MONOGRÁFICO
MÈTODE SCIENCE STUDIES JOURNAL (2018). Universitat de València.
DOI: 10.7203/metode.9.11145
Artículo recibido: 12/12/2017, aceptado: 23/07/2018.

HACIA LA INTELIGENCIA ARTIFICIAL

PROGRESOS, RETOS Y RIESGOS

RAMON LÓPEZ DE MÁNTARAS

Este artículo contiene algunas reflexiones acerca de la inteligencia artificial (IA). En primer lugar se distingue entre IA fuerte y débil y los conceptos relacionados de IA general y específica. A continuación, se describen brevemente los principales modelos existentes. También se discute la necesidad de poder dotar de conocimientos de sentido común a las máquinas para avanzar hacia el objetivo de construir IA general. Después hablamos de las tendencias en IA basada en el análisis de grandes cantidades de datos que han permitido alcanzar progresos espectaculares muy recientemente. Para finalizar, tratamos otros temas que son y continuarán siendo clave en IA y concluimos con una breve reflexión sobre los riesgos de la IA.

Palabras clave: inteligencia artificial fuerte, inteligencia artificial débil, conocimientos de sentido común, aprendizaje profundo.

INTRODUCCIÓN

El objetivo último de la inteligencia artificial (IA), conseguir que una máquina tenga una inteligencia de tipo general similar a la humana, es una de las metas más ambiciosas que se ha planteado la ciencia. Plantea una dificultad comparable a otros grandes objetivos científicos como explicar el origen de la vida o el origen del universo, o bien 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 se plasmó en los sistemas telefónicos, ya que parecía que sus conexiones se podían asimilar a una red neuronal. Actualmente el modelo dominante es el computacional basado en el ordenador digital y por tanto es el modelo al que nos referiremos en este artículo.

> «LA IA ES EL CAMPO CIENTÍFICO DEDICADO A VERIFICAR SI UN ORDENADOR ES CAPAZ O NO DE TENER UNA CONDUCTA INTELIGENTE DE TIPO GENERAL»

INTELIGENCIA ARTIFICIAL DÉBIL 'VERSUS' FUERTE

Allen Newell y Herbert Simon formularon la hipótesis de que todo sistema de símbolos físicos posee los medios necesarios y suficientes para llevar a cabo acciones inteligentes (Newell y Simon, 1976). Por otro lado, dado que los seres humanos somos capaces de mostrar conductas inteligentes, de acuerdo con la hipótesis, nosotros somos también sistemas de símbolos físicos. Conviene aclarar a qué se refieren Newell y Simon. Un sistema de símbolos físicos consiste en un conjunto de entidades llamadas símbolos que, mediante relaciones, pueden ser combinados para formar estructuras mayores –como los átomos que se combinan formando moléculas– y que pueden ser transformados aplicando un conjunto de procedimientos. Estos procedimientos pueden crear nuevos símbolos, crear y modificar relaciones entre estos, almacenar, comparar si dos son iguales o diferentes, etcétera. Estos símbolos son físicos en tanto que tienen un sustrato fisicoelectrónico (en el caso de los ordenadores) o fisicobiológico (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 del sistema de símbolos físicos, la naturaleza del sustrato (circuitos electrónicos o redes neuronales) no tiene importancia, siempre que este permita procesar símbolos. No olvide-

mos que se trata de una hipótesis y por tanto su validez o refutación se deberá verificar de acuerdo con el método científico. La inteligencia artificial es precisamente el campo científico dedicado a intentar verificar esta hipótesis en el contexto de los ordenadores, es decir, verificar si un ordenador convenientemente programado es capaz o no de tener una conducta inteligente de tipo general.

Es importante matizar que se debería tratar de inteligencia de tipo general, y no una específica, ya que la inteligencia de los seres humanos es de tipo general. Exhibir inteligencia específica es algo muy diferente. Por ejemplo, los programas que juegan al ajedrez al nivel de gran maestro son incapaces de jugar a las damas. Se requiere un programa diferente para que el mismo ordenador juegue a las damas; es decir, este no puede aprovechar el hecho de que juega al ajedrez para adaptarse y jugar también a las damas. En el caso de los seres humanos, cualquier jugador de ajedrez puede aprovechar sus conocimientos sobre este juego para jugar a las damas perfectamente. La inteligencia artificial que únicamente muestra comportamiento inteligente en un ámbito muy específico está relacionada con lo que se conoce como «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.

Quien introdujo esta distinción entre IA débil y fuerte fue el filósofo John Searle en un artículo crítico con la inteligencia artificial publicado en 1980 (Searle, 1980) que provocó, y continúa provocando, mucha polémica. La IA fuerte implicaría que un ordenador convenientemente programado no simula una mente sino que «es una mente» y por tanto tendría que ser capaz de pensar igual que un ser humano. 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 solo en un sentido; es decir, que toda IA fuerte será necesariamente general pero puede haber IA generales que no sean fuertes, esto es, que simulen la capacidad de exhibir inteligencia general de la mente pero sin ser mentes.

La IA débil, por otro lado, consistiría, según Searle, en construir programas que realicen tareas específicas. La capacidad de los ordenadores para realizar tareas específicas incluso mejor que las personas ya se ha demostrado sobradamente en ciertos dominios, como buscar soluciones a fórmulas lógicas con muchas variables y otros aspectos relacionados con la toma de decisiones. También se asocia con la IA débil el hecho de formular y probar hipótesis sobre aspectos relacionados con la mente (por ejemplo, la capacidad de razonar deductivamente, de aprender inductivamente, etc.) mediante la construcción de programas que llevan a cabo estas funciones aunque sea mediante procesos completamente diferentes de los que lleva a cabo el cerebro. Absolutamente todos los avances conseguidos hasta ahora en el campo de la IA son manifestaciones de la IA débil y específica.

LOS PRINCIPALES MODELOS EN INTELIGENCIA ARTIFICIAL

Hasta muy recientemente, el modelo dominante en IA ha sido el simbólico. Este modelo tiene las raíces en la hipótesis del sistema de símbolos físicos. Aún continúa siendo muy importante y actualmente se considera el modelo «clásico» en IA. Es un modelo descendiente (*top-down*) que se basa en el razonamiento lógico y la investigación 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

Universidad Carnegie Mellon

En la década de los setenta, Allen Newell y Herbert Simon plantearon que todo sistema de símbolos físicos –ya sean estos fisicoelectrónicos en el caso de los ordenadores o fisicobiológicos en el de los seres humanos– posee los medios necesarios para llevar a cabo acciones inteligentes. En la imagen, los profesores Simon (a la izquierda) y Newell (a la derecha), trabajando en la programación de ajedrez a finales de los años cincuenta, en la Universidad Carnegie Mellon en Pittsburgh (EE UU).

IBM

Los seres humanos poseen una inteligencia de tipo general, mientras que los programas que juegan al ajedrez al nivel de gran maestro, como es el caso del ordenador Deep Blue que consiguió ganar al campeón Kasparov en 1997, tienen una inteligencia de tipo específico. Eso quiere decir que son incapaces de utilizar sus conocimientos para jugar, por ejemplo, a las damas. En la imagen, el equipo de IBM que desarrolló el ordenador Deep Blue, en una imagen de 1996.

mediante lenguajes de representación basados principalmente en la lógica matemática y sus extensiones. Por este motivo, los primeros sistemas inteligentes resolvían sobre todo problemas que no requerían interactuar directamente con el entorno, como demostrar sencillos teoremas matemáticos o jugar a ajedrez. Eso no quiere decir que la IA simbólica no se pueda usar para programar el módulo de razonamiento de un robot físico situado en un entorno real, pero en los primeros años de la IA no había lenguajes de representación del conocimiento ni de programación que permitiesen hacerlo de forma eficiente. Actualmente, la IA simbólica se sigue usando para demostrar teoremas o jugar a ajedrez, pero también para aplicaciones que requieren percibir el entorno y actuar sobre él, como el aprendizaje y la toma de decisiones en robots autónomos.

Simultáneamente con la IA simbólica también empezó a desarrollarse una IA bioinspirada denominada conexionista. Contrariamente a la IA simbólica, se trata de una modelización ascendente (*bottom-up*), ya que se basa 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. McCulloch y Pitts (1943) propusieron un modelo simplificado de neurona de acuerdo con la idea de que esta es esencialmente una unidad lógica. Este modelo es una abstracción matemática con entradas y salidas, que se corresponderían, respectivamente, con las dendritas y los axones. El valor de la salida se calcula en función del resultado de una suma ponderada de las entradas, de forma que si esta 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. De acuerdo con lo que ya se sabía entonces sobre la afirmación de las sinapsis entre neuronas biológicas, se vio que estas redes neuronales artificiales se pueden entrenar para que aprendan funciones que relacionen las entradas con las salidas mediante el ajuste de los pesos que sirven para ponderar la fuerza de las conexiones entre neuronas. Por este motivo se pensó que la cognición y la memoria serían mejores modelos para el aprendizaje 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.

«LA CAPACIDAD DE LOS ORDENADORES PARA REALIZAR TAREAS ESPECÍFICAS INCLUSO MEJOR QUE LAS PERSONAS YA SE HA DEMOSTRADO SOBRADAMENTE»

Por otro lado, el 90 % de las células del cerebro no son neuronas sino las llamadas células gliales, que no solamente regulan el funcionamiento de las neuronas sino que también poseen potenciales eléctricos, generan ondas de calcio y se comunican entre ellas, lo cual parece indicar que también representan un papel muy importante en los procesos cognitivos. No obstante, no hay ningún modelo conexionista que incluya estas células; por tanto, en el mejor de los casos, estos modelos son muy incompletos. Eso hace pensar que la llamada singularidad, es decir, futuras superinteligencias artifi-

ciales que, basadas en réplicas del cerebro, superasen por mucho la inteligencia humana en un plazo de unos veinte años, es una predicción con poco fundamento.

Otra modelización bioinspirada, también compatible con la hipótesis del sistema de símbolos físicos, y no corpórea, es la computación evolutiva. El éxito de la biología para hacer evolucionar organismos complejos hizo que a principios de los años sesenta algunos investigadores se planteasen la posibilidad de imitar la evolución para que los programas de ordenador, mediante un proceso evolutivo, mejorasen automáticamente las soluciones a los problemas para los cuales habían sido programados. La idea es que estos programas, gracias a operadores de mutación y cruce de los «cromosomas» que modelan los programas, produzcan nuevas generaciones de programas modificados de tal forma que sus soluciones sean 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 aplicar la programación evolutiva para encontrar estos 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.

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 en vez de que un programador proporcione descripciones abstractas de este entorno codificadas mediante un lenguaje de representación de conocimientos. Sin un cuerpo, estas representaciones abstractas no tienen contenido semántico para la máquina. No obstante, mediante la interacción directa con el entorno, el agente puede relacionar las señales que percibe mediante sus sensores con representaciones simbólicas generadas a partir de lo que ha percibido.

En 1965, el filósofo Hubert Dreyfus publicó un artículo titulado «Alchemy and artificial intelligence» (Dreyfus, 1965) en el que 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 argumentaba que el cerebro procesa la información de forma global y continua mientras que un ordenador utiliza un conjunto finito y discreto de operaciones deterministas, es decir, aplicando reglas a un conjunto finito de datos. En este aspecto podemos ver un argumento similar al de Searle, pero Dreyfus, en artículos y libros posteriores, usó también otro argumento basado en el papel crucial que el cuerpo representa en la inteligencia (Dreyfus, 1992). Fue, pues, uno de

Softbank Robotics

Una de las críticas de los modelos no corpóreos de IA se basa en que un agente inteligente necesita un cuerpo para poder tener experiencias directas con su entorno. En la imagen, el robot humanoide Romeo desarrollado por Softbank Robotics.

«POR MUY SOFISTICADAS QUE LLEGUEN A SER, LAS INTELIGENCIAS DE LAS MÁQUINAS SERÁN DIFERENTES DE LAS NUESTRAS»

IBM

La disponibilidad de enormes cantidades de datos y el acceso a la computación de altas prestaciones para analizarlos ha permitido desarrollar nuevos sistemas de inteligencia artificial como Watson, capaz de responder a preguntas formuladas en lenguaje natural. Según la compañía IBM, que lo ha desarrollado, Watson es capaz de aprender en cada experiencia.

los primeros en abogar por la necesidad de que la inteligencia forme parte de un cuerpo con el que poder interaccionar 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. De hecho, esta necesidad de corporeidad se basa en la fenomenología de Heidegger, que enfatiza la importancia del cuerpo con sus necesidades, deseos, placeres, penas, formas de moverse, de actuar, etc. Según Dreyfus, la IA debería modelar todos estos aspectos para alcanzar el objetivo último de la IA fuerte. Es decir, que Dreyfus no niega completamente la posibilidad de la IA fuerte, pero afirma que no es posible con los métodos clásicos de la IA no corpórea.

¿LOS PROGRESOS DE LA IA ESPECÍFICA NOS ACERCAN A LA INTELIGENCIA ARTIFICIAL GENERAL?

Prácticamente todos los proyectos en IA se han centrado en construir inteligencias artificiales especializadas y los éxitos alcanzados en solo sesenta años de existencia, y en particular durante el último decenio, son muy impresionantes, principalmente gracias 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 AlphaGo (Silver et al., 2016), Watson (Ferrucci, Levas, Bagchi, Gondek y Mueller, 2013) y los avances en vehículos autónomos han sido posibles gracias a esta capacidad para analizar grandes cantidades de datos. No obstante, no hemos avanzado nada hacia la consecución de IA general. De hecho, 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, o jugar a ajedrez y a Go al más alto nivel) ha resultado factible y lo que parecía más fácil (comprender el significado profundo del lenguaje o interpretar una escena visual) aún no se ha alcanzado.

La explicación a esta aparente contradicción hay que buscarla en la dificultad de dotar a las máquinas de conocimientos de sentido común. El sentido común es el requisito fundamental para conseguir una IA similar a la humana en lo que respecta a generalidad y profundidad. Los conocimientos de sentido común son fruto de vivencias y experiencias obtenidas interactuando con nuestro entorno. 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. Las capacidades más complicadas de alcanzar son aquellas que requieren interaccionar con entornos no restrictivos 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 sobre muchos tipos diferentes de objetos, situaciones, acciones, etc., así como de sus propiedades y de las relaciones entre ellos.

> «EL SENTIDO COMÚN ES EL REQUISITO FUNDAMENTAL PARA CONSEGUIR IA SIMILAR A LA HUMANA EN LO QUE RESPECTA A GENERALIDAD Y PROFUNDIDAD»

También necesitamos nuevos algoritmos que, a partir de estas representaciones, puedan responder, de forma robusta y eficiente, a preguntas sobre prácticamente cualquier tema. Finalmente, como necesitarán conocer un número prácticamente ilimitado de cosas, estos sistemas tienen que ser capaces de aprender nuevos conocimientos de forma continua a lo largo de toda su existencia. En definitiva, es imprescindible diseñar sistemas que integren percepción, representación, razonamiento, acción y aprendizaje. Solo combinando estos elementos dentro de sistemas cognitivos integrados podremos empezar a construir IA general.

PASADO RECIENTE Y FUTURO A CORTO PLAZO DE LA INTELIGENCIA ARTIFICIAL

Entre las actividades futuras, creemos que los temas de investigación más importantes continuarán basándose

en lo que se conoce en inglés por *massive data-driven AI*, es decir, explotar la posibilidad de acceder a cantidades masivas de datos y poderlos procesar con *hardware* cada vez más rápido para descubrir relaciones entre ellos, detectar patrones y realizar inferencias y aprendizaje mediante modelos probabilísticos como los sistemas de aprendizaje profundo (Bengio, 2009). No obstante, estos sistemas basados en el análisis de enormes cantidades de datos tendrán que incorporar en el futuro módulos que permitan explicar cómo se ha llegado a los resultados y conclusiones que proponen, ya que la capacidad de explicación es una característica irrenunciable en cualquier sistema inteligente, pues permite comprender cómo funciona el sistema y evaluar su confiabilidad. Por otro lado, esto también es necesario para corregir posibles errores de programación y detectar si los datos de entrenamiento están sesgados.

Waymo

El desarrollo de tecnologías relacionadas con la inteligencia artificial hace necesario analizar los riesgos que pueden conllevar. Por ejemplo, en el caso del pilotaje automático de vehículos autónomos necesitamos una metodología para probar y validar que los sistemas de aprendizaje automático alcancen altos niveles de precisión.

Hay que saber si las respuestas que nos dan son correctas por las razones correctas o se deben a coincidencias que puede haber en el conjunto de datos de entrenamiento. Por eso, uno de los temas de investigación más importantes en aprendizaje profundo es diseñar aproximaciones interpretables de estos sistemas complejos de aprendizaje. Una aproximación consiste no solo en entrenar el sistema de aprendizaje profundo sino que, con los mismos datos, también se entrena otro sistema que lo mimetiza usando una representación sencilla y transparente.

Otro tema de investigación muy actual es la verificación y validación del *software* que implementa el algoritmo de aprendizaje. Eso es especialmente importante en aplicaciones de alto riesgo como el pilotaje automático de vehículos autónomos. En estos casos, necesitamos una metodología para probar y validar que estos sistemas de aprendizaje automático alcanzan altos niveles de precisión. Una idea que se está explorando actualmente se conoce como aprendizaje adversario (*adversarial learning* en inglés) y consiste en entrenar un segundo sistema de IA que trata de «romper» el *software* de aprendizaje intentando encontrar los puntos débiles. Por ejemplo, en el caso del reconocimiento visual, generando imágenes que provoquen que el sistema tome la decisión equivocada.

«INCLUSO SUPONIENDO QUE FUERA POSIBLE DESARROLLAR ‘SOFTWARE’ COMPLETAMENTE FIABLE, HAY ASPECTOS ÉTICOS QUE LOS PROGRAMADORES DEBEN TENER EN CUENTA»

OTROS TEMAS CLAVE EN INTELIGENCIA ARTIFICIAL

Otras áreas de la IA que continuarán 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, la robótica social y las nuevas tendencias en robótica del desarrollo que pueden ser clave para dotar las máquinas de sentido común. También veremos progresos significativos gracias a las aproximaciones biomiméticas para reproducir en máquinas el comportamiento de animales. Algunos biólogos están interesados en los intentos de fabricar un cerebro artificial lo más complejo posible porque consideran que es una forma de comprender mejor el órgano y los ingenieros, por su parte, buscan información biológica para hacer diseños más eficaces.

Otras áreas importantes de interés para la IA, y en particular para la robótica, son la ciencia de ma-

teriales y la nanotecnología. Por ejemplo, para el desarrollo de músculos artificiales, cartílagos artificiales y sistemas sensoriales como pieles artificiales.

Por lo que respecta a las aplicaciones, algunas de las más importantes continuarán siendo las relacionadas con la red, los videojuegos, y los robots autónomos (en particular vehículos autónomos, robots sociales, robots para la exploración de planetas, etc.). Las aplicaciones en medio ambiente y ahorro energético también serán importantes, así como en la economía y la sociología.

Finalmente, las aplicaciones de la IA en el arte cambiarán de forma importante la naturaleza del proceso creativo. Los ordenadores ya no son solo herramientas de ayuda a la creación, sino que empiezan a ser agentes creativos. Eso ha dado lugar a una nueva y muy prometedora área de aplicación de la inteligencia artificial denominada creatividad computacional que ya ha producido resultados muy interesantes (Colton, López de Mántaras y Stock, 2009; Colton et al., 2015; López de Mántaras, 2016) en música, artes plásticas y narrativa, entre otras actividades creativas.

LOS RIESGOS DE LA INTELIGENCIA ARTIFICIAL, UNA REFLEXIÓN FINAL

Por muy inteligentes que lleguen a ser las futuras inteligencias artificiales, en particular las de tipo general, nunca serán iguales a las inteligencias humanas, ya que, 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. Eso, junto al hecho de que las máquinas no seguirán procesos de socialización y culturización como los nuestros, incide aún más en el hecho de que, por muy sofisticadas que lleguen a ser, serán inteligencias diferentes de las nuestras. El hecho de ser inteligencias ajenas a la humana y por tanto ajenas a los valores y necesidades humanas nos tendría que hacer reflexionar sobre posibles limitaciones éticas al desarrollo de la inteligencia artificial. En particular, estamos de acuerdo con Weizenbaum (1976) en que ninguna máquina debería tomar decisiones de manera 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.

La IA se basa en programación compleja, y por tanto necesariamente contendrá errores. Pero incluso suponiendo que fuera posible desarrollar *software* completamente fiable, hay aspectos éticos que los programadores deben tener en cuenta a la hora de diseñarlo. Estos aspectos éticos hacen que muchos expertos en IA señalen la necesidad de regular su desarrollo. Sin embargo, además de regular, es imprescindible educar a los ciudadanos sobre los riesgos de las tecnologías inteligentes, dotándolos de las competencias necesarias para controlarlas en vez de ser controlados por ellas. Este proceso de formación debe empezar 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 conseguiremos una sociedad que pueda aprovechar las ventajas de las tecnologías inteligentes minimizando los riesgos.

REFERENCIAS

Bengio., Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–127. doi: 10.1561/2200000006

Colton, S., Halskov, J., Ventura, D., Gouldstone, I., Cook, M., & Pérez-Ferrer, B. (2015). The Painting Fool sees! New projects with the automated painter. En *International Conference on Computational Creativity (ICCC 2015)* (pp. 189–196). Utah, UT: Brighma Young University.

Colton, S., López de Mántaras, R., & Stock, O. (2009). Computational creativity: Coming of age. *AI Magazine, 30*(3), 11–14. doi: 10.1609/aimag.v30i3.2257

Dreyfus, H. L. (1965). *Alchemy and artificial intelligence*. Santa Mónica, CA: RAND Corporation.

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

Ferrucci, D. A., Levas, A., Bagchi, S., Gondek, D., & Mueller, E. T. (2013). Watson: Beyond Jeopardy! *Artificial Intelligence*, *199*, 93–105. doi: 10.1016/j.artint.2012.06.009

López de Mántaras, R. (2016). Artificial intelligence and the arts: Toward computational creativity. En *The next step: Exponential life* (pp. 100–125). Madrid: BBVA.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133. doi: 10.1007/BF02478259

Newell, A., & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126. doi: 10.1145/360018.360022

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. doi: 10.1017/S0140525X00005756

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van den Driessche, ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. doi: 10.1038/nature16961

Weizenbaum, J. (1976). *Computer power and human reasoning: From judgment to calculation*. San Francisco, CA: W. H. Freeman and Co.

Ramon López de Mántaras. Profesor de investigación y director del Instituto de Investigación en Inteligencia Artificial del CSIC (Bellaterra, España). Doctor en Física por la Universidad Paul Sabatier de Toulouse, *master of Science* en Informática por la Universidad de California-Berkeley y doctor en Informática por la Universidad Politécnica de Cataluña. Es miembro numerario del Institut d'Estudis Catalans. Actualmente investiga en razonamiento por analogía, en técnicas de aprendizaje automático en robots humanoides y en inteligencia artificial aplicada a la música, áreas en las que ha publicado cerca de 300 artículos científicos. Publicó en 2017 el libro de divulgación *Inteligencia artificial* dentro de la colección de libros «Que sabemos de» (Los Libros de la Catarata). ✉ mantaras@iiia.csic.es