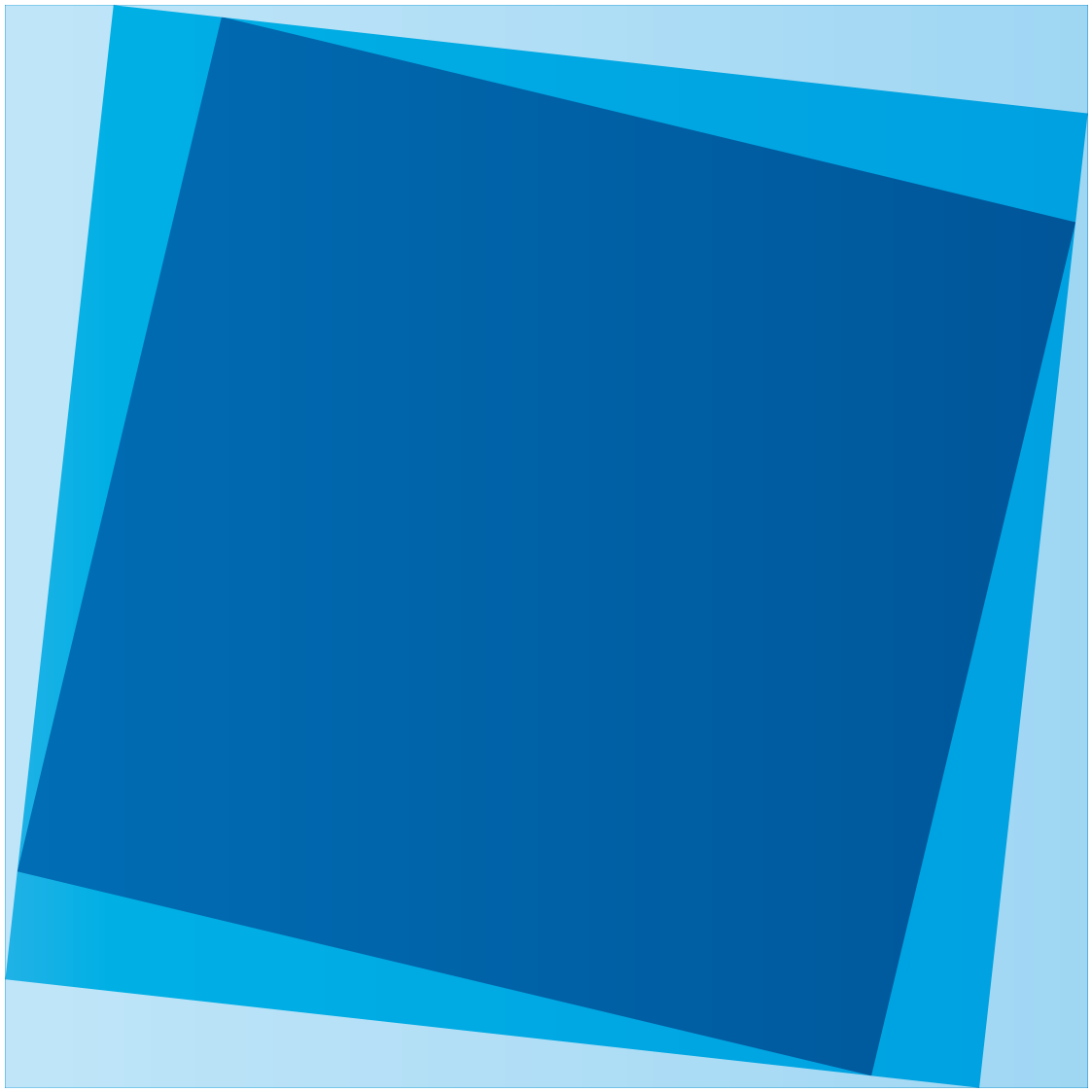
¿Hacia una nueva Ilustración? Una década trascendente



El futuro de la IA: hacia inteligencias artificiales realmente inteligentes

Ramón López de Mántaras

Ramón López de Mántaras
Consejo Superior de Investigaciones Científicas (CSIC)

Profesor investigador del Centro Superior de Investigaciones Científicas (CSIC) y director del Instituto de Investigación de inteligencia artificial (IIIA). Máster en Ingeniería Informática por la Universidad de California Berkeley, doctor en Física (Control Automático) por la Universidad de Toulouse y en Ingeniería Informática por la Universidad Politécnica de Barcelona. Es pionero de la inteligencia artificial (IA) en España. Autor de casi 300 artículos. Conferenciante plenario en numerosos congresos internacionales. Exeditor jefe de la revista *Artificial Intelligence Communications*, es miembro del consejo editorial de varias publicaciones internacionales de prestigio. Ha sido coganador de cinco premios a la mejor ponencia en congresos internacionales. Ha recibido, entre otros premios, el Ciudad de Barcelona a la investigación en 1981; el Robert S. Engelmore Memorial Award de la American Association of Artificial Intelligence (AAAI) en 2011; el nacional de la Sociedad Científica Informática en 2012, el Distinguished Service Award de la European Association of Artificial Intelligence, en 2016 y el IJCAI Donald E. Walker Distinguished Service Award, en 2017. También es miembro del Institut d'Estudis Catalans. Forma parte de distintos paneles de expertos y consejos asesores de instituciones públicas y privadas de Estados Unidos y Europa, tales como el EC Joint Research Center High-Level Peer Group. Su trabajo actual se centra en el razonamiento basado en casos, el aprendizaje automático y las aplicaciones de IA en la música.

Libros recomendados: Domingos, Pedro (2015): *The Master Algorithm* [El algoritmo maestro], Nueva York, Basic Books; y López de Mántaras, Ramón y Meseguer, Pedro (2017): *Inteligencia Artificial*, Madrid, Los Libros de la Catarata.

Este capítulo contiene algunas reflexiones sobre inteligencia artificial (IA). En primer lugar, se explica la distinción entre la IA fuerte y la débil, así como los conceptos relacionados de IA general y específica, dejando claro que todas las manifestaciones existentes de IA son débiles y específicas. Se describen brevemente los principales modelos, insistiendo en la importancia de la corporalidad como aspecto clave para conseguir una IA de naturaleza general. A continuación se aborda la necesidad de proporcionar a las máquinas conocimientos de sentido común que hagan posible avanzar hacia el ambicioso objetivo de construir IA de tipo general. También se comentan las últimas tendencias en IA basadas en el análisis de grandes cantidades de datos que han hecho posibles progresos espectaculares en épocas muy recientes, con una alusión a las dificultades presentes hoy en los enfoques de la IA. Por último, se comentan otras cuestiones que son y continuarán siendo clave en la IA, antes de cerrar con una breve reflexión sobre los riesgos de la inteligencia artificial.

Introducción

El objetivo último de la IA, lograr que una máquina tenga una inteligencia de tipo *general* similar a la humana, es uno de los objetivos más ambiciosos que se ha planteado la ciencia. Por su dificultad, es comparable a otros grandes objetivos científicos como explicar el origen de la vida, el origen del universo o conocer la estructura de la materia. A lo largo de los últimos siglos, este afán por construir máquinas inteligentes nos ha conducido a inventar modelos o metáforas del cerebro humano. Por ejemplo, en el siglo XVII, Descartes se preguntó si un complejo sistema mecánico compuesto de engranajes, poleas y tubos podría, en principio, emular el pensamiento. Dos siglos después, la metáfora fueron los sistemas telefónicos ya que parecía que sus conexiones se podían asimilar a una red neuronal. Actualmente el modelo dominante es el modelo computacional basado en el ordenador digital y, por consiguiente, es el modelo que se contempla en este artículo.

La hipótesis del Sistema de Símbolos Físicos: IA débil versus IA fuerte

En una ponencia, con motivo de la recepción del prestigioso Premio Turing en 1975, Allen Newell y Herbert Simon (Newell y Simon, 1975) formularon la hipótesis del Sistema de Símbolos Físicos según la cual «todo sistema de símbolos físicos posee los medios necesarios y suficientes para llevar a cabo acciones inteligentes». Por otra parte, dado que los seres humanos somos capaces de mostrar conductas inteligentes en el sentido general, entonces, de acuerdo con la hipótesis, nosotros somos también sistemas de símbolos físicos. Conviene aclarar a que se refieren Newell y Simon cuando hablan de *Sistema de Símbolos Físicos* (SSF). Un SSF consiste en un conjunto de entidades denominadas símbolos que, mediante relaciones, pueden ser combinados formando estructuras más grandes —como los átomos que se combinan formando moléculas— y que pueden ser transformados aplicando un conjunto de procesos. Estos procesos pueden generar nuevos símbolos, crear y modificar relaciones entre símbolos, almacenar símbolos, comparar si dos símbolos son iguales o distintos, etcétera. Estos símbolos son físicos en tanto que tienen un sustrato físico-electrónico (en el caso de los ordenadores) o físico-biológico (en el caso de los seres humanos). Efectivamente, en el caso de los ordenadores, los símbolos se realizan mediante circuitos electrónicos digitales y en el caso de los seres humanos mediante redes de neuronas. En definitiva, de acuerdo con la hipótesis SSF, la naturaleza del sustrato (circuitos electrónicos o redes neuronales) carece de importancia siempre y cuando dicho sustrato permita procesar símbolos. No olvidemos que se trata de una hipótesis y, por lo tanto, no debe de ser ni aceptada ni rechazada *a priori*. En cualquier caso, su validez o refutación se deberá verificar de acuerdo con el método científico, con ensayos experimentales. La IA es precisamente el campo científico dedicado a intentar verificar esta hipótesis en el contexto de los ordenadores digitales, es decir, verificar si un ordenador convenientemente programado es capaz o no de tener conducta inteligente de tipo general.

Es importante el matiz de que debería tratarse de inteligencia de tipo general y no una inteligencia específica ya que la inteligencia de los seres humanos es de tipo general. Exhibir inteligencia específica es otra cosa bien distinta. Por ejemplo, los programas que juegan al ajedrez a nivel de Gran Maestro son incapaces de jugar a las damas a pesar de ser un juego mucho más sencillo. Se requiere diseñar y ejecutar un programa distinto e independiente del que le permite jugar al ajedrez para que el mismo ordenador juegue también a las damas. Es decir, que no puede aprovechar su capacidad para jugar al ajedrez para adaptarla a las

damas. En el caso de los seres humanos no es así ya que cualquier jugador de ajedrez puede aprovechar sus conocimientos sobre este juego para, en cuestión de pocos minutos, jugar a las damas perfectamente. El diseño y realización de inteligencias artificiales que únicamente muestran comportamiento inteligente en un ámbito muy específico, está relacionado con lo que se conoce por *IA débil* en contraposición con la *IA fuerte* a la que, de hecho, se referían Newell y Simon y otros padres fundadores de la IA. Aunque estrictamente la hipótesis SSF se formuló en 1975, ya estaba implícita en las ideas de los pioneros de la IA en los años cincuenta e incluso en las ideas de Alan Turing en sus escritos pioneros (Turing, 1948, 1950) sobre máquinas inteligentes.

Quien introdujo esta distinción entre IA débil y fuerte fue el filósofo John Searle en un artículo crítico con la IA publicado en 1980 (Searle, 1980) que provocó, y sigue provocando, mucha polémica. La IA fuerte implicaría que un ordenador convenientemente diseñado no simula una mente sino que *es una mente* y por consiguiente debería ser capaz de tener una inteligencia igual o incluso superior a la humana. Searle en su artículo intenta demostrar que la IA fuerte es imposible. En este punto conviene aclarar que no es lo mismo IA general que IA fuerte. Existe obviamente una conexión pero solamente en un sentido, es decir que toda IA fuerte será necesariamente general pero puede haber IA generales, es decir multitarea, que no sean fuertes, que emulen la capacidad de exhibir inteligencia general similar a la humana pero sin experimentar estados mentales.

La IA débil, por otro lado, consistiría, según Searle, en construir programas que realicen tareas específicas y, obviamente sin necesidad de tener estados mentales. La capacidad de los ordenadores para realizar tareas específicas, incluso mejor que las personas, ya se ha demostrado ampliamente. En ciertos dominios, los avances de la IA débil superan en mucho la pericia humana, como por ejemplo buscar soluciones a formulas lógicas con muchas variables o jugar al ajedrez, o al Go, o en diagnóstico médico y muchos otros aspectos relacionados con la toma de decisiones. También se asocia con la IA débil el hecho de formular y probar hipótesis acerca de aspectos relacionados con la mente (por ejemplo la capacidad de razonar deductivamente, de aprender inductivamente, etcétera) mediante la construcción de programas que llevan a cabo dichas funciones aunque sea mediante procesos completamente distintos a los que lleva a cabo el cerebro. Absolutamente todos los avances logrados hasta ahora en el campo de la IA son manifestaciones de IA débil y específica.

Los principales modelos en IA: simbólico, conexionista, evolutivo y corpóreo

El modelo dominante en IA ha sido el simbólico, que tiene sus raíces en la hipótesis SSF. De hecho, sigue siendo muy importante y actualmente se considera el modelo clásico en IA (también denominado por el acrónimo GOFAI, de *Good Old Fashioned AI*). Es un modelo *top-down* que se basa en el razonamiento lógico y la búsqueda heurística como pilares para la resolución de problemas, sin que el sistema inteligente necesite formar parte de un cuerpo ni estar situado en un entorno real. Es decir, la IA simbólica opera con representaciones abstractas del mundo real que se modelan mediante lenguajes de representación basados principalmente en la lógica matemática y sus extensiones. Por este motivo, los primeros sistemas inteligentes resolvían principalmente problemas que no requieren interactuar directamente con el entorno como, por ejemplo, demostrar sencillos teoremas matemáticos o jugar al ajedrez —los programas que juegan al ajedrez no necesitan de hecho la percepción visual para ver las piezas en el tablero ni actuadores para mover las piezas—. Ello no significa que la IA simbólica no pueda ser usada para, por ejemplo, programar el módulo de razonamiento

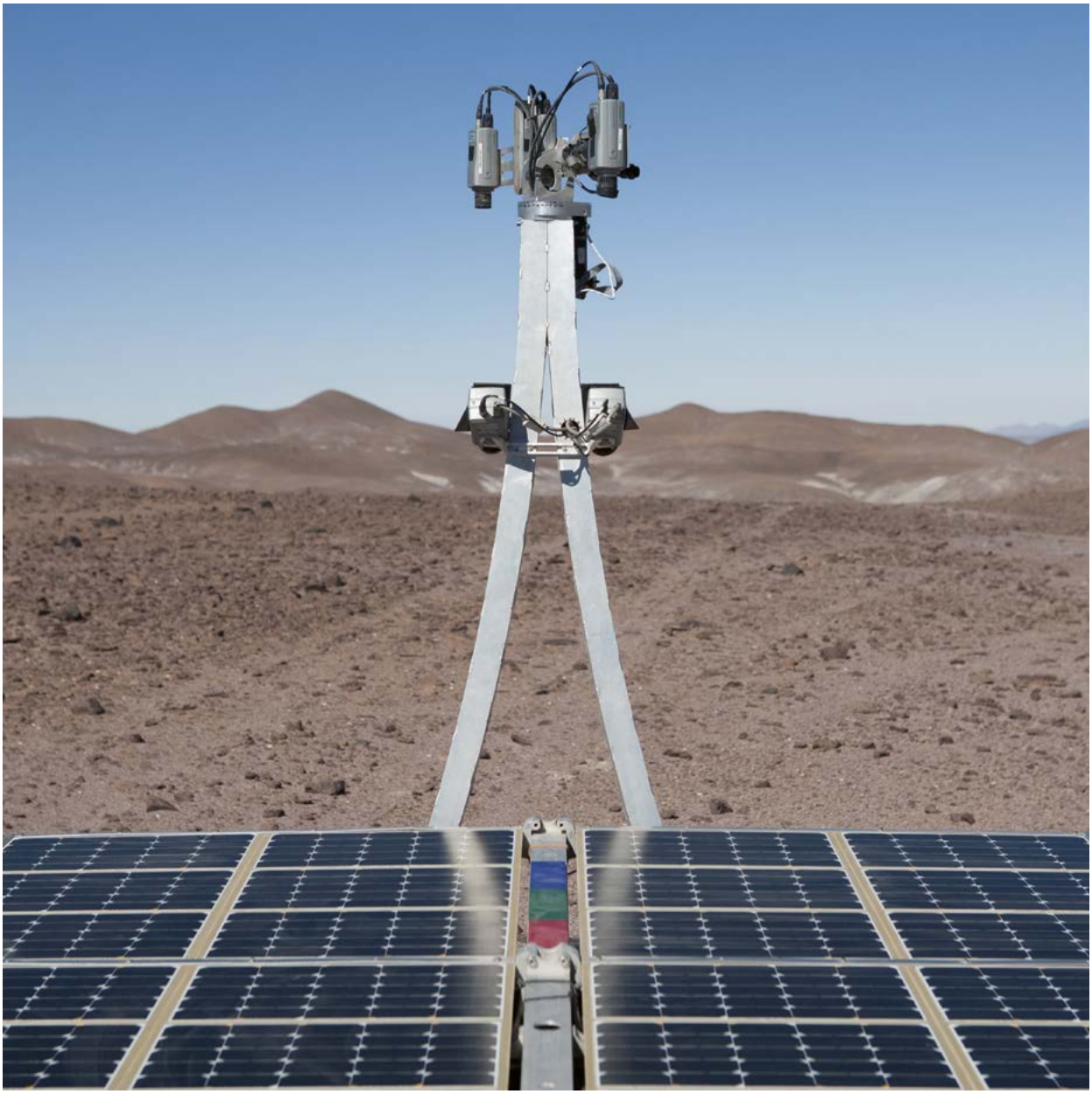
de un robot físico situado en un entorno real, pero en los primeros años los pioneros de la IA no disponían de lenguajes de representación del conocimiento ni de programación que permitieran hacerlo de forma eficiente y por este motivo los primeros sistemas inteligentes se limitaron a resolver problemas que no requerían interacción directa con el mundo real. Actualmente, la IA simbólica se sigue usando para demostrar teoremas o jugar al ajedrez, pero también para aplicaciones que requieren percibir el entorno y actuar sobre él como por ejemplo el aprendizaje y la toma de decisiones en robots autónomos.

El objetivo último de la IA, lograr que una máquina tenga una inteligencia de tipo general similar a la humana, es uno de los objetivos más ambiciosos que se ha planteado la ciencia. Por su dificultad, es comparable a explicar el origen de la vida, el origen del universo o conocer la estructura de la materia

Simultáneamente con la IA simbólica también empezó a desarrollarse una IA bioinspirada llamada conexionista. Los sistemas conexionistas no son incompatibles con la hipótesis SSF pero, contrariamente a la IA simbólica, se trata de una modelización *bottom-up*, ya que se basan en la hipótesis de que la inteligencia emerge a partir de la actividad distribuida de un gran número de unidades interconectadas que procesan información paralelamente. En la IA conexionista estas unidades son modelos muy aproximados de la actividad eléctrica de las neuronas biológicas.

Ya en 1943, McCulloch y Pitts (McCulloch y Pitts, 1943) propusieron un modelo simplificado de neurona en base a la idea de que una neurona es esencialmente una unidad lógica. Este modelo es una abstracción matemática con entradas (dendritas) y salidas (axones). El valor de la salida se calcula en función del resultado de una suma ponderada de las entradas, de forma que si dicha suma supera un umbral preestablecido entonces la salida es un «1», en caso contrario la salida es «0». Conectando la salida de cada neurona con las entradas de otras neuronas se forma una red neuronal artificial. En base a lo que ya se sabía entonces sobre el reforzamiento de las sinapsis entre neuronas biológicas se vio que estas redes neuronales artificiales se podían entrenar para aprender funciones que relacionaran las entradas con las salidas mediante el ajuste de los pesos que sirven para ponderar las conexiones entre neuronas, por este motivo se pensó que serían mejores modelos para el aprendizaje, la cognición y la memoria, que los modelos basados en la IA simbólica. Sin embargo, los sistemas inteligentes basados en el conexionismo tampoco necesitan formar parte de un cuerpo ni estar situados en un entorno real y, desde este punto de vista, tienen las mismas limitaciones que los sistemas simbólicos. Por otra parte, las neuronas reales poseen complejas arborizaciones dendríticas con propiedades no solo eléctricas sino también químicas nada triviales. Pueden contener conductancias iónicas que producen efectos no lineales. Pueden recibir decenas de millares de sinapsis variando en posición, polaridad y magnitud. Además, la mayor parte de las células del cerebro no son neuronas, son células *gliales*, que no solamente regulan el funcionamiento de las neuronas, también poseen potenciales eléctricos, generan ondas de calcio y se comunican entre ellas, lo que parece indicar que juegan un papel muy importante en los procesos cognitivos. Sin embargo, no existe ningún modelo conexionista que incluya a dichas células por lo que, en el mejor de los casos, estos modelos son muy incompletos y, en el peor, erróneos. En definitiva, toda la enorme complejidad del cerebro queda muy lejos de los modelos actuales. Esta inmensa complejidad del cerebro

Ingenieros de la Universidad Carnegie Mellon desarrollaron esta robot llamada *Zoe* para que detectara vida en entornos aparentemente deshabitados. *Zoe* incorpora un sistema puntero de detección de moléculas orgánicas que quizá ayude a encontrar vida en Marte; y es 20 veces más rápida que los robots exploradores de marte *Spirit* y *Opportunity*. Desierto de Atacama, Chile, 2005




DENSO
04:58
電王戦

El diseño y la realización de inteligencias artificiales que solo muestran comportamiento inteligente en un ámbito muy específico están relacionados con lo que se conoce por *IA débil* en contraposición con la *IA fuerte*

Masayuki Toyoshima, jugador profesional de *shogi*, el ajedrez japonés, juega contra el programa de ordenador YSS, que mueve las piezas mediante un brazo robótico. Osaka, Japón, marzo de 2014

también conduce a pensar que la llamada *singularidad*, es decir, futuras superinteligencias artificiales que, basadas en réplicas del cerebro, superarán con mucho la inteligencia humana en un plazo de unos veinticinco años, es una predicción con poco fundamento científico.

Otra modelización bioinspirada, también compatible con la hipótesis SSF, y no corpórea, es la *computación evolutiva* (Holland, 1975). Los éxitos de la biología evolucionando organismos complejos, hizo que a primeros de los años sesenta algunos investigadores se plantearan la posibilidad de imitar la evolución con el fin de que los programas de ordenador, mediante un proceso evolutivo, mejorasen automáticamente las soluciones a los problemas para los que habían sido programados. La idea es que estos programas, gracias a operadores de mutación y cruce de «cromosomas» que modelan a los programas, generan nuevas generaciones de programas modificados cuyas soluciones son mejores que las de los programas de las generaciones anteriores. Dado que podemos considerar que el objetivo de la IA es la búsqueda de programas capaces de producir conductas inteligentes, se pensó que se podría usar la programación evolutiva para encontrar dichos programas dentro del espacio de programas posibles. La realidad es mucho más compleja y esta aproximación tiene muchas limitaciones aunque ha producido excelentes resultados, en particular en la resolución de problemas de optimización.

La complejidad del cerebro dista mucho de los modelos de IA y conduce a pensar que la llamada *singularidad* —superinteligencias artificiales basadas en réplicas del cerebro que superarán con mucho la inteligencia humana— es una predicción con poco fundamento científico

Una de las críticas más fuertes a estos modelos no corpóreos se basa en que un agente inteligente necesita un cuerpo para poder tener experiencias directas con su entorno (diríamos que el agente está «situado» en su entorno) en lugar de que un programador proporcione descripciones abstractas de dicho entorno codificadas mediante un lenguaje de representación de conocimientos. Sin un cuerpo, estas representaciones abstractas no tienen contenido semántico para la máquina. Sin embargo, gracias a la interacción directa con el entorno, el agente puede relacionar las señales que percibe a través de sus sensores con representaciones simbólicas generadas a partir de lo percibido. Algunos expertos en IA, en particular Rodney Brooks (Brooks, 1991) incluso llegaron a afirmar que no era ni siquiera necesario generar dichas representaciones internas, esto es, que no es necesario que un agente tenga que tener una representación interna del mundo que le rodea ya que el propio mundo es el mejor modelo posible de sí mismo y que la mayor parte de las conductas inteligentes no requieren razonamiento sino que emergen a partir de la interacción entre el agente y su entorno. Esta idea generó mucha polémica y el propio Brooks, unos años más tarde, admitió que hay muchas situaciones en las que una representación interna del mundo es necesaria para que el agente tome decisiones racionales.

En 1965, el filósofo Hubert Dreyfus afirmó que el objetivo último de la IA, es decir, la IA fuerte de tipo general, era tan inalcanzable como el objetivo de los alquimistas del siglo XVII que pretendían transformar el plomo en oro (Dreyfus, 1965). Dreyfus argumentaba que el cerebro procesa la información de manera global y continua mientras que un ordenador utiliza un conjunto finito y discreto de operaciones deterministas aplicando reglas a un conjunto finito de datos. En este aspecto podemos ver un argumento similar al de Searle, pero Dreyfus, en

posteriores artículos y libros (Dreyfus, 1992), usó también otro argumento consistente en que el cuerpo juega un papel crucial en la inteligencia. Fue pues uno de los primeros en abogar la necesidad de que la inteligencia forme parte de un cuerpo con el que poder interactuar con el mundo. La idea principal es que la inteligencia de los seres vivos deriva del hecho de estar situados en un entorno con el que pueden interactuar gracias a sus cuerpos. De hecho esta necesidad de corporeidad está basada en la Fenomenología de Heidegger que enfatiza la importancia del cuerpo con sus necesidades, deseos, placeres, penas, formas de moverse, de actuar, etcétera. Según Dreyfus, la IA debería modelar todos estos aspectos para alcanzar el objetivo último de la IA fuerte. Dreyfus no niega completamente la posibilidad de la IA fuerte pero afirma que no es posible con los métodos clásicos de la IA simbólica y no corpórea, en otras palabras considera que la hipótesis del Sistema de Símbolos Físicos no es correcta. Sin duda se trata de una idea interesante que hoy en día comparten muchos investigadores en IA. Efectivamente, la aproximación corpórea con representación interna ha ido ganando terreno en la IA y actualmente muchos la consideramos imprescindible para avanzar hacia inteligencias de tipo general. De hecho, basamos una gran parte de nuestra inteligencia en nuestra capacidad sensorial y motora. En otras palabras, el cuerpo conforma a la inteligencia y por lo tanto sin cuerpo no puede haber inteligencia de tipo general. Esto es así porque el *hardware* del cuerpo, en particular los mecanismos del sistema sensorial y del sistema motor, determinan el tipo de interacciones que un agente puede realizar. A su vez, estas interacciones conforman las habilidades cognitivas de los agentes dando lugar a lo que se conoce como *cognición situada*. Es decir, se sitúa a la máquina en entornos reales, como ocurre con los seres humanos, con el fin de que tengan experiencias interactivas que, eventualmente, les permitan llevar a cabo algo similar a lo que propone la teoría del desarrollo cognitivo de Piaget (Inhelder y Piaget, 1958), según la cual un ser humano sigue un proceso de maduración mental por etapas y quizá los distintos pasos de este proceso podrían servir de guía para diseñar máquinas inteligentes. Estas ideas ha dado lugar a una nueva subárea de la IA llamada *robótica del desarrollo* (Weng *et al*., 2001).

Éxitos de la IA especializada

Todos los esfuerzos de la investigación en IA se han centrado en construir inteligencias artificiales especializadas y los éxitos alcanzados son muy impresionantes, en particular durante el último decenio gracias sobre todo a la conjunción de dos elementos: la disponibilidad de enormes cantidades de datos y el acceso a la computación de altas prestaciones para poder analizarlos. Efectivamente, el éxito de sistemas, como por ejemplo AlphaGo (Silver *et al*., 2016), Watson (Ferrucci *et al.*, 2013) y los avances en vehículos autónomos o en diagnóstico médico basado en imágenes, han sido posibles gracias a esta capacidad para analizar grandes cantidades de datos y detectar patrones eficientemente. Sin embargo, prácticamente no hemos avanzado hacia la consecución de IA general. De hecho, podemos afirmar que los actuales sistemas de IA son una demostración de lo que Daniel Dennet llama «competencia sin comprensión» (Dennet, 2018).

Posiblemente la lección más importante que hemos aprendido a lo largo de los sesenta años de existencia de la IA es que lo que parecía más difícil (diagnosticar enfermedades, jugar al ajedrez y al Go al más alto nivel) ha resultado ser relativamente fácil y lo que parecía más fácil ha resultado ser lo más difícil. La explicación a esta aparente contradicción hay que buscarla en la dificultad de dotar a las máquinas de conocimientos de sentido común. Sin estos conocimientos no es posible una comprensión profunda del lenguaje ni una interpretación profunda de lo que capta un sistema de percepción visual, entre otras limitaciones. De hecho,

La corporación tecnológica IBM ha abierto una división dedicada a su Watson idC (internet de las cosas) dentro de las torres Highlight, en Múnich, Alemania. El desarrollo de nuevas soluciones para la idC se hace mediante inteligencia artificial

el sentido común es requisito fundamental para alcanzar una IA similar a la humana en cuanto a generalidad y profundidad. Los conocimientos de sentido común son fruto de nuestras vivencias y experiencias. Algunos ejemplo son: «el agua siempre fluye de arriba hacia abajo», «para arrastrar un objeto atado a una cuerda hay que tirar de la cuerda, no empujarla», «un vaso se puede guardar dentro de un armario pero no podemos guardar un armario dentro de un vaso», etcétera. Hay millones de conocimientos de sentido común que las personas manejamos fácilmente y que nos permiten entender el mundo en el que vivimos. Una posible línea de investigación que podría dar resultados interesantes en adquisición de conocimientos de sentido común es la robótica del desarrollo mencionada anteriormente. Otra línea de trabajo muy interesante es la que tiene como objetivo la modelización matemática y el aprendizaje de relaciones causa-efecto, es decir, el aprendizaje de causales y, por lo tanto, asimétricos del mundo. Los sistemas actuales basados en aprendizaje profundo simplemente pueden aprender funciones matemáticas simétricas, no pueden aprender relaciones asimétricas y por consiguiente no son capaces de diferenciar entre causas y efectos, como por ejemplo que la salida del sol es la causa del canto del gallo y no lo contrario (Pearl, 2018; Lake *et al.*, 2016).

Futuro: hacia inteligencias artificiales realmente inteligentes

Las capacidades más complicadas de alcanzar son aquellas que requieren interaccionar con entornos no restringidos ni previamente preparados. Diseñar sistemas que tengan estas capacidades requiere integrar desarrollos en muchas áreas de la IA. En particular, necesitamos lenguajes de representación de conocimientos que codifiquen información acerca de muchos tipos distintos de objetos, situaciones, acciones, etcétera, así como de sus propiedades y de las relaciones entre ellos, en particular relaciones causa-efecto. También necesitamos nuevos algoritmos que, en base a estas representaciones, puedan, de forma robusta y eficiente, resolver problemas y responder preguntas sobre prácticamente cualquier tema. Finalmente, dado que necesitarán adquirir un número prácticamente ilimitado de conocimientos, estos sistemas deberán ser capaces de aprender de forma continua a lo largo de toda su existencia. En definitiva, es imprescindible diseñar sistemas que integren percepción, representación, razonamiento, acción y aprendizaje. Este es un problema muy importante en IA, ya que todavía no sabemos como integrar todos estos componentes de la inteligencia. Necesitamos arquitecturas cognitivas (Forbus, 2012) que integren estos componentes de forma adecuada. Los sistemas integrados son un paso previo fundamental para conseguir algún día inteligencias artificiales de tipo general.

Las capacidades más complicadas de alcanzar son aquellas que requieren interaccionar con entornos no restringidos ni previamente preparados. Diseñar sistemas que tengan estas capacidades requiere integrar desarrollos en muchas áreas de la IA

Entre las actividades futuras, creemos que los temas de investigación más importantes pasarán por sistemas híbridos que combinen las ventajas que poseen los sistemas capaces de razonar en base a conocimientos y uso de la memoria (Graves *et al.*, 2016) y las ventajas de la IA basada en análisis de cantidades masivas de datos, en lo que se conoce por aprendizaje

profundo (Bengio, 2009). Actualmente, una importante limitación de los sistemas de aprendizaje profundo es el denominado «olvido catastrófico», lo cual significa que si una vez han sido entrenados para llevar a cabo una tarea (por ejemplo, jugar al Go), si a continuación los entrenamos para llevar a cabo otra tarea distinta (por ejemplo, distinguir entre imágenes de perros y de gatos) olvidan completamente la tarea anteriormente aprendida (en este caso jugar al Go). Esta limitación es una prueba contundente de que en realidad estos sistemas no aprenden nada, por lo menos en el sentido humano de aprender. Otra importante limitación de estos sistemas es que son «cajas negras» sin capacidad explicativa, por ello un objetivo interesante de investigación será como dotar de capacidad explicativa a los sistemas de aprendizaje profundo incorporando módulos que permitan explicar como se ha llegado a los resultados y conclusiones propuestas, ya que la capacidad de explicación es una característica irrenunciable en cualquier sistema inteligente. También es necesario desarrollar nuevos algoritmos de aprendizaje que no requieran enormes cantidades de datos para ser entrenados así como un *hardware* mucho mas eficiente en consumo energético para implementarlos, ya que el consumo de energía podría acabar siendo una de las principales barreras al desarrollo de la IA. En comparación, el cerebro es varios órdenes de magnitud más eficiente que el *hardware* actual necesario para implementar los algoritmos de IA más sofisticados. Una posible vía a explorar es la computación Neuromórfica basada en memristores (Saxena *et al*., 2018).

Otras técnicas más clásicas de IA que seguirán siendo objeto de investigación extensiva son los sistemas multiagente, la planificación de acciones, el razonamiento basado en la experiencia, la visión artificial, la comunicación multimodal persona-máquina, la robótica humanoide y sobre todo las nuevas tendencias en *robótica del desarrollo* que puede ser la clave para dotar a las máquinas de sentido común y, en particular, aprender la relación entre sus acciones y los efectos que estas producen en el entorno. También veremos progresos significativos gracias a las aproximaciones biomiméticas para reproducir en máquinas el comportamiento de animales. No se trata únicamente de reproducir el comportamiento de un animal sino de comprender como funciona el cerebro que produce dicho comportamiento. Se trata de construir y programar circuitos electrónicos que reproduzcan la actividad cerebral que genera este comportamiento. Algunos biólogos están interesados en los intentos de fabricar un cerebro artificial lo más complejo posible porque consideran que es una manera de comprender mejor el órgano, y los ingenieros buscan información biológica para hacer diseños más eficaces. Mediante la biología molecular y los recientes avances en optogenética será posible identificar qué genes y qué neuronas juegan un papel clave en las distintas actividades cognitivas.

La *robótica del desarrollo* puede ser la clave para dotar a las máquinas de sentido común y, en particular, aprender la relación entre sus acciones y los efectos que estas producen en el entorno

En cuanto a las aplicaciones, algunas de las más importantes seguirán siendo aquellas relacionadas con la web, los videojuegos, los asistentes personales y los robots autónomos (en particular vehículos autónomos, robots sociales, robots para la exploración de planetas, etcétera). Las aplicaciones al medio ambiente y ahorro energético también serán importantes, así como las dirigidas a la economía y la sociología.

Por último, las aplicaciones de la IA al arte (artes visuales, música, danza, narrativa) cambiarán de forma importante la naturaleza del proceso creativo. Los ordenadores ya no son solamente herramientas de ayuda a la creación, los ordenadores empiezan a ser agentes creativos. Ello ha dado lugar a una nueva y muy prometedora área de aplicación de la IA denominada *Creatividad Computacional* que ya ha producido resultados muy interesantes (Colton *et al*., 2009, 2015; López de Mántaras, 2016) en ajedrez, música, artes plásticas y narrativa, entre otras actividades creativas.

Reflexión final

Por muy inteligentes que lleguen a ser las futuras inteligencias artificiales, incluidas las de tipo general, nunca serán iguales a las inteligencias humanas ya que, tal como hemos argumentado, el desarrollo mental que requiere toda inteligencia compleja depende de las interacciones con el entorno y estas interacciones dependen a su vez del cuerpo, en particular del sistema perceptivo y del sistema motor. Ello, junto al hecho de que las máquinas no seguirán procesos de socialización y culturización como los nuestros, incide todavía más en que, por muy sofisticadas que lleguen a ser, serán inteligencias distintas a las nuestras. El que sean inteligencias ajenas a la humana y, por lo tanto, ajenas a los valores y necesidades humanas nos debería hacer reflexionar sobre posibles limitaciones éticas al desarrollo de la IA. En particular, estamos de acuerdo con la afirmación de Weizenbaum (Weizenbaum, 1976) de que ninguna máquina debería nunca tomar decisiones de forma completamente autónoma o dar consejos que requieran, entre otras cosas, de la sabiduría, producto de experiencias humanas, así como de tener en cuenta valores humanos.

Por muy inteligentes que lleguen a ser las futuras inteligencias artificiales nunca serán como la humana; el desarrollo mental que requiere toda inteligencia compleja depende de las interacciones con el entorno y estas dependen a su vez del cuerpo, en particular de los sistemas perceptivo y motor

El verdadero peligro de la IA no es la muy improbable singularidad tecnológica debida a la existencia de unas futuras hipotéticas superinteligencias artificiales, los verdaderos peligros ya están aquí. Actualmente los algoritmos en que se basan los motores de búsqueda en internet, los sistemas de recomendación y los asistentes personales de nuestros teléfonos móviles, conocen bastante bien lo que hacemos, nuestras preferencias y nuestros gustos e incluso pueden llegar a inferir el qué pensamos y cómo nos sentimos. El acceso a cantidades masivas de información, que voluntariamente generamos, es fundamental para que esto sea posible, ya que mediante el análisis de estos datos provenientes de fuentes diversas es posible encontrar relaciones y patrones que serían imposibles de detectar sin las técnicas de IA. Todo esto resulta en una pérdida alarmante de privacidad. Para evitarlo deberíamos tener derecho a poseer una copia de todos los datos personales que generamos, controlar su uso y decidir a quién le permitimos el acceso y bajo qué condiciones, en lugar de que estén en manos de grandes corporaciones sin poder saber realmente qué uso hacen de nuestros datos.

La IA está basada en programación compleja, y por lo tanto necesariamente cometerá errores. Pero incluso suponiendo que fuera posible desarrollar un *software* completamente fiable, hay dilemas éticos que los desarrolladores de *software* deben tener en cuenta a la hora de diseñarlo. Por ejemplo, un vehículo autónomo podría decidir atropellar a un peatón para evitar una colisión que podría causar daños a sus ocupantes. Equipar las empresas con sistemas avanzados de IA para hacer la gestión y la producción más eficientes requerirá menos empleados humanos y generará más paro. Estos dilemas éticos hacen que muchos expertos en IA señalen la necesidad de regular su desarrollo. En algunos casos se debería incluso de prohibir el uso de la IA. Un ejemplo claro son las armas autónomas. Los tres principios básicos que rigen los conflictos armados: discriminación (la necesidad de discernir entre combatientes y civiles o entre un combatiente en actitud de rendirse y uno dispuesto a atacar), proporcionalidad (evitar el uso desmedido de fuerza) y precaución (minimización del número de víctimas y daños materiales) son extraordinariamente difíciles de evaluar y, por lo tanto, casi imposibles de cumplir por los sistemas de IA que controlan las armas autónomas. Pero incluso en el caso de que a muy largo plazo las máquinas tuvieran esta capacidad, sería indigno delegar en una máquina la decisión de matar. Pero, además de regular, es imprescindible educar a los ciudadanos sobre los riesgos de las tecnologías inteligentes, dotándolos de las competencias necesarias para controlarla en lugar de ser controlados por ella. Necesitamos futuros ciudadanos mucho más informados, con más capacidad para evaluar los riesgos tecnológicos, con más sentido crítico y dispuestos a hacer valer sus derechos. Este proceso de formación debe comenzar en la escuela y tener continuación en la universidad. En particular es necesario que los estudiantes de ciencia e ingeniería reciban una formación ética que les permita comprender mejor las implicaciones sociales de las tecnologías que muy probablemente desarrollarán. Solo si invertimos en educación lograremos una sociedad que pueda aprovechar las ventajas de las tecnologías inteligentes minimizando los riesgos. La IA tiene sin duda un extraordinario potencial para beneficiar a la sociedad siempre y cuando hagamos un uso adecuado y prudente. Es fundamental aumentar la conciencia de las limitaciones de la IA, así como actuar de forma colectiva para garantizar que la IA se utilice en beneficio del bien común con seguridad, fiabilidad y responsabilidad.

El camino hacia la IA realmente inteligente seguirá siendo largo y difícil, al fin y al cabo la IA tiene apenas sesenta años y, como diría Carl Sagan, sesenta años son un brevísimo momento en la escala cósmica del tiempo; o, como muy poéticamente dijo Gabriel García Márquez: «Desde la aparición de vida visible en la Tierra debieron transcurrir 380 millones de años para que una mariposa aprendiera a volar, otros 180 millones de años para fabricar una rosa sin otro compromiso que el de ser hermosa, y cuatro eras geológicas para que los seres humanos fueran capaces de cantar mejor que los pájaros y morirse de amor».

Bibliografía

—Bengio, Yoshua (2009): «Learning deep architectures for AI», en *Foundations and Trends in Machine Learning*, vol. 2, n.º 1, pp. 1-127.
—Brooks, Rodney A. (1991): «Intelligence without reason», *Proceedings of the 12th International Joint Conference on Artificial Intelligence* (IJCAI'91), vol. 1, pp. 569-595.
—Colton, S.; López de Mántaras, R. y Stock, O. (2009): «Computational creativity: coming of age», *en AI Magazine*, vol. 30, n.º 3, pp. 11-14.
—Colton, S.; Halskov, J.; Ventura, D.; Gouldstone, I.; Cook, M. y Pérez-Ferrer, B. (2015): «The painting fool sees! New projects with the automated painter», I*nternational Conference on Computational Creativity* (ICCC 2015), pp. 189-196.
—Dennet, D. C. (2018): *From Bacteria to Bach and Back: The Evolution of Minds*, Londres, Penguin Random House.
—Dreyfus, Hubert L. (1965): *Alchemy and Artificial Intelligence*, Santa Mónica, California, Rand Corporation.
—(1992): *What Computers Still Can't Do*, Nueva York, MIT Press.
—Ferrucci, D. A.; Levas, A.; Bagchi, S.; Gondek, D. y Mueller, E. T. (2013): «Watson: beyond jeopardy!», en *Artificial Intelligence*, n.º 199, pp. 93-105.
—Forbus, Kenneth. D. (2012): «How minds will be built», en *Advances in Cognitive Systems*, n.º 1, pp. 47-58.
—Graves, A.; Wayne, G.; Reynolds, M.; Harley, T.; Danihelka, I.; Grabska-Barwińska, A.; Gómez-Colmenarejo, S.; Grefenstette, E.; Ramalho, T.; Agapiou, J.; Puigdomènech-Badia, A.; Hermann, K. M.; Zwols, Y.; Ostrovski, G.; Cain, A.; King, H.; Summerfield, C.; Blunsom, P.; Kavukcuoglu, K. y Hassabis, D. (2016): «*Hybrid* computing using a neural network with dynamic external memory», *en Nature*, n.º 538, pp. 471-476.
—Holland, John. H. (1975): *Adaptation in natural and artificial systems*, Michigan, University of Michigan Press.
—Inhelder, Bärbel y Piaget, Jean (1958): *The Growth of Logical Thinking from Childhood to Adolescence*, Nueva York, Basic Books.
—Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B. y Gershman, S. J. (2017): «Building machines that learn and think like people», en *Behavioral and Brain Sciences*, vol. 40, e253.
—López de Mántaras, R. (2016): «Artificial intelligence and the arts: toward computational creativity», en AA VV, *The Next Step: Exponential Life*, Madrid, BBVA/Turner, pp. 100-125.
—McCulloch, Warren. S. y Pitts, Walter (1943): «A logical calculus of ideas immanent in nervous activity», en *Bulletin of Mathematical Biophysics*, n.º 5, pp. 115-133.
—Newell, Allen y Simon, Herbert A. (1976): «Computer science as empirical inquiry: symbols and serch», en *Communications of the ACM*, vol. 19, n.º 3, pp. 113-126.
—Pearl, Judea y Mackenzie, Dana (2018): *The Book of Why: The New Science of Cause and Effect*, Nueva York, Basic Books.
—Saxena, V.; Wu, X.; Srivastava, I. y Zhu, K. (2018): «Towards neuromorphic learning machines using emerging memory devices with brain-like energy efficiency, preprints». Disponible en www.preprints.org
—Searle, John R. (1980): «Minds, brains, and programs», en *Behavioral and Brain Science*, vol. 3, n.º 3, pp. 417-457.
—Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Ven den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T. y Hassabis, D. (2016): «Mastering the game of go with deep neural networks and tree search», en *Nature*, vol. 529, n.º 7.587, pp. 484-489.
—Turing, Alan M. (1948): *Intelligent Machinery, National Physical Laboratory Report*, reimpreso en B. Meltzer y D. Michie (eds.) (1969): *Machine Intelligence 5*, Edimburgo, Edinburgh University Press, 1969.
—(1950): «Computing machinery and intelligence», en *Mind*, vol. 59, n.º 236, pp. 433-460.
—Weizenbaum, Joseph (1976): *Computer Power and Human Reasoning: From Judgment to Calculation*, San Francisco, W. H. Freeman and Company.
—Weng, J.; McClelland, J.; Pentland, A.; Sporns, O.; Stockman, I.; Sur, M. y Thelen, E. (2001): «Autonomous mental development by robots and animals», *en Science*, n.º 291, pp. 599-600.

OpenMind

ESCUCHA EL AUDIO DE ESTE CAPÍTULO

ACCEDE AL LIBRO COMPLETO

- ¿Hacia una nueva Ilustración? Una década trascendente
- Towards A New Enlightenment? A Trascendent Decade

ACCESO AL ARTÍCULO EN INGLÉS

The Future of AI: Toward Truly Intelligent Artificial Intelligences

CÓMO CITAR ESTE ARTÍCULO

López de Mántaras, R., "El futuro de la IA: hacia inteligencias artificiales realmente inteligentes", en ¿Hacia una nueva Ilustración? Una década trascendente, Madrid, BBVA, 2018.

ARTÍCULOS RELACIONADOS

LEER MÁS SOBRE #TECNOLOGÍA #INTELIGENCIA ARTIFICIAL

- El futuro de la comunicación humano-máquina: el test de Turing
  El próximo paso: la vida exponencial, Huma Shah y Kevin Warwick

- El futuro de la inteligencia artificial y la cibernética
  Hay futuro: visiones para un mundo mejor, Kevin Warwick

- La inteligencia artificial y las artes. Hacia una creatividad computacional
  El próximo paso: la vida exponencial, Ramón López de Mántaras

TODOS LOS TÍTULOS DE LA COLECCIÓN OPENMIND