TECNOLOGÍA

EL TRAJE NUEVO DE LA INTELIGENCIA ARTIFICIAL

Cada vez más decisiones de calado se están dejando en manos de supuestas máquinas inteligentes que no comprenden absolutamente nada. Por el bien de todos, urge una revisión crítica de los logros de este campo de investigación

Ramon López de Mántaras

Ramon López de Mántaras es fundador y exdirector del Instituto de Investigación en Inteligencia Artificial del CSIC, en Barcelona. Ha destacado por sus trabajos en reconocimiento de patrones, razonamiento basado en casos y aprendizaje basado en la experiencia. En 2018 recibió el Premio Nacional de Investigación.

HOY EN DÍA ESTAMOS VIVIENDO UNA NUEVA PRIMAVERA de la inteligencia artificial. Y, al igual que en primaveras anteriores, abundan las predicciones de que la llegada de máquinas dotadas de una inteligencia general igual o superior a la humana será cuestión de algunos decenios, y de que esto nos llevará a la llamada «singularidad»: el momento en que las máquinas lo harán todo mucho mejor que nosotros, incluida la propia investigación científica, lo que dará lugar a una nueva etapa evolutiva conocida como posthumanismo.

¿Es esta primavera de la IA, vestida con un traje nuevo, el indicador de que, efectivamente, estamos cerca de alcanzar el sueño de la inteligencia artificial general? ¿O quizá la inteligencia artificial está desnuda, como el emperador del cuento de Hans Christian Andersen, y el momento actual no es sino una etapa más del larguísimo camino hacia ese sueño?

En las líneas que siguen argumentaré que, en efecto, la inteligencia artificial (IA) sigue estando desnuda. Para entender por qué, es necesario analizar el origen de la fiebre que estamos viviendo, cuáles son las aplicaciones concretas que han dado lugar a todo tipo de declaraciones y titulares exagerados, y cómo funcionan realmente tales aplicaciones y de qué adolecen. Como veremos, la IA actual está muy lejos de alcanzar el objetivo de la IA general. Y ello no se debe a que aún queden por afinar unos pocos detalles o a una falta de potencia de cómputo, sino al enfoque que desde hace unos años ha adoptado esta disciplina. Lo que debería darnos miedo no es ninguna singularidad futura debido a la hipotética existencia de superinteligencias artificiales, sino un presente en el que estamos encomendando cada vez más decisiones a máquinas estúpidas.

EL NACIMIENTO DE UNA FIEBRE

El actual entusiasmo por la IA se debe a los recientes logros de la técnica conocida como aprendizaje profundo (el «traje nuevo») en el contexto del reconocimiento de imágenes, los juegos de tablero y el procesamiento del lenguaje.

Todo comenzó en 2012, cuando un equipo de la Universidad de Toronto liderado por Geoffrey Hinton consiguió que un tipo de red neuronal, llamada «convolucional», alcanzara un 85 por ciento de aciertos al clasificar, entre mil categorías posibles, 150.000 imágenes de la base ImageNet. Tales redes habían sido introducidas en 1980 a partir de los trabajos del investigador japonés Kunishiko Fukushima, quien había desarrollado el «neocognitrón», una red neuronal artificial inspirada, a su vez, en los estudios de David Hubel y Torsten Wiesel sobre el sistema visual de los animales, trabajos por los que en 1981 estos investigadores recibieron el premio Nobel [*véase* «Mecanismos cerebrales de la visión», por David H. Hubel y Torsten N. Wiesel; Investigación y Ciencia, noviembre de 1979].

Hubel y Wiesel descubrieron que nuestra corteza visual se encuentra organizada según una jerarquía de capas, de tal manera que las neuronas contenidas en cada capa detectan características de complejidad creciente en los objetos de una imagen. Por ejemplo, las neuronas de la primera capa se activan cuando detectan rasgos simples, como los bordes de los objetos. Después transmiten su nivel de activación a las neuronas de la segunda capa, donde se detectan características algo más complejas, que, en esencia, corresponden a combinaciones de los rasgos detectados en la capa anterior (por ejemplo, un conjunto de bordes que dan lugar a un polígono, un círculo, una elipse, etcétera). El proceso continúa hasta llegar a la última capa, la cual detecta objetos enteros y hace posible identificarlos. Por ejemplo, si la imagen contiene un rostro, las elipses detectadas en una de las capas intermedias corresponderían a los ojos y en la última capa se reconocería la cara entera.

Las redes convolucionales implementan computacionalmente este proceso jerárquico. Pero, si se conocen desde 1980, ¿por qué tuvieron que transcurrir más de treinta años para

EN SÍNTESIS

La inteligencia artificial (IA) vive un nuevo auge. Los éxitos de la técnica conocida como aprendizaje profundo han sido presentados por muchos científicos, compañías y medios de comunicación como una prueba de que la IA general está cerca. ¿Es cierto?

La realidad es muy otra. Un análisis pausado revela que los algoritmos actuales siguen siendo propensos a errores catastróficos, carecen de capacidad de razonamiento y contextualización, y no poseen nada remotamente parecido al sentido común humano.

Ello se explica porque, en los últimos años, la investigación en IA se ha centrado en construir máquinas eficientes para fines concretos y muy lucrativos, pero también extremadamente limitados. El coste social y científico de semejante deriva podría ser enorme.

PÁGINAS ANTERIORES: GETTY IMAGES/GREMLIN/ISTOCK

que el equipo de Toronto lograra su espectacular resultado? La razón se debe a que estas redes han de entrenarse primero con una enorme cantidad de imágenes. Y, hasta hace poco, ni había bases de imágenes lo suficientemente grandes ni existía la potencia de cómputo necesaria para poder entrenar redes multicapa en un tiempo razonable.

Dicho entrenamiento consiste en ajustar los valores numéricos correspondientes a los «pesos» de las conexiones que unen las neuronas artificiales de la red. Para ello, a la máquina se le proporciona una gran cantidad de imágenes ya etiquetadas, y un algoritmo va ajustando los valores de los pesos en función de los errores que comete la red al clasificar las imágenes de entrenamiento. Dicho algoritmo propaga el error desde la última capa hasta la anterior, de esta a la precedente, y así hasta llegar a la primera. Antes de comenzar el entrenamiento los valores asignados a las conexiones son aleatorios, y el proceso finaliza cuando los pesos alcanzan valores estables.

«ES CAPAZ DE CREAR CONOCIMIENTO POR SÍ MISMA»: Con estas palabras, reproducidas en su momento por varios medios de comunicación, describía el investigador principal del algoritmo AlphaGo los éxitos de su sucesor, AlphaGo Zero, el cual aprendió a jugar al go practicando contra sí mismo. En marzo de 2016 (*fotografía*), AlphaGo saltó a la fama al vencer a Lee Sedol, uno de los mejores jugadores de go del mundo.

Así pues, si queremos que la red aprenda a discernir entre gatos y perros, durante el entrenamiento le iremos mostrando imágenes de estos animales. La red clasificará cada una con un grado de confianza: por ejemplo, 70 por ciento «perro» y 30 por ciento «gato». Pero, si la imagen era la de un gato, la red hubiera tenido que responder en su lugar 100 por cien «gato» y 0 por ciento «perro», por lo que el algoritmo propagará hacia atrás dicho error, cambiando los pesos de las conexiones para que la próxima vez que se le muestre esa misma imagen los grados de confianza se acerquen más a los correctos. Sin embargo, es necesario mostrar un enorme número de veces cada una de las imágenes de entrenamiento para que la red neuronal converja a los valores correctos y podamos después emplearla para reconocer imágenes nuevas.

Por supuesto, todo ello requiere partir de una representación numérica de la imagen. Esto se consigue asociando un número a cada píxel, de modo que, desde el punto de vista de la máquina, una imagen no es más que una enorme matriz de números. En el caso de imágenes en color, se trata de una matriz tridimensional en la que cada dimensión corresponde a uno de los tres colores primarios (rojo, verde y azul, por ejemplo). Cada entrada puede tomar un valor entero entre 0 y 255. Así, el color negro se representa mediante (0, 0, 0), el blanco por (255, 255, 255), etcétera. Ello implica un total de $256^3 = 16.777.216$ colores. La red efectúa millones de operaciones matemáticas (básicamente, sumas y multiplicaciones de matrices) que, en el caso de redes con muchas capas, pueden llegar a los miles de millones. Esta enorme cantidad de cálculos ilustra la necesidad de disponer de la enorme potencia de cómputo mencionada antes.

Una propiedad interesante de estas redes es que, por su propia estructura, son invariantes en escala y traslación. Eso significa que no importan ni el tamaño ni la posición de los objetos presentes en la imagen. Sin embargo, no son invariantes frente a rotaciones, por lo que fallan a la hora de reconocer la misma imagen boca abajo. De hecho, la búsqueda de algoritmos capaces de identificar objetos rotados constituye hoy en día un activo campo de investigación.

Poco después del éxito del grupo de Hinton, otros equipos aumentaron el porcentaje de aciertos hasta el 98 por ciento. Eso provocó que comenzaran a aparecer titulares como este:

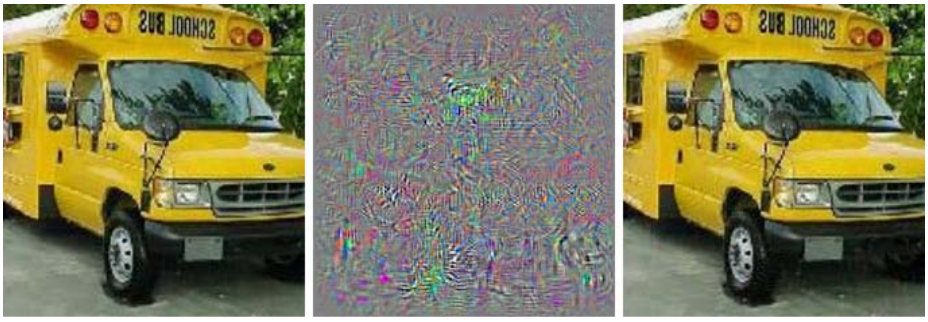
Los ordenadores superan a los humanos en el reconocimiento y clasificación de imágenes

—*The Guardian*, 13 de mayo de 2015

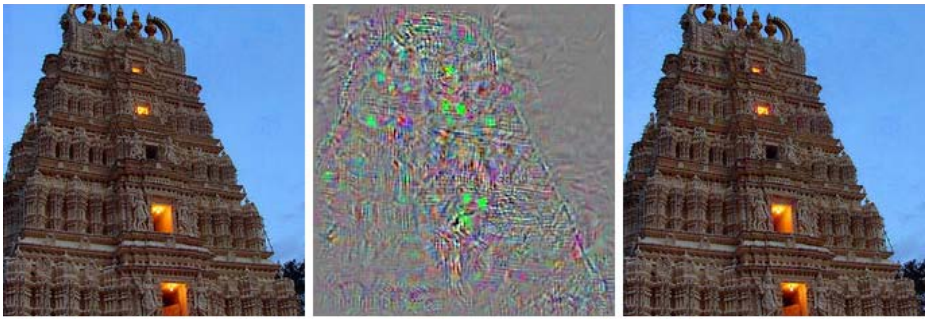
Pero vayamos por partes. ¿Seguro que las máquinas son mejores que nosotros reconociendo imágenes?

El titular citado se basaba en un estudio que afirmaba que los humanos cometen un 5 por ciento de errores al clasificar imágenes de ImageNet. Pero, si analizamos el trabajo referido, comprobaremos que a la máquina se le permitía dar una lista de cinco categorías, ordenadas de mayor a menor confianza, y se consideraba que había acertado si la categoría correcta era una de esas cinco, aunque fuera la quinta. No obstante, si solo se consideraba la primera, el error aumentaba hasta el 18 por ciento, mucho mayor que el humano. Pero, además, en el estudio había participado solo una persona, por lo que un titular más apropiado hubiera sido «Los ordenadores superan a *X* en el reconocimiento y clasificación de imágenes» donde *X* tendría que haber indicado el nombre y apellidos del individuo en cuestión. Por tanto, el titular de *The Guardian* no es cierto, aunque no se puede negar que es llamativo.

Por otro lado, a menudo se nos ha querido hacer creer que las redes neuronales artificiales aprenden por sí mismas. Sin embargo, se requiere un enorme esfuerzo por parte de los programadores para preparar una red antes de que pueda empezar a aprender. No se trata solo de etiquetar la ingente cantidad de datos necesarios para el entrenamiento, sino también de definir todo tipo de aspectos de la arquitectura de la red, lo que se conoce como «hiperparámetros». Entre estos se encuentran el número de capas intermedias, las funciones de activación de las



«Autobús escolar» Distorsión «Avestruz, *Struthio camelus*»



«Edificio» Distorsión «Avestruz, *Struthio camelus*»

CONOCIMIENTO FRÁGIL: En 2013, un trabajo demostró lo fácil que resulta engañar a los algoritmos de reconocimiento de imágenes. La modificación de unos pocos píxeles (una distorsión imperceptible para el ojo humano, *columna central*) bastaba para que una red neuronal clasificara las imágenes de un autobús escolar o de un edificio (*izquierda*) como un avestruz (*derecha*).

neuronas, así como diversos parámetros asociados al algoritmo de propagación del error. En redes complejas, el número de hiperparámetros que deben fijarse puede ser muy elevado y, además, deben cambiarse por completo para cada nueva tarea que queramos que la red aprenda. De hecho, es una destreza que a los programadores les cuesta mucho tiempo adquirir, ya que se trata básicamente de un proceso de ensayo y error. Algunos afirman que es casi un arte.

RECONOCIMIENTO DE IMÁGENES: APRENDIZAJE SUPERFICIAL

No cabe duda de que estamos muy lejos de poder resolver el problema del reconocimiento de imágenes. Las disparidades entre el aprendizaje humano y el artificial, incluido el aprendizaje profundo, son aún enormes.

La primera gran diferencia es que nosotros, para aprender a reconocer categorías, solo necesitamos haber visto unos cuantos ejemplos, no millones. Además, no aprendemos de forma pasiva, sino que interactuamos con nuestro entorno por medio de todos los sentidos; es decir, con nuestro cuerpo. «Ver» consiste en mucho más que reconocer cosas: implica extraer vínculos entre los objetos que vemos y entender cómo estos se relacionan con otros elementos que no están necesariamente presentes en la imagen. Si, además, vemos seres animados, como personas o animales, sabemos interpretar sus movimientos e incluso predecir de manera aproximada sus acciones inmediatas. Estas facultades resultan esenciales a la hora de tomar decisiones, como cuando vamos conduciendo y nos vemos obligados a frenar porque una persona está a punto de cruzar la calle.

La investigación actual está abordando estos problemas. Pero, como ha ocurrido siempre con la IA, lo que resulta más sencillo para los humanos acaba siendo extremadamente complejo para las máquinas. Esta aparente paradoja se explica porque, en los humanos, la capacidad de analizar una imagen no puede desligarse del resto de las facultades que conforman la inteligencia; en particular, de la capacidad de abstracción, de entender el lenguaje y de razonar con sentido común. Nada de eso puede aprenderse a partir de las imágenes de una base de datos, sino que exige interaccionar con el mundo real: una cuestión clave sobre la que volveremos más adelante.

No en vano, una gran mayoría de los investigadores en IA creemos que el aprendizaje supervisado no constituye el mejor camino para alcanzar la IA general. El motivo principal es que, incluso limitándonos a objetos físicos, no resulta nada realista pretender etiquetar absolutamente todas las cosas que podemos llegar a observar en el mundo, y sin etiquetas no puede haber aprendizaje supervisado. Otra dificultad guarda relación con el hecho de que, en prácticamente todos los dominios de la IA, antes o después aparece el llamado «problema de cola larga»: aunque existe un gran número de situaciones que suceden con una probabilidad elevada, a menudo hay una gran «cola» de situaciones (que, en conjunto, pueden llegar a sumar muchas más) que tienen una probabilidad muy pequeña de ocurrir. Eso provoca que tales situaciones no aparezcan casi nunca en los datos de entrenamiento, por lo que un sistema de aprendizaje supervisado errará estrepitosamente ante ellas.

Otro grave problema de estos sistemas es que, incluso cuando una red funciona correctamente y, por ejemplo, clasifica con acierto la imagen de un gato, nunca podemos estar seguros de que realmente haya detectado la presencia del animal. Tal vez haya localizado algún otro objeto que, en las imágenes de entrenamiento, aparecía con frecuencia junto a los felinos, como una pelota. Por tanto, a menudo ni siquiera los diseñadores de los sistemas de aprendizaje profundo saben con exactitud por qué la máquina funciona cuando acierta ni por qué falla cuando se equivoca. Este serio inconveniente, conocido como «problema de la caja negra», hace que sea prácticamente imposible explicar las decisiones que toman estos sistemas. Por ello, una línea de investigación muy activa hoy en día es la llamada «IA explicable», la cual persigue que las máquinas puedan explicar las decisiones que toman en un lenguaje que las personas logremos entender con facilidad. No basta con listar las operaciones matemáticas que ha efectuado la red.

Este último obstáculo afecta a la confianza que otorgamos a la IA. Es cierto que las personas tampoco podemos explicar siempre nuestras decisiones. Sin embargo, hay una diferencia fundamental: los humanos tendemos a confiar unos en otros porque creemos que los mecanismos de pensamiento de los de-

más son similares a los nuestros. Es lo que los psicólogos llaman tener una «teoría de la mente» sobre los demás. No obstante, ninguno de nosotros tiene una teoría de la mente sobre ninguna máquina, ni desde luego ninguna máquina la tiene sobre nosotros. Por ello, resulta perfectamente razonable exigir más explicaciones a una máquina que a una persona.

Ese problema de falta de confianza puede agravarse debido a lo sencillo que resulta engañar a una red neuronal. Para ello basta con modificar unos pocos píxeles en la imagen que el sistema debe reconocer; distorsiones imperceptibles para el ojo humano, pero que pueden provocar colosales errores de clasificación. Un ejemplo célebre lo hallamos en una fotografía de ImageNet que mostraba un autobús escolar. Tras ser mínimamente distorsionada, la imagen fue sorprendentemente clasificada como un avestruz.

A la vista de su enorme fragilidad y de su escasa relación con la visión y el aprendizaje humanos, resulta muy difícil entender por qué hoy en día existen tantas aplicaciones basadas en esta tecnología, como ya ocurre con el reconocimiento facial. En mi opinión, tales aplicaciones constituyen una falta absoluta de prudencia y de ética. Cuando los humanos vemos un objeto, vamos mucho más allá del objeto en sí: tenemos en cuenta el contexto en el que aparece, recordamos otras situaciones en que hemos visto objetos similares, sabemos para qué sirve, cómo se relaciona con otros elementos y con nosotros mismos, y un larguísimo etcétera imposible de enumerar. Sin todo ese «sentido común», los sistemas de visión artificial siempre serán frágiles y poco fiables.

> Hoy por hoy, no existe ningún sistema de IA capaz de contextualizar y de hacer el tipo de inferencias básicas que incluso un niño realiza sin esfuerzo

Por ello, desde el punto de vista ético habría que adoptar una actitud de enorme prudencia con respecto a la IA. En este sentido, la *Declaración de Barcelona para un desarrollo y uso adecuados de la IA en Europa*, un documento elaborado en 2017 con el concurso de varios expertos, recomienda este principio de prudencia, entre otros aspectos.

JUEGOS DE TABLERO: INCAPACIDAD DE GENERALIZAR

Otra técnica de IA que ha ganado gran popularidad en los últimos años es el aprendizaje por refuerzo. Este enfoque se ha combinado con éxito con otros, en particular con las redes convolucionales profundas, para desarrollar programas que han aprendido a jugar al *backgammon*, a distintos juegos de Atari o incluso al go, llegando a superar a los mejores jugadores humanos. Esta combinación, conocida como «aprendizaje profundo por refuerzo» también ha contribuido de manera significativa a la reciente fiebre de la IA.

En este tipo de aprendizaje, un agente aprende a partir de las consecuencias de las acciones que ejecuta, ya sea sobre la base de su experiencia previa —si la tiene— o de una selección aleatoria de las acciones que puede tomar en cada situación, o «estado». El agente recibe un valor numérico (refuerzo) que codifica el éxito o el fracaso, y su objetivo consiste en seleccionar aquellas acciones que maximicen el refuerzo acumulado. En este caso no se trata de un aprendizaje supervisado, pues no se proporcionan ejemplos de pares estado-acción.

Un ejemplo de este proceso nos lo proporciona un ratón que debe aprender a recorrer un laberinto a cuya salida hay un trozo de queso, el cual servirá como recompensa. El punto de partida constituye el estado inicial del problema, la salida corresponde al estado final, y las situaciones entre uno y otro (cruces de pasillos y callejones sin salida) a los estados intermedios del problema. Al llegar a una encrucijada, el ratón ha de tomar la decisión de qué ruta seguir. Dado que al principio el animal no conoce el resultado de las posibles acciones (tomar un camino u otro al llegar a una cruce, por ejemplo), la manera de hallar la salida consiste en aprender el camino; es decir, a asociar acciones con estados. El animal comenzará con una estrategia de ensayo y error que le hará recorrer el laberinto de forma aleatoria. Cuando llegue a la salida y encuentre la recompensa, la acción que le condujo hasta allí desde el penúltimo estado recibirá un valor de refuerzo elevado. Este se propagará a su vez hasta la acción que le condujo desde el antepenúltimo estado hasta el penúltimo, y así sucesivamente.

Tras un elevado número de intentos, el algoritmo de aprendizaje por refuerzo converge a unos valores que siempre conducen al resultado deseado de manera eficiente. Las variaciones y extensiones de estos algoritmos son múltiples. Una de ellas fue la que sirvió de base a la máquina AlphaGo Zero, que en 2017 aprendió a jugar al go practicando contra sí misma y acabó alcanzando el máximo nivel tras jugar millones de partidas. En este caso, los estados corresponden a las posiciones de las piedras en el tablero, y las acciones son los movimientos reglamentarios asociados a cada una.

Una extensión importante de estos algoritmos es la que permite que el agente solo posea un conocimiento imperfecto del estado en el que se encuentra. Esto no ocurre en los juegos de tablero, donde la situación completa es perfectamente observable, pero sí constituye un problema en otros casos, como el de un robot móvil que ha de aprender a desempeñar tareas en un entorno físico. Por regla general, los sistemas de percepción no serán lo suficientemente precisos para determinar con total seguridad en qué posición se encuentra el robot ni dónde se hallan todos los objetos de su alrededor. Otra dificultad del aprendizaje por refuerzo es el compromiso entre aprovechar lo aprendido o seguir explorando —es decir, aprendiendo— con la esperanza de encontrar una estrategia mejor. Por último, el mayor problema de esta técnica reside en su escalabilidad. Cuando el número de estados y acciones posibles es muy elevado, el aprendizaje resulta extremadamente lento y hace falta extender los algoritmos con el concurso de otras técnicas.

Los juegos constituyen un excelente campo de aplicación para poner a prueba los algoritmos de aprendizaje. Pero ¿pueden ampliarse estas técnicas más allá de dicho ámbito? En otras palabras, ¿son realmente generales estos algoritmos? La verdad es que no mucho. AlphaZero, por ejemplo, una versión extendida de AlphaGo Zero que, además de al go, aprendió a jugar al ajedrez y al shogi (el «ajedrez japonés»), requería una red convolucional separada para cada uno de esos juegos y tuvo que ser entrenada desde cero para cada uno de ellos. Es decir, fue incapaz de aprovechar lo que había aprendido en un juego para transferirlo a otro; ni siquiera entre el shogi y el ajedrez, a pesar de su similitud. Esta imposibilidad de generalizar supone un elemento más que añadir a la larga lista de diferencias entre el aprendizaje artificial y el humano.

La facultad de transferir a una nueva tarea lo aprendido previamente en otra constituye un aspecto esencial del aprendizaje humano, así como de nuestra capacidad de generalizar y de razonar. Hoy, el aprendizaje por transferencia constituye un área de investigación muy activa en IA, ya que somos muchos quienes creemos que se trata de un paso importante hacia la IA general. Por ejemplo, en un trabajo cuyos resultados se publicaron en 2015, nuestro grupo de investigación enseñó a un robot a mantener en equilibrio un doble péndulo invertido. Después, un sistema de aprendizaje por transferencia trasladó ese conocimiento a un segundo robot que debía aprender a caminar. Aunque esta última máquina también podía aprender a andar sin saber cómo equilibrar el péndulo, disponer de tales conocimientos aceleró de manera significativa el proceso. Con todo, los resultados en esta área de investigación son todavía incipientes.

También aquí los diseñadores han de invertir enormes esfuerzos. Además, hablamos de inteligencias artificiales específicas, que operan en entornos limitados y que solo ejecutan tareas muy bien definidas, lo que las inhabilita para llevar a cabo acciones que nosotros ejecutamos sin apenas esfuerzo. A modo de ejemplo, pensemos en un robot doméstico que tuviese que aprender a cargar un lavavajillas. La variedad de objetos que tendría que reconocer, incluso solo en la cocina, es enorme. Y no todos irían al lavavajillas, ya que además de platos, vasos o cubiertos, en la mesa puede haber servilletas, teléfonos móviles, periódicos o restos orgánicos. También puede haber enseres que sí deban ir al lavavajillas pero que no estén encima de la mesa, sino apilados en el fregadero, por lo que no todos serán visibles. Como consecuencia, resulta prácticamente imposible enumerar a priori todos los estados en que puede hallarse el sistema o las diferentes acciones que habría de ejecutar en cada uno. De hecho, incluso predecir la fragilidad de un objeto solo con verlo constituye un problema aún sin resolver, por lo que la manipulación de todos esos utensilios tendría que hacerse con mucho tiento si no queremos tener que comprar miles de piezas de vajilla para el entrenamiento.

Todas estas dificultades impiden que un robot pueda aprender semejante tarea en una cocina real. En su lugar, debería hacerlo en una cocina simulada con realidad virtual. Pero una vez más, un simulador de un entorno tan complejo como una cocina debería tener en cuenta las propiedades físicas de cada objeto, cómo se comporta en cada situación posible según la ley de la gravedad, así como contemplar situaciones inesperadas, como que un niño entre corriendo mientras el robot lleva a cabo su tarea. Hoy por hoy, absolutamente nadie en el campo de la IA sabe cómo resolver estos problemas, ni siquiera en el caso de un entorno tan limitado como una cocina.

PROCESAMIENTO DEL LENGUAJE; NI CONTEXTO NI SEMÁNTICA

Por último, otro ámbito de la IA conocido por sus espectaculares progresos es el del procesamiento del lenguaje natural. El uso de las llamadas «redes neuronales recurrentes profundas» ha permitido grandes avances tanto en traducción automática como en asistentes personales capaces de responder a preguntas formuladas por el usuario.

Al contrario que las redes convolucionales, cuya matriz de entrada tiene un tamaño prefijado determinado por el número de píxeles, estas redes pueden procesar secuencias de datos de longitud variable, que, en el caso del lenguaje, corresponden a secuencias ordenadas de palabras. Las neuronas de cada una de sus capas no solo están conectadas a las capas anterior y posterior, sino también entre ellas. En un instante dado, una neurona calcula su grado de activación dependiendo de las señales que recibe de la capa anterior, pero también de las que en el instante previo le proporcionaron otras neuronas de su misma capa. Esto añade una cierta «memoria», gracias a la cual la red puede relacionar cada palabra con la precedente.

Al igual que ocurría con el caso de las imágenes, también el procesamiento automático del lenguaje requiere convertir primero las palabras en números. Esto se logra transformándolas en vectores de números reales, de tal manera que dos vectores asociados a palabras que estadísticamente aparezcan juntas en muchos textos estén también cercanos entre sí (dado que se trata de vectores numéricos, siempre es posible definir una distancia entre ellos). Dicho espacio de vectores es multidimensional, ya que una misma palabra puede tener diferentes significados en función del contexto. Por ejemplo, el vector asociado a la palabra *banco* estará cerca del correspondiente a *asiento* en una de las dimensiones, pero cerca del asignado a *caja de ahorros* en otra. Sin embargo, los vectores asociados a *asiento* y a *caja de ahorros* estarán alejados entre sí.

Este tipo de representación ha permitido mejorar las prestaciones de los sistemas de procesamiento del lenguaje natural hasta el punto de dar lugar a declaraciones como estas:

El nuevo servicio de Google traduce idiomas casi tan bien como las personas

—*MIT Technology Review*, 27 de septiembre 2016

Watson, de IBM, ya habla con fluidez nueve idiomas (y subiendo)

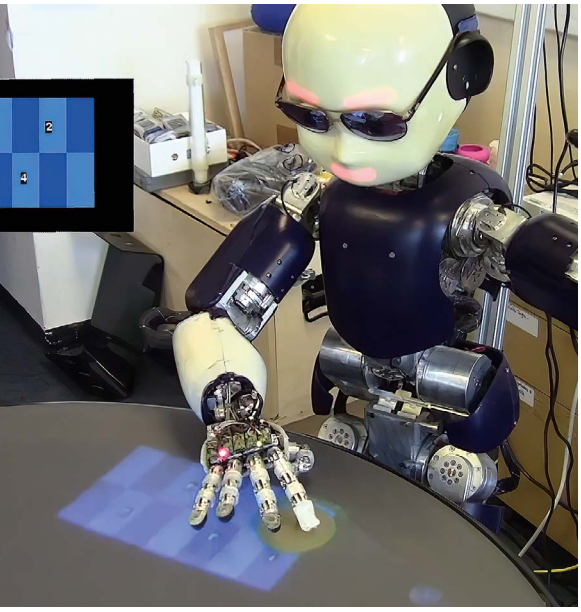
—*Wired*, 6 de octubre 2016

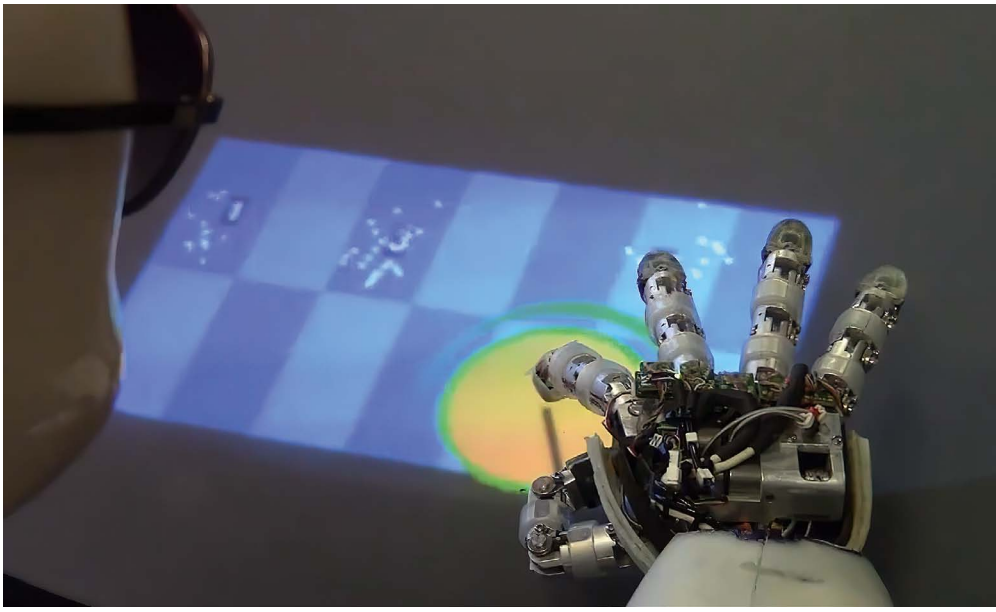
«Nuestras redes neuronales han desarrollado un asombroso sentido de la comprensión»

—Gereon Frahling, presidente de la compañía de traducción automática DeepL, 20 de marzo 2018

Estas afirmaciones responden a algunos estudios comparativos con traducciones humanas. Dado que un mismo texto puede admitir varias traducciones, todas ellas correctas, se pidió a un grupo de personas bilingües que evaluaran la calidad de cientos de frases traducidas tanto automáticamente como por traductores humanos profesionales. A continuación, calcularon la media aritmética de las notas con que cada evaluador había calificado los dos tipos de traducciones, y calcularon una media global, que era la media de las medias de cada evaluador, tanto para las traducciones automáticas como para las humanas.

La primera observación es que todos sabemos que las medias son engañosas: si la mitad de las traducciones son malas y la otra mitad muy buenas, el resultado global será que las traducciones son bastante buenas. Sin embargo, sería claramente preferible un sistema que proporcionara siempre traducciones bastante buenas y que no hiciera nunca traducciones pésimas. Otra crítica es que se trataba de traducir frases aisladas, no párrafos extensos, donde el análisis del discurso desempeña un papel fundamental para traducir correctamente. También hay que añadir que el corpus usado estaba formado por frases extraídas de noticias y de entradas en Wikipedia, las cuales suelen estar escritas evitando al máximo las ambigüedades.





¿MENTE SIN CUERPO? Un número creciente de investigadores consideran que el objetivo de una inteligencia artificial general solo será posible en máquinas dotadas de un cuerpo que les permita interaccionar con el entorno, con el fin de aprender a partir de dichas interacciones. En 2015, un grupo de investigación del Instituto de Investigación en Inteligencia Artificial de Barcelona demostró que un robot iCub (*imagen*) podía aprender a relacionar posiciones de dedos (causas) con sonidos de notas (efectos) interactuando con un teclado musical virtual. Una vez aprendidas dichas relaciones, el robot fue capaz de reproducir secuencias de notas de forma robusta.

Pero incluso cuando un sistema como Google Translate nos proporciona una traducción estupenda, no lo hace mediante un análisis semántico profundo. De forma similar a lo que ocurre con la clasificación de imágenes, lo consigue a partir de cálculos matriciales gracias a haberse entrenado con un corpus de millones de frases traducidas correctamente. Si hubiera comprensión semántica, estos sistemas deberían poder inferir una gran cantidad de relaciones que los facultarían para responder preguntas incluso cuando la respuesta no aparece explícitamente en el texto. Para entenderlo, consideremos el siguiente fragmento:

Juan no tenía dinero en efectivo. Era viernes. Después de cenar tomó la tarjeta y fue al único cajero que hay en su pueblo. Introdujo la tarjeta, pero el cajero no funcionaba y se la tragó. Mientras caminaba hacia su casa, se encontró con un amigo y fueron al bar del pueblo a tomar una cerveza. Cayó más de una, y cuando volvió a casa estaba bastante alegre.

Cualquier persona reconocerá fácilmente los siguientes hechos no explicitados en el texto: se trata de una tarjeta bancaria; un cajero es una máquina que dispensa dinero; Juan no consiguió sacar efectivo; debe vivir en un pueblo muy pequeño; su amigo pagó las cervezas o, tal vez, al ser un pueblo muy pequeño los dueños del bar le conocían y le fiaron; los bares ofrecen bebidas a cambio de dinero; la cerveza es una bebida alcohólica; el alcohol puede producir euforia en las personas; bebió un poco más de lo debido; volvió a casa bien entrada la noche y sin la tarjeta; no la pudo recuperar antes del lunes; etcétera. Por tanto, cualquiera de nosotros podría responder numerosas preguntas relativas a, entre otras cosas, la tarjeta, el cajero, el pueblo, el dinero o los efectos del alcohol.

Hoy por hoy no existe ningún sistema de procesamiento automático del lenguaje capaz de hacer estas inferencias. Para que una máquina pueda deducir todo lo anterior necesitaría disponer de una enorme cantidad de conocimientos de sentido común que nosotros adquirimos casi sin esfuerzo a lo largo de la vida. Así pues, no resulta sorprendente que la comprensión profunda del lenguaje natural constituya uno de los mayores desafíos a los que se enfrenta la IA. El lenguaje no solo es ambiguo y dependiente del contexto, sino que presupone una gran cantidad de conocimientos generales. Incluso en el reconocimiento del habla, el ámbito en el que posiblemente haya habido más progresos, la falta de comprensión es manifiesta. Es conocido el caso del usuario de Siri que le dijo: «Siri, apunta lo siguiente en la lista de la compra». A lo que Siri respondió: «“Lo siguiente”, apuntado en la lista de compra».

Es cierto que la transcripción de voz a texto sí funciona bien, pues hoy estas aplicaciones llegan a transcribir correctamente hasta el 95 por ciento de las palabras pronunciadas. No obstante, ello se debe a que para identificar fonemas no es necesario comprender el significado de las palabras. Y tales porcentajes se logran solo en entornos con muy poco ruido de fondo. Con bullicio, la tasa de acierto cae de manera drástica.

Así pues, y a pesar de los éxitos del aprendizaje profundo aplicado al procesamiento del lenguaje, vemos que, contrariamente a lo que ha llegado a afirmarse, seguimos estando muy lejos del nivel humano. La razón de dichas exageraciones seguramente obedece a la feroz competencia entre empresas para hacerse con la parte más grande de un pastel, el de la traducción y los asistentes personales, extremadamente lucrativo. Y aunque aún falte mucho para lograr traducciones automáticas de calidad similar a las de un humano profesional, no cabe duda de que una herramienta como Google Translate resulta muy útil si no somos muy exigentes con el resultado; por ejemplo, cuando nos basta con tener una idea aproximada del contenido de un texto redactado en un idioma que ignoramos por completo. Pero si

INSTITUTO DE INVESTIGACIÓN EN INTELIGENCIA ARTIFICIAL/CSIC

somos un poco más estrictos, siempre es necesario un trabajo de edición posterior para corregir los errores que inevitablemente cometen los traductores automáticos y que tan a menudo nos hacen sonreír.

Conversar con máquinas de forma robusta en lenguaje natural y sobre una amplia diversidad de temas sigue siendo, por tanto, una meta muy lejana. Y no parece que vaya a llegar de la mano de técnicas que se basen exclusivamente en analizar enormes cantidades de datos sin una comprensión real del lenguaje; en otras palabras: sin atacar el problema de cómo dotar a las máquinas de sentido común.

EL PROBLEMA DEL SENTIDO COMÚN

Dediquemos unas líneas a explicar a qué nos referimos cuando hablamos de sentido común. Esta clase de conocimiento se adquiere en las primeras etapas de nuestra vida. Por ejemplo, cualquier niño pequeño sabe que, para mover un tren de juguete atado a una cuerda, hay que tirar de la cuerda, no empujarla. O también que un objeto inanimado no cambiará de posición a menos que alguien lo mueva directamente. Para adquirir tales conocimientos hay que entender entre otras cosas las relaciones de causa y efecto, así como ser capaces de razonar sobre ellas. Los niños aprenden también muy pronto que hay situaciones que enojan a sus padres, lo que implica que tienen un modelo mental de los demás y que pueden razonar sobre dichos modelos mentales.

Judea Pearl y Adnan Darwiche, expertos en IA de la Universidad de California en Los Ángeles, han subrayado que las técnicas actuales de aprendizaje profundo detectan correlaciones, pero no relaciones de causa y efecto. Por ejemplo, no pueden aprender que es la salida del sol lo que provoca el canto del gallo, y no al revés. Estos investigadores han argumentado a favor de cierto tipo de modelos para integrar en las máquinas las relaciones de causa y efecto. Se trata de un enfoque de enorme interés, puesto que combina razonamiento y aprendizaje; un paso que, en mi opinión, resultará imprescindible para progresar hacia la IA general. En los últimos años se han propuesto otras técnicas para integrar razonamiento y aprendizaje, como las que plantean añadir memoria y capacidad de razonamiento a las redes neuronales. Estas tendencias son una buena noticia, ya que parecía que la comunidad de IA había renunciado a responder a una pregunta clave en toda actividad científica: la pregunta del porqué.

Con el éxito del aprendizaje profundo, muchos investigadores e ingenieros parecían haber olvidado que, en última instancia, lo importante es preguntarse por qué algo funciona, no conformarse con el hecho de que funcione. Si la IA específica ya permite desarrollar sistemas de reconocimiento visual, traducción automática y asistentes personales, que, además, son extremadamente lucrativos, ¿por qué deberíamos preocuparnos por la IA general? La respuesta es doble. Por un lado, por el fin puramente científico de entender de una vez por todas qué es la inteligencia. Por otro, por el hecho práctico de que la IA general nos permitirá desarrollar aplicaciones mucho mejores y más útiles.

Las máquinas «inteligentes» actuales constituyen un ejemplo de lo que el filósofo de la mente Daniel Dennet llama «habilidad sin comprensión». Y, de hecho, creo que esta podría ser incluso una buena definición de lo que es la propia IA hoy en día: «El área de la computación que consiste en dotar a las máquinas de habilidades sin capacidad de comprender». Otro filósofo, John Searle, señaló hace cuarenta años la imposibilidad de que los sistemas de IA basados en la mera manipulación sintáctica de símbolos lleguen a entender nada. Su argumento, hoy famoso, se conoce con el nombre de la «habitación china». El propio Searle lo resumió en estos términos:

Supongamos que un angloparlante que no tiene ni idea de chino se encierra en una habitación en la que dispone de un conjunto muy completo de reglas, escritas en inglés, sobre cómo manipular caracteres chinos y cómo generar otros a partir de tales manipulaciones. A continuación, desde el exterior se le proporcionan una serie de caracteres en ese idioma y él, aplicando las reglas mencionadas, procede a transformarlos en otros caracteres chinos que devuelve al exterior, de manera que estos resulten ser respuestas a los caracteres de entrada indistinguibles de las que daría alguien que habla chino con fluidez.

> La IA actual podría definirse como «la disciplina que consiste en dotar a las máquinas de habilidades sin capacidad de comprender»

Searle afirma que esta «habitación china» (el sistema formado por la persona que no habla chino, por las reglas de transformación de símbolos y por los símbolos objeto de manipulación) no entiende chino, puesto que lo único que está haciendo es una manipulación puramente sintáctica; es decir, sin consideraciones semánticas. Searle fue, y todavía es, muy criticado por la comunidad de IA. Pero, en mi opinión, el tiempo le ha dado la razón y hoy podemos decir que, efectivamente, la habitación china no entiende el chino, de igual modo que Google Translate no entiende los textos que traduce.

CUERPO E INTELIGENCIA GENERAL

Con anterioridad a los éxitos de las redes neuronales profundas, el modelo dominante en IA había sido el simbólico. Este tiene sus raíces en la hipótesis conocida como «sistema de símbolos físicos» (SSF), formulada en 1975 por Allen Newell y Herbert Simon en su ponencia de recepción del premio Turing, considerado el galardón de mayor prestigio en las ciencias de la computación.

Por «sistema de símbolos físicos» Newell y Simon se referían a un conjunto de entidades (símbolos) que, mediante una serie de reglas, pueden combinarse para formar estructuras mayores y transformarse en otras. Tales procedimientos permiten crear símbolos nuevos, generar y modificar las relaciones entre ellos, almacenarlos, determinar si dos símbolos son iguales o no, etcétera. Tales símbolos son además físicos, en el sentido de que tienen un sustrato material (electrónico, en el caso de los ordenadores, o químico-biológico, en el de los humanos). Según la hipótesis de Newell y Simon, todo sistema de símbolos físicos posee los medios necesarios y suficientes para llevar a cabo acciones inteligentes. Por otra parte, dado que los humanos somos capaces de mostrar conductas inteligentes, también nosotros deberíamos ser sistemas de símbolos físicos. Por último, la naturaleza del sustrato (electrónica o biológica) carece de importancia: un sistema puede ser inteligente siempre y cuando dicho sustrato le permita procesar símbolos.

No olvidemos que se trata de una hipótesis, por lo que no debe ser aceptada ni rechazada a priori, sino evaluada de acuerdo con el método científico. También es importante matizar que en ningún momento Newell y Simon limitaron su hipótesis a que el procesamiento de símbolos debiera ser únicamente sintáctico, por lo que su validez no contradiría el argumento de Searle de que la habitación no entiende chino. En mi opinión, el objetivo científico de la IA general no es otro que intentar demostrar esta hipótesis en el contexto de los ordenadores. Es decir, averiguar si una computadora convenientemente programada es capaz de mostrar una conducta inteligente de tipo general.

El modelo simbólico sigue siendo muy importante hoy en día, y de hecho se considera el modelo «clásico» en IA. Puede calificarse como un modelo que opera «de arriba abajo» (*top-down*), puesto que trabaja con representaciones abstractas del mundo, las cuales se procesan mediante lenguajes basados principalmente en la lógica matemática y sus extensiones. Las redes neuronales, en cambio, constituyen una modelización «de abajo arriba» (*bottom-up*), basada en la hipótesis de que la inteligencia emerge a partir de la actividad distribuida de un gran número de unidades interconectadas (las neuronas artificiales). Estos sistemas no son incompatibles con la hipótesis del SSF, ya que, al fin y al cabo, también procesan símbolos. Sin embargo, estos no son explícitos, sino que se encuentran repartidos por toda la red. Por esta razón, la IA simbólica permite explicar con mayor facilidad el funcionamiento de las máquinas, ya que, al contar con símbolos explícitos, es posible analizar cómo estos intervienen en el proceso de razonamiento, algo irrealizable con las redes neuronales actuales.

Con todo, ni la IA simbólica ni la basada en redes neuronales requieren que el sistema disponga de un cuerpo situado en un entorno real. Una de las críticas más duras a estos modelos «no corpóreos» se basa en el hecho de que, para muchos investigadores, un agente inteligente necesita un cuerpo que le permita tener experiencias directas con el entorno. No basta con que un programador le proporcione descripciones abstractas de ese entorno codificadas en un lenguaje de representación (como en el caso de la IA simbólica) o millones de datos de entrenamiento (como ocurre con las redes neuronales artificiales).

En mi opinión, en ausencia de un cuerpo, ni las representaciones abstractas de la IA simbólica ni el estado interno de una red neuronal artificial podrán adquirir contenido semántico para una máquina. En cambio, una interacción directa con el entorno permitiría que un agente con cuerpo, como un robot, relacionase las señales que captan sus sensores con las representaciones simbólicas generadas a partir de lo percibido con anterioridad. Hubert Dreyfus, filósofo de la Universidad de California en Berkeley, fue uno de los primeros en abogar por la necesidad de asociar la inteligencia a un cuerpo capaz de interactuar con el mundo. La idea es que la inteligencia de los seres vivos deriva del hecho de estar situados en un ambiente con el que pueden interaccionar. Según Dreyfus, la IA debería modelizar todos estos aspectos para alcanzar el objetivo de la IA general. Sin duda, se trata de una idea interesante que hoy en día compartimos cada vez más investigadores.

REPENSAR EL FUTURO

No puedo imaginar qué avances harán falta para construir una IA general corpórea, ni si dicho objetivo será nunca posible. Sin embargo, no deberíamos renunciar a él. Al fin y al cabo, son las grandes preguntas las que dan a la ciencia su razón de ser. Y la pregunta de si es posible lograr una IA general es de complejidad equiparable a otras grandes preguntas de la ciencia, como la relativa al origen de la vida o al inicio del universo.

¿Qué ocurrirá si algún día logramos la IA general? Stuart Russell, de la Universidad de California en Berkeley, ha argumentado en su reciente libro *Human compatible* que, en tal caso, será de vital importancia asegurar que las máquinas persigan los objetivos que realmente queremos, en lugar de ejecutar de manera inflexible comportamientos preprogramados claros pero incorrectos. Para ello, deberán aprender a entender nuestros deseos observando las decisiones que tomamos y comprendiendo por qué las tomamos. Según Russell, la solución pasará por introducir un margen de duda (incertidumbre) en las especificaciones de las máquinas. De esta forma, se verían obligadas a pedir aclaraciones a los humanos antes de tomar sus decisiones y, llegado el caso, permitirían que las desconectásemos.

Volviendo al presente, es innegable que los sistemas de IA actuales siguen sin comprender absolutamente nada. Al igual que el emperador del cuento de Andersen, hoy la IA continúa desnuda. Por tanto, deberíamos replantearnos con la máxima seriedad algunas de las aplicaciones que tan alegremente estamos desplegando. Me refiero al uso de algoritmos de reconocimiento facial o a los que pretenden predecir el futuro, como los que evalúan la probabilidad de reincidencia de un criminal para que un juez decida si concederle o no la libertad condicional. O lo que sería aún peor: aplicar la IA para desarrollar armas letales autónomas. ¿Somos conscientes de que estamos dejando decisiones clave en manos de artefactos estúpidos?

Lo que debería aterrorizarnos no es un futuro dominado por una hipotética IA superior. Dejemos esto para quienes creen en la singularidad y confunden la ciencia con la ciencia ficción. Lo que realmente debería preocuparnos es la situación presente, en la que estamos delegando cada vez más tareas en una IA tan limitada como la actual. Por ello, es necesario regular lo antes posible el desarrollo y uso de la IA. De lo contrario, acabaremos pagando un precio excesivamente alto.

PARA SABER MÁS

What computers still can't do: A critique of artificial reason. Hubert L. Dreyfus. MIT Press, 1992.

From bacteria to Bach and back: The evolution of minds. Daniel C. Dennet. Penguin Random House, 2018.

The Barcelona declaration for the proper development and usage of artificial intelligence in Europe. Luc Steels y Ramon López de Mántaras en *AI Communications*, vol. 31, págs. 485–494, diciembre de 2018.

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Cynthia Rudin en *Nature Machine Intelligence*, vol. 1, págs. 206-215, mayo de 2019.

Human compatible: Artificial intelligence and the problem of control. Stuart J. Russell. Viking, 2019.

EN NUESTRO ARCHIVO

¿Es la mente un programa informático? John R. Searle en *IyC*, marzo de 1990. Reeditado para *La ciencia después de Alan Turing*, colección *Temas* de *IyC*, n.º 68, 2012.

El valor de la experiencia para los robots. Ramon López de Mántaras en *IyC*, agosto de 2016.

¿Hemos de temer a los robots superinteligentes? Stuart Russell en *IyC*, agosto de 2016.

A favor de los robots desobedientes. Gordon Briggs y Matthias Scheutz en *IyC*, marzo de 2017.

El problema de la caja negra. Davide Castelvecchi en *IyC*, abril de 2017.

Ética en la inteligencia artificial. Ramon López de Mántaras en *IyC*, agosto de 2017.

La IA, en manos privadas. Yochai Benkler en *IyC*, octubre de 2019.