantearse seriamente el objetivo explícito de producir máquinas con características tan nebulosas y controvertidas como emoción y consciencia* (Moravec 1988, p. 44). Sin embargo, la tercera condición mencionada ni siquiera la discute.

Conviene subrayar antes de nada que si se cumplen estas condiciones y estamos realmente ante un caso de competencia interespecífica entre robots y hombres, hay algo en el discurso de Moravec que no encaja en absoluto. Me refiero a la imagen de unos padres orgullosos de ver cómo sus hijos (mentales) les sustituyen en el manejo de los negocios. Si se da alguna vez tal competencia, un final más probable sería el que recoge la película *Terminator*: la guerra total entre hombres y máquinas. Pero dejemos de lado esa imagen extraña de «los hijos mentales» y sigamos con la metáfora biológica.

Muchas veces, cuando dos especies compiten en la naturaleza, el resultado final no es una exclusión competitiva, es decir, la desaparición de una especie en favor de la otra. De hecho, la mayoría de los casos de competencia interespecífica que pueden ser observados y estudiados son de especies que coexisten sin que ninguna de ellas termine por eliminar a la competidora. En tales casos lo habitual es que ambas especies hayan reducido su nicho ecológico fundamental o precompetitivo para disponer de un nicho efectivo o poscompetitivo marcadamente diferente del de la especie competidora. De este modo, se evita la competencia dentro del mismo nicho y ambas especies reparten los recursos existentes o los explotan de forma distinta (cf. Begon *et al.* 1999, cap. 7 y Rodríguez 1999, cap. 14).

Esto puede ser entendido mejor si lo analizamos en términos del modelo de competencia interespecífica de Lotka-Volterra. Tomemos a los seres humanos como la especie 1 y a los robots inteligentes como la especie 2 y supongamos que entran alguna vez en competencia. Designemos como K_1 al número de individuos que tendría la especie 1 en una situación de equilibrio en ausencia de toda competencia, es decir, al número de individuos que constituye la *capacidad de carga del medio* para dicha especie 1. Designemos como K_2 a la capacidad de carga del medio para la especie 2. Finalmente, sea α_{12} un *coeficiente de competencia* que mide la intensidad del efecto (negativo) de la competencia por individuo de la especie 2 sobre la especie 1 (en relación a la

competencia intraespecífica en la especie 1); y sea \square_{21} el coeficiente de competencia de la especie 1 sobre la 2.

De acuerdo con el modelo de competencia de Lotka-Volterra, hay cuatro resultados posibles en la competencia interespecífica:

- a) Gana la especie 1 y la 2 desaparece.
- b) Gana la especie 2 y la 1 desaparece.
- c) Ambas especies coexisten.
- d) Se da una dominancia indeterminada en la que el resultado final dependerá de las densidades relativas de cada especie en el momento de entrar en competencia.

El resultado que prevé Moravec es el segundo: la especie 2 (los robots) excluye competitivamente a la especie 1 (los humanos). Según el modelo citado, esto ocurrirá únicamente en el caso en que $\square_{12} > K_1/K_2$ y $\square_{21} < K_2/K_1$. Obsérvese que la tasa de crecimiento de cada especie es aquí irrelevante (y lo mismo sucede en los otros tres resultados posibles), es decir, que la rapidez con que se reproduzcan los robots en comparación con los seres humanos no influye en el resultado.

Es evidente que si suponemos que los robots son mucho mejores competidores que los seres humanos (es decir, que $\square_{12} \gg \square_{21}$) y, en especial, que el medio puede sostener a un número mucho más elevado de robots que de seres humanos (es decir, que $K_2 \gg K_1$), entonces se cumplirán fácilmente las anteriores desigualdades y el resultado será que los robots excluyan competitivamente a los humanos. Pero lo cierto es que no hay modo de saber qué valores podrían llegar a tomar estos coeficientes; y mientras no sea posible atribuirles un valor fiable, los cuatro resultados posibles están abiertos.

La reducción del espacio ecológico producida por la coexistencia de las dos especies es una posibilidad que se parece más a las previsiones de Moravec en su trabajo más reciente. En él, como hemos dicho, no se da por segura la extinción de los seres humanos, pero sus vidas quedan reducidas a una existencia plácida y estúpida que recuerda a la que llevaban los Eloi en la novela de H. G. Wells *La máquina del tiempo*.

No obstante, pese a los deseos de Moravec, tampoco es posible predecir cómo habría de ser la eventual reducción del espacio ecológico

de los seres humanos motivada por una competencia con los robots inteligentes. Una alternativa que no puede ser descartada de antemano es que, al contrario del panorama que él describe, las actividades que exigieran más creatividad e imaginación quedaran reservadas para los humanos, mientras que los robots se encontrarán más «a su gusto» en actividades repetitivas o en las que se requiriera una gran capacidad de procesamiento de información.

Por otra parte, si los robots inteligentes fueran tan superiores a los seres humanos como afirma —llega a especular con la posibilidad de máquinas con una inteligencia 10^{30} veces más potente que la humana (cf. 1988, p. 74)—, entonces no parece que pueda hablarse de competencia interespecífica por un mismo nicho ecológico. En efecto, a mayor diferencia fenotípica, mayor grado de diferencia en las necesidades y los recursos explotados y menor grado de competencia interespecífica. De las tesis de Moravec parece seguirse que, a la larga al menos, las necesidades y los recursos utilizados por nuestra «progenie mental» serían muy diferentes a los nuestros. Si esto es así, los robots no serían nuestros competidores, porque en realidad su nicho ecológico no sería el mismo que el nuestro. Incluso su tamaño ideal podría ser finalmente de una escala muy distinta a la de los seres humanos. El poseer una misma característica fenotípica, en este caso la inteligencia, no es una condición suficiente para que se establezca la competencia. Si no hay una similitud en los fenotipos suficiente como para que exista coincidencia en la explotación de los recursos, no se produce tal competencia.

Es más, aventurándonos todavía un poco más de la mano de esa ecología imaginada de los robots, puede argüirse que ninguno de los ecosistemas de la Tierra parece un lugar ideal para dichas máquinas. No sólo tendrían que enfrentarse a una atmósfera rica en oxígeno y en vapor de agua, con el consiguiente efecto degradante y corrosivo sobre sus componentes, sino que en el universo hay lugares mucho mejores para obtener energía de forma más fácil y directa, sin la criba, por ejemplo, de una densa capa atmosférica que refleja gran parte de la radiación solar que recibe. Cabe pensar, por tanto, que tampoco competirían con nosotros por el espacio, ya que muy posiblemente abandonarían este planeta a las primeras de cambio y nos dejarían en él igual que el granjero deja tras de sí a los ratones de su viejo granero cuando se va a la ciudad. Ésta es una posibilidad que, en justicia hay que decirlo, también contempla Moravec. Claro que para él, querer perma-

necer ligado a este planeta es de un provincianismo insufrible (cf. 1988, p. 102).

¿QUIÉNES SON ESOS INMORTALES?

La idea de que podemos alcanzar la inmortalidad personal traspasando nuestra mente a una máquina ha tomado forma en la imaginación de algunos propiciada por la adopción generalizada del funcionalismo en las ciencias cognitivas y, por tanto, también en la IA. El funcionalismo se opone a la identificación que hace materialismo entre el cerebro y la mente o, para ser más precisos, entre tipos de procesos mentales y tipos de procesos cerebrales. Según el funcionalismo, los procesos mentales son estados funcionales de los organismos y, por tanto, se caracterizan no por su soporte material, sino por la función que desempeñan. De acuerdo con esto, una máquina dotada de un programa capaz de simular a la perfección el patrón de entradas y salidas que presenta un cerebro humano cuando se produce en él un determinado proceso mental (el recuerdo de una cara o la visión de un color, por ejemplo), presenta también ese proceso mental (cf. Putnam 1981, cap. 4).

Si el funcionalismo es correcto, daría igual que nuestra mente fuera el resultado del funcionamiento de un cerebro biológico, constituido por células nerviosas, o de un cerebro mecánico, constituido por chips de silicio. Suponiendo que todos los estados funcionales del primero sean realizables por el segundo, ambos cerebros tendrían los mismos procesos mentales. Así pues, sería en principio posible que los seres humanos hubiéramos poseído inteligencia si en lugar estar dotados de un cerebro constituido por materia orgánica hubiésemos estado dotados de un «cerebro» de cualquier otro material capaz de adoptar los mismos estados funcionales.

Hay que ser cuidadosos, sin embargo, con lo que esto significa y lo que no significa. El funcionalismo rechaza la identidad entre tipos de procesos mentales y tipos de procesos cerebrales (*type-type identity*), y en este sentido rechaza la reducción de los procesos mentales a procesos físico-químicos, pero acepta la identidad entre un proceso mental concreto y un estado funcional concreto en un sistema físico, ya sea un cerebro o una máquina (*token-token identity*). En la medida en que un proceso mental se caracteriza funcionalmente, tendrá propiedades

no físicas, y por tanto no se reducirá a procesos fisicoquímicos; pero para que exista un proceso mental se requiere un soporte que sea capaz de presentar el estado funcional que caracteriza a dicho proceso, ya que ese caso de proceso mental consiste precisamente en el caso de estructura causal que adopta el soporte. Esto implica que si tenemos dos soportes materiales en dos estados funcionales iguales, tendremos dos estados mentales *iguales*, pero no *un único* estado mental. Por la sencilla razón de que un mismo estado mental no puede identificarse con dos estados funcionales en soportes diferentes. En eso radica el compromiso materialista del funcionalismo entendido en sentido estricto.

Dicho de otro modo, siendo coherentes con el funcionalismo, una máquina capaz de simular todos los estados funcionales de mi cerebro tendrá los mismos procesos mentales que yo, podrá recordar las mismas cosas que yo o formar los mismos juicios que yo, pero mis procesos mentales y los suyos no serán idénticos, es decir, yo no seré la máquina ni la máquina será yo. Mi proceso mental concreto se identifica con mi estado funcional concreto y el de la máquina con el suyo propio. En otras palabras, una copia mecánica exacta de mi mente no será yo mismo, y el que esa copia pueda sobrevivir a mi muerte no me convierte en inmortal, ni disminuye un ápice el hecho de que la persona que yo soy ha dejado de existir (al menos en este mundo) en el momento de la muerte. Andrew Brook y Robert Stainton han sabido ilustrar la cuestión con un impactante ejemplo:

Imagine que ha ido usted al Centro para la Vida Eterna a hacer que le rejuvenezcan el cuerpo y que transfieran su mente a ese cuerpo reparado. Sube usted a la mesa, oye algunos zumbidos, y las luces se apagan. Cuando desciende de la mesa, un desconcertado celador le explica que ha habido un leve fallo técnico. Le dice: «El modo en que esta tecnología funciona normalmente es el siguiente: se crea un nuevo cuerpo, la información de su cerebro se pasa al cerebro del nuevo cuerpo, y su viejo cuerpo es destruido. El problema es que, aunque hemos creado un nuevo cuerpo y se le ha puesto su programa, desafortunadamente la corriente eléctrica se fue antes de que el cuerpo viejo (i. e. ¡usted!) pudiera ser atomizado». Ahora bien, no se puede permitir que los dos abandonen el centro, así que el celador le hace un simple ruego: «Por favor, vuelva a la mesa para que podamos destruir el cuerpo viejo». (Brook y Stainton 2000, pp. 131-2).

Brook y Stainton afirman que «muchas gente se resistiría a ese ruego». Yo dudo de que alguien, incluidos Moravec y Minsky, lo aceptara. La razón de este rechazo es, según estos autores, que *a pesar de las apariencias iniciales, 'mudar' su mente a otro cuerpo podría ser realmente un modo de morir, no un modo de continuar vivo* (p. 132). Brook y Stainton ponen el dedo en la llaga. No es sólo que una copia de mi mente no sea yo mismo, es que muy posiblemente mi propia mente en otro cuerpo no sería yo mismo. Esto es al menos lo que habría que pensar si consideramos la identidad personal como algo más que la posesión de una mente y una mente como algo más un conjunto de informaciones o como un programa de ordenador³.

Pero esto es precisamente lo que Moravec niega. Para él, el rechazo de la idea de que una copia de mí mismo sea yo mismo, así como el negarse a aceptar que yo no muero mientras quede viva una copia de mí mismo, provienen de una opinión común pero errónea a la que denomina «posición de la identidad-cuerpo» (*body-identity*). Según esta opinión, una persona se define por el material del que está hecho el cuerpo humano. Sin un cuerpo humano que mantenga una continuidad, una persona deja de ser ella misma. Por tanto, no sería ella misma en otro cuerpo que no fuera el suyo, ya sea humano o mecánico. Frente a esa opinión Moravec propone otra que permitiría salvar su visión de la inmortalidad: la «posición de la identidad-patrón» (*pattern-identity*). A semejanza de lo mantiene el funcionalismo, para esta segunda posición lo importante no sería el material del que estamos hechos, sino «el patrón y el proceso» que se dan en el cuerpo y el cerebro. «Si el proceso queda preservado —escribe—, yo quedo preservado. El resto es simple gelatina» (Moravec 1988, p. 117).

Moravec compara esta situación con lo que sucede en nuestros cuerpos con el paso del tiempo. En unos años un ser humano ha cambiado todos y cada uno de los átomos que constituían su cuerpo en un momento dado, y sin embargo, ese ser humano sigue siendo la misma persona a pesar de haber cambiado totalmente la materia que lo integraba. Esto querría decir que la identidad personal reside en lo único que se conserva, o sea, el patrón o la estructura modelo.

Entre las consecuencias que saca de esta idea está la de que la mente y el cuerpo pueden ser separados. Una posición que con toda razón

³ Esta es también una conclusión que parece seguirse de la concepción de la mente como un sistema dinámico encarnado (o incorporado). Ver nota 1.

califica de dualista, ya que, en efecto, las tesis de Moravec van más allá del funcionalismo para caer de lleno en el dualismo. Un funcionalista no podría admitir que la misma mente está simultáneamente en diferentes soportes materiales, un dualista sí.

¿Qué puede decirse de estos argumentos? Es cierto, en primer lugar, que una persona sigue siendo la misma a pesar de los cambios que el tiempo produce en su cuerpo, sin embargo de ahí no se puede concluir que su cuerpo no sea parte de su identidad personal, ni que ésta se pueda reducir a un mero patrón o estructura funcional. No conviene olvidar que los cambios de la edad no hacen que los individuos tengan otro cuerpo, sino que tienen el mismo cuerpo (en el sentido de que no ha sido sustituido por otro), aunque sea un cuerpo distinto al de la juventud (en el sentido de que ha sufrido cambios en su apariencia, en sus componentes y en sus capacidades). Precisamente si la identidad personal se mantiene a través de los cambios del cuerpo es, entre otras razones, porque dichos cambios son experimentados como cambios en el propio cuerpo, no como un cambio de cuerpo. No obstante, es necesario reconocer que la posesión del mismo cuerpo durante toda la vida no excluye la posibilidad de un cambio de identidad personal causado por fuertes trastornos mentales. Es decir, el cuerpo no basta para garantizar la identidad por sí sólo.

Pero tampoco la mente basta para garantizarla. A no ser que se asuma un dualismo radical mente/cuerpo, no es fácil aceptar que sigamos siendo la misma persona, es decir, que se preservara nuestra identidad personal, si nuestra mente dejara nuestro cuerpo o si la cambiáramos a otro cuerpo. Y mucho menos si este otro cuerpo no es humano. Mas bien, como señala Putnam, parece una ingenuidad concebir la mente como una especie de fantasma, capaz de habitar cuerpos diferentes (pero sin ningún cambio en el modo de pensar, de sentir, de recordar y de exhibir la personalidad, si se ha de juzgar según el torrente de libros populares sobre la reencarnación y los «recuerdos de vidas anteriores») o incluso capaz de existir sin un cuerpo (y continuar pensando, sintiendo, recordando y exhibiendo una personalidad) (Putnam 1981, p. 77). Esta es una de las razones que hace del dualismo una postura minoritaria entre los científicos cognitivos.

Por otro lado, si la teoría de la *pattern-identity* fuera correcta podríamos decir que una determinada función matemática capaz de modelar un sistema físico, es idéntica a ese sistema físico, cosa que obviamente no puede hacerse.

La imposibilidad de aceptar que una copia de mí mismo pueda ser yo mismo no proviene de la tesis de la *body-identity*, sino del concepto mismo de identidad. Como ya señaló Kant en su respuesta al principio de los indiscernibles de Leibniz, basta con que dos cosas estén en lugares distintos para que no puedan ser consideradas indiscernibles o idénticas. Si se trata realmente de *dos* cosas, ya no son idénticas. La identidad sólo puede ser de una cosa consigo misma. Una copia puede ser muy parecida al original del que es copia, pero en sentido estricto no puede ser *idéntica* al original, es decir, no puede ser simultáneamente el original y la copia.

En definitiva, si realmente las máquinas superinteligentes llegan a existir y compiten con nosotros por el mismo nicho ecológico (cosa poco probable si aceptamos las consideraciones que hemos hecho acerca del poco interés que nuestro modesto nicho tendría previsiblemente para ellas), la hipotética transmisión de nuestra mente a las máquinas no representaría una alternativa viable.

CONCLUSIONES

Las previsiones de Moravec acerca de la competencia entre seres humanos y robots inteligentes por un mismo nicho ecológico descansan sobre supuestos muy cuestionables desde el punto de vista biológico. En particular los dos siguientes:

- 1) La mera posesión de la inteligencia por parte de las máquinas las convertiría en competidoras de los seres humanos por el mismo nicho ecológico.
- 2) El resultado de esta competencia sería necesariamente la sustitución de una especie (la humana) por la otra (los robots).

En cuanto a la solución propuesta –la búsqueda de la inmortalidad por medio del transvase de nuestra mente a un cuerpo mecánico–, se basa en un dualismo sustentado por una concepción sumamente problemática de la identidad personal.

Hay otras posibilidades que no han sido tenidas en cuenta por Moravec sin que se nos dé razón de este olvido y que merecen una exploración. Podría ocurrir, por ejemplo, que los procesos mentales en los que las máquinas fueran realmente buenas consistieran en proce-

mentos mentales muy distintos de aquéllos en los que los seres humanos fueran realmente buenos. Podría ocurrir que los robots computerizados del futuro fueran mucho más inteligentes que los humanos en procesos de cálculo, de análisis de datos, de elaboración de planes basados en análisis de situaciones complejas, de almacenamiento y recuperación de la información, etc., pero que no tuvieran la autonomía mental y/o física suficiente como para representar una amenaza desde el punto de vista evolutivo para los seres humanos. Su inteligencia podría ser mayor en todo, incluso podrían tener preferencias a la hora de ejecutar planes concretos y, sin embargo, podrían al mismo tiempo carecer de un control voluntario sobre su propio destino.

Una inquietud adicional que no creo que se deba soslayar es la que suscita lo inapropiado de las actitudes que hemos reseñado, que lleguen hasta el punto de considerar un motivo de orgullo y de satisfacción el fin de la especie humana bajo el dominio de la máquina. El ser humano es considerado como algo sumamente deficiente que reclama una profunda transformación, un ser que debe ser reconstruido (o recreado), que debe renacer bajo una nueva forma propiciada por la tecnología y que sólo estos profetas de la tecnología están en disposición de conocer. Su desaparición como especie biológica no significaría además ninguna pérdida digna de lamentarse mientras las máquinas preservaran su cultura.

Es posible que a quienes han magnificado el sentido y la importancia de su trabajo con robots y ordenadores, pensando que con él ponen término a una era y abren otra nueva en la que se alcanzará una meta sublime, pueda servirles de consuelo ante la perspectiva del fin el hecho de que las máquinas inteligentes que nos sustituyan serán nuestros «hijos mentales». A otros, en cambio, más bien les parecerá que esta metáfora no da para consuelo alguno.

Se trata, por otra parte, de actitudes que se manifiestan con la misma contundencia que los viejos moralistas como enemigas del cuerpo, al que sólo se ve como fuente de limitaciones, como un lastre que aparta al hombre de su más alto ideal. Incluso superan a los viejos moralistas en su rígido ascetismo cuando ni siquiera se detienen a reconocer, aunque sea para condenarlo, el poder de la sensualidad corporal como aliante de la vida humana.

¿Cuáles son las razones para llevar a cabo un proyecto semejante de destrucción del ser humano (porque en esto consiste al fin y al cabo

el futuro que se nos anuncia, por mucho que se lo intente pasar por una redención)? Como ya señalara Weizenbaum en su crítica de las ambiciones desmedidas de ciertos representantes destacados de la IA, y como podemos confirmar en los textos de Moravec, sólo se aducen dos razones (cf. Weizenbaum 1984, p. 252-253). En primer lugar se insiste en que si no lo hacemos nosotros, lo harán otros peores que nosotros. En segundo lugar se dice que el progreso tecnológico es imparable y que ninguna limitación o control puede modificar su curso. La primera razón es moralmente inaceptable y resulta incapaz de justificar ninguna acción responsable. La segunda razón da por sentado lo que está en cuestión, a saber, que el control de la tecnología no es posible. Esta tesis resulta empíricamente refutable, ya que vemos habitualmente como ciertas líneas de desarrollo tecnológico no llegan a su ejecución o fracasan al poco de ser realizadas porque chocan con la opinión pública o con otros factores sociales. Es evidente que no podemos renunciar a la tecnología, pero sí podemos –contra lo que defiende el determinista– desobedecer el imperativo tecnológico que convierte en necesario todo lo que es técnicamente posible (cf. Ropohl 1983, Sanmartín 1990 y Niiniluoto 1990). El desarrollo tecnológico es controlable mediante una adecuada política tecnológica y mediante su condicionamiento a una serie de valores aceptados.

Finalmente, todo este discurso apocalíptico sobre la exclusión competitiva de la especie humana frente a las máquinas inteligentes contribuye en mucho a desviar la atención de otros peligros más inmediatos y reales en relación con la computadora electrónica. La dependencia de las máquinas a la hora de tomar decisiones en ámbitos de especial importancia social, el carácter incuestionable con el que se asumen ciertos fines ligados a su uso y difusión, la extensión de su dominio sobre cada vez más aspectos de nuestras vidas, y la dilución de responsabilidades que el propio sistema tecnológico impone, pueden conseguir que el control efectivo del ser humano sobre sus propias acciones disminuya seriamente. Creo que este es un peligro al que debemos por el momento prestarle más atención.

REFERENCIAS

- Begon, M., J. L. Harper & C. R. Townsend (1999): *Ecología*, (trad. M. Riba y R. Salvador), Barcelona: Omega, (3ª edición).

- Brook, A. y R. J. Staiton (2000): *Knowledge and Mind*, Cambridge, Mass.: The MIT Press.
- Brooks, R. A. (1997): «Intelligence without Representation», en J. Haugeland (ed.), *Mind Desing II*, Cambridge, Mass.: The MIT Press, pp. 396-420.
- Clark, A. (1997): *Being There. Putting Brain, Body, and World Together Again*, Cambridge, Mass.: The MIT Press.
- Copeland, J. (1996): *Inteligencia artificial*, (trad. J. C. Armero), Madrid: Alianza.
- Dreyfus, H. L. (1993): *What Computers Still Can't Do*, Cambridge, Mass.: The MIT Press. (Edición revisada de la obra de 1979 *What Computers Can't Do*).
- Jastrow, R. (1985): *El telar mágico*, (trad. D. Santos), Barcelona: Salvat. 1ª ed. en inglés en 1981.
- Martínez Freire, P. (1995): *La nueva filosofía de la mente*, Barcelona: Gedisa.
- (1996): «El futuro de las máquinas pensantes», en *Diálogo Filosófico* 35, pp. 235-250.
- McCORDUCK, P. (1991): *Máquinas que piensan*, (trad. D. Cañamero), Madrid: Tecnos. 1ª ed. en inglés en 1979.
- Minsky, M. (1986): «Nuestro futuro robotizado» en M. Minsky et al., *Robótica*, trad. M. M. Moya i Tasis, Barcelona: Planeta, pp. 241-258. 1ª ed. en inglés en 1985.
- (1994): «¿Serán los robots quienes hereden la Tierra?», en *Investigación y Ciencia*, Diciembre, pp. 87-92.
- Moravec, H. (1986): «Los vagabundos», en M. Minsky et al., *Robótica*, trad. M. M. Moya i Tasis, Barcelona: Planeta, pp. 99-119. 1ª ed. en inglés en 1985.
- 1988: *Mind Children. The Future of Robots and Human Intelligence*, Cambridge, Mass.: Harvard University Press.
- 2000: «El apogeo de los robots», en *Investigación y Ciencia*, Enero, pp. 78-86.
- NIINILUOTO, I. (1990): «Should Technological Imperatives be Obeyed?», *International Studies in the Philosophy of Science*, vol. 4, nº 2, pp. 181-189.
- Noble, D. F. (1997): *The Religion of Technology*, New York: A. Knopf.
- Putnam, H. (1981): *Reason, Truth and History*, Cambridge, Mass.: Cambridge University Press.
- Rodríguez, J. (1999): *Ecología*, Madrid: Pirámide.
- Ropohl, G. (1993): «A Critique of Technological Determinism», en P. T. Durbin y F. Rapp (eds.), *Philosophy and Technology*, Dordrecht: Reidel, pp. 83-96.
- Sanmartín, J. (1990): *Tecnología y futuro humano*, Barcelona: Anthropos.
- Searle, J. (1980): «Mind, Brains, and Programs», *The Behavioral and Brain Sciences*, 3, pp. 417-424. Reimpreso en J. Haugeland (ed.), *Mind Desing II*, Cambridge, Mass.: The MIT Press, 1997.

- Van Gelder, T. (1997): «Dinamics and Cognition», en J. Haugeland (ed.), *MInd Desing II*, Cambridge, Mass.: The MIT Press, pp. 421-450.
- Varela, F., E. Thomson & E. Rosch (1991): *The Embodied Mind: Cognitive Science and Human Experience*, Cambridge, Mass.: The MIT Press.
- Weizenbaum, J. (1984): *Computer Power and Human Reason*, London: Penguin Books.