Cuatro problemas irresolubles de la IA simbólica

(Four unsolvable problems of symbolic AI)

Manuel Carabantes López

Recibido: 14 de enero de 2014
Aceptado: 9 de julio de 2014

Resumen

Dentro de la corriente fuerte de la inteligencia artificial (IA), que es la que pretende crear máquinas pensantes con competencias intelectuales parecidas a las del ser humano, el programa de investigación más explorado es el de la IA simbólica, que se define como el intento de utilizar computadoras electrónicas para duplicar la mente humana, bien suponiendo una semejanza estructural y funcional entre ambas, o bien intentando duplicar la conducta producida por la mente humana mediante procesos computacionales con una estructura también intencional pero sólo instrumentalmente equivalentes. En el presente artículo mostraremos que la IA simbólica fuerte, en cualquiera de sus dos mencionadas variantes, es imposible, ya que los sistemas formales, que es lo que en el fondo son todos los programas informáticos, no son suficientes para duplicar de manera intencional cuatro facultades intelectuales del ser humano imprescindibles para la conducta inteligente: sentido de la situación, sentido común, habilidades procedimentales y abducción de teorías.

Palabras clave: Inteligencia artificial, IA simbólica, problema de la pertinencia, problema de la cualificación, habilidades procedimentales, abducción.

Abstract

Within the strong branch of artificial intelligence (AI), which is aimed at creating thinking machines with intellectual powers like those of man, the most explored research program is symbolic AI, defined as the attempt to use electronic comput-

ers to replicate the human mind, either assuming a structural and functional similarity between them, or trying to replicate the behavior produced by the human mind through computational processes that also have an intentional structure but are only instrumentally equivalent. In this paper we show that strong symbolic AI, in either of its two variants, is impossible, since formal systems, that is what all computer programs ultimately are, are not sufficient to replicate in an intentional way four intellectual faculties of man that are essential to intelligent behavior: sense of the situation, common sense, procedural skills and theory abduction.

Keywords: artificial intelligence, symbolic AI, pertinence problem, qualification problem, procedural skills, abduction.

1. El deseo de algo imposible

El imaginario popular de nuestro tiempo está poblado de inteligencias artificiales: máquinas pensantes de las que se espera que por su condición de mercancías sirvan a sus propietarios, y por su condición de inteligentes, que lo hagan con unas destrezas intelectuales parecidas a las de un ser humano. La ciencia ficción, la mercadotecnia y algunos científicos implicados en su investigación llevan décadas prometiendo que las inteligencias artificiales serán una realidad muy pronto. Por supuesto, el mito de la reproducción artificial del hombre por el hombre no es nuevo, sino que viene de muy atrás (Robinet, 1973, p. 27), pero en nuestra época tiene una presencia particularmente procelosa debido a la disponibilidad de unas máquinas, las computadoras electrónicas, que, desde su invención a mediados del siglo XX, se han presentado como el soporte perfecto para duplicar el intelecto humano.

La duplicación de nuestro intelecto mediante computadoras electrónicas puede plantearse de dos maneras: creando modelos informáticos de la mente o del cerebro. Dentro de la inteligencia artificial (IA), la *IA simbólica* es el programa de investigación centrado en la duplicación de la *mente*, mientras que del *cerebro* se ocupa la *IA subsimbólica*. Por supuesto, esta nomenclatura no es la única posible. Así, por ejemplo, a la IA simbólica también se la conoce por otros nombres, tales como el de *IA del lenguaje del pensamiento* (Haugeland, 1996, p. 16). Una segunda distinción pertinente para nuestros intereses es la que refiere al alcance de la duplicación (Searle, 1980, p. 282). Así se diferencia, por un lado, a la *IA débil*, que realiza simulaciones informáticas de ciertos procesos mentales o cerebrales con el propósito de que sirvan para refutar o corroborar hipótesis y, en general, hacer avanzar el conocimiento científico; mientras que la *IA fuerte* tiene un objetivo mucho más ambicioso, ya que no se conforma con simular ciertos procesos, sino

que aspira a la creación de un verdadero autómata del pensamiento semejante al ser humano. Adicionalmente, se puede distinguir un tercer enfoque, el de la *IA aplicada*, cuyo interés es crear sistemas informáticos que, si bien no son tan inteligentes como un ser humano, sí que reproducen ciertas capacidades intelectuales con aplicaciones comerciales útiles, tales como el reconocimiento de patrones faciales para videovigilancia.

Sobre la base de estas definiciones, podemos declarar ya que el propósito del presente artículo es demostrar la imposibilidad de la *IA simbólica fuerte*. En contra de la creencia popular, la cual por cierto también es compartida por buena parte de los científicos de la IA, sostenemos que la mente por sí sola no puede ser reproducida mediante una computadora electrónica. Demostraremos esta tesis a través de la exposición de cuatro facultades esenciales del intelecto humano y de las respectivas razones particulares por las cuales no han podido ni podrán ser jamás reproducidas por una computadora electrónica de manera exclusivamente simbólica. Estas facultades son: sentido de la situación, sentido común, habilidades procedimentales y abducción de teorías. La causa, en general, por la que no pueden ser reproducidas por una supuesta computadora que duplique la mente humana, ya sea de manera realista o instrumental, es doble. Primero, porque si la duplicación se intenta de manera realista, es imposible debido a que se trata de facultades que, en el caso de los seres humanos, implican procesos exclusivamente cerebrales, es decir, de los que no emerge ningún correlato mental. Y segundo, porque si la duplicación se pretende de manera instrumental, esto es, creando una mente que, aunque no opere como la nuestra, sea conductualmente indistinguible, es imposible debido a que el único modelo de la mente con el que pueden operar las computadoras, que es el de un sistema formal, no es suficiente para reemplazar a los citados procesos cerebrales.

En otras palabras, la IA simbólica fuerte es imposible porque comete el error de creer que la mente, entendida como una especie de lenguaje del pensamiento sujeto a reglas de formación y transformación de las expresiones, es suficiente para producir la conducta inteligente. Para la filosofía, lo más interesante del caso es que no se trata de un error accidental, sino que es una concreción sintomática de la obsesión logocéntrica de nuestra cultura. Creer que la mente es calculable como el universo físico y creer que ese cálculo es suficiente para la inteligencia son dos ideas con precedentes en filósofos como Platón, Descartes y Leibniz (Dreyfus, 1992, p. 67). La finalidad de este artículo es, como decimos, demostrar la imposibilidad de la IA simbólica fuerte, una tarea que nos conducirá en última instancia al descubrimiento de las razones por las cuales algo que es imposible lleva proclamándose como posible durante el último medio siglo. Y no sólo como posible: sino como un avance de la técnica que no tardará en hacerse realidad para automatizar el trabajo intelectual que las computadoras electrónicas todavía no han podido asumir.

2. Características de las computadoras electrónicas

Dejando a un lado las características materiales derivadas de su condición de aparatos electrónicos, por ser irrelevantes respecto a la esencia de su funcionamiento como bien señaló Alan Turing al afirmar su equivalencia con la Máquina Analítica de Charles Babbage (Turing, 1950, p. 446), las computadoras lo único que hacen es ejecutar programas informáticos, y los programas informáticos no son más que sistemas formales (Weizenbaum, 1976, p. 92). Son sistemas formales de una gran complejidad, es cierto, de millones de líneas de código diseñadas por decenas de ingenieros en algunos casos, pero sistemas formales al fin y al cabo. Un *sistema formal* se compone de dos partes: lenguaje formal y mecanismo deductivo (Falguera & Martínez, 1999, p. 62). Un *lenguaje formal* es un lenguaje exento de interpretación semántica en su definición (Ibíd., p. 61), y se divide en un vocabulario y un conjunto de reglas que establecen cómo han de combinarse los símbolos del vocabulario para formar expresiones correctas. En cuanto al *mecanismo deductivo*, consiste en esencia en un conjunto de reglas que indican cómo transformar unas expresiones en otras.

El filósofo norteamericano John Haugeland ofrece otra definición de sistema formal menos académica, pero que quizás ayuda mejor a comprender qué es lo que hacen las computadoras. Un sistema formal, dice Haugeland, «es como un juego en el que las fichas (*token*, en el original, significa ficha, símbolo, señal) son manipuladas de acuerdo a reglas para ver qué configuraciones pueden ser obtenidas» (Haugeland, 1981, p. 5). Las configuraciones son expresiones cuasi-lingüísticas, como por ejemplo sucede en el ajedrez, un juego en el que todo cuanto sucede en el espacio del tablero está siendo procesado por la máquina jugadora de una manera no analógica, esto es, no haciéndose una representación interna homomórfica en sentido etimológico de ese objeto espacial, que es lo que hacemos nosotros, sino en forma proposicional, utilizando una notación, la de un lenguaje formal de programación, que es parecida a la algebraica que emplean los aficionados a este juego para ir tomando nota de lo que sucede a lo largo de la partida. Este punto es fundamental para lo que nos proponemos demostrar: las computadoras son sólo sistemas formales, y como tales lo único que hacen es manipular expresiones cuasi-lingüísticas de acuerdo a reglas.

3. El supuesto ontológico

Ciertamente, tanto a nivel simbólico como subsimbólico, los programas de ordenador, siempre en el sentido de programas en ejecución, hacen lo mismo: manipular símbolos. Sin embargo, entre la IA simbólica y la IA subsimbólica hay una

diferencia muy importante respecto a cuáles son las *entidades semánticamente interpretables*. En la IA simbólica dichas entidades son algunos de los símbolos con los que se opera, aún cuando el lenguaje al que pertenecen, como hemos dicho, haya sido definido sin interpretación semántica. Por ejemplo, en el caso de una IA jugadora de ajedrez, ésta opera con expresiones como Af1-c4, en la que el símbolo "A" significa "alfil", el guión "-" significa "mueve a", y los pares de símbolos "f1" y "c4" significan dos casillas. En la IA subsimbólica, en cambio, los símbolos manipulados no refieren a contenidos mentales como alfiles o casillas de un tablero de ajedrez, sino a las neuronas y a sus propiedades, tales como la conductividad o el umbral de disparo. Así pues, en la IA subsimbólica, así como en las redes de neuronas naturales del cerebro, las entidades semánticamente interpretables no son las entidades manipuladas, sino sus patrones de activación (Smolensky, 1989, p. 239; Benítez, 2011b, p. 276). He aquí la diferencia crucial que hace que la IA simbólica no sea capaz de reemplazar, ni si quiera de manera instrumental, a las operaciones cerebrales sin correlato mental.

La IA simbólica puede adquirir dos grados de compromiso. Por un lado está la vertiente *realista*, que se fundamenta en una versión *fuerte* de la *hipótesis del sistema de símbolos* (Copeland, 1993, p. 273), según la cual la inteligencia es una propiedad exclusiva de los sistemas formales, de lo que se concluye que la IA simbólica fuerte ha de ser posible, pues la mente humana misma no sería más que un sistema formal, y no habría por tanto ninguna objeción de principio contra la posibilidad de descubrir las reglas de ese sistema formal que es la mente y suministrárselas a una computadora electrónica para que las ejecutase. Por otro lado está la vertiente *instrumental* de la IA simbólica, que se basa en una versión *débil* de la *hipótesis del sistema de símbolos*, según la cual los sistemas formales no son necesarios para la inteligencia, pero sí suficientes (Ibíd., p. 273), y por tanto es posible que la mente no sea un sistema formal, pero en cualquier caso un sistema formal es suficiente para producir una conducta inteligente indistinguible de la producida por la mente humana.

Como se puede apreciar, la vertiente instrumental adquiere un menor grado de compromiso. Hubert Dreyfus, uno de los filósofos más críticos contra la IA fuerte, denomina *supuesto psicológico* al compromiso de la IA realista (Dreyfus, 1992, p. 163), también conocida como *IA humana*, *IA de modo teórico* o *IA de simulación cognitiva*. Correlativamente, da el nombre de *supuesto epistemológico* al compromiso asumido por la *IA instrumental* (Ibíd., p. 189), que por su parte recibe los nombres alternativos de *IA ajena* o *IA de modo operativo* (Copeland, 1993, p. 54). Sin embargo, a pesar de esta diferencia, tanto la IA realista como la IA instrumental se apoyan en un supuesto común, de tal manera que refutando este supuesto, queda refutada toda la IA fuerte simbólica. Esa columna maestra, que en este artículo derribaremos señalando cuatro problemas insostenibles para ella, es el *supuesto*

ontológico. Dreyfus lo define así: «El supuesto ontológico (establece) que todo lo que es esencial para la conducta inteligente debe en principio ser comprensible en términos de un conjunto de elementos independientes determinados» (Dreyfus, 1992, p. 206).

La razón por la cual la IA simbólica se basa necesariamente en este supuesto es que las computadoras, en tanto que son sistemas formales, sólo pueden operar con elementos discretos, explícitos y determinados. Sea o no sea la mente humana un programa informático, lo indiscutible es que la IA simbólica fuerte necesita suponer que las expresiones cuasi-lingüísticas de un lenguaje formal son suficientes para que una computadora, operando transformaciones sobre ellas a nivel exclusivamente simbólico, produzca una especie de discurso mental capaz de originar una conducta inteligente indistinguible de la de un ser humano.

En la psicología cognitivista, que es el paradigma de la psicología (Carpintero, 1996, p. 404) que nació hermanado con la IA simbólica, se puede discutir con cuántos tipos de representaciones opera la mente humana (Gardner, 1985, p. 149). Ampliar el espectro a representaciones que no sean las cuasi-lingüísticas de los lenguajes formales es algo que, ciertamente, va en contra de una de las tesis nucleares del cognitivismo, a saber, la *metáfora computacional*, también llamada *tesis del procesamiento de información* (García, 2001, p. 18), según la cual la mente procesa la información de modo semejante a como lo hace un programa informático. Sin embargo, es una transgresión que la psicología puede permitirse, aunque sea a costa de poner en peligro la continuidad de un paradigma, pues la metáfora computacional no es más que una idea, y como todas las ideas, puede ser reemplazada por otra en el curso histórico de una ciencia. En cambio, la IA simbólica no puede permitirse ni tan siquiera el considerar otro tipo de representaciones mentales que no sean proposicionales, porque las máquinas que emplea en su intento por reproducir la inteligencia humana sólo ejecutan sistemas formales. Incluso cuando opera con figuras volumétricas, la computadora lo hace de manera proposicional, es decir, algebraica. Todo aquello que no pueda adoptar la forma de un conjunto de proposiciones es por definición incomputable. Aquí reside la limitación de estas máquinas, la carencia que no les permite reproducir simbólicamente ciertas conductas humanas.

Ésta es, en resumen, la tesis del supuesto ontológico: el mundo, tal y como nos lo representamos los seres humanos en la mente, debe ser descomponible en proposiciones. Si hubiera alguna facultad intelectual que no fuese obtenible como resultado de manipular esas proposiciones, entonces no sería reproducible a nivel simbólico por una computadora electrónica. Aquí está la clave de nuestro argumento: vamos a demostrar, mediante las cuatro mencionadas facultades intelectuales del ser humano, que en todas ellas la mente opera sobre la base de procesos cerebrales inconscientes, de los que no emerge correlato mental, y que no son reproducibles mediante un discurso simbólico, es decir, proposicional.

3.1. Sentido de la situación

La primera característica del intelecto humano imprescindible para producir una conducta como la nuestra pero que, sin embargo, no es reproducible de manera simbólica por una computadora, es el *sentido de la situación*, una expresión que refiere a nuestra capacidad para tener presente el conjunto de memorias adecuado para comprender la situación actual. Por poner un ejemplo tomado de Roger Schank, uno de los más notorios investigadores de la IA, comprender lo que es un restaurante de comida rápida es saber cosas tales como que son restaurantes con la particularidad de que se paga antes de comer (Schank, 1999, p. 27). Las memorias sobre los restaurantes de comida rápida conforman un *marco* (*frame*), término acuñado por uno de los padres de la IA, Marvin Minsky, en su artículo *A framework for representing knowledge* (Minsky, 1975, p. 96); aunque con algo de polémica, ya que otros dos fundadores de la disciplina, Allen Newell y Herbert Simon, reclamaron haberlo propuesto antes con el nombre de *esquema* (*schema*) (Crevier, 1993, p. 174).

El problema de encontrar el marco de memorias pertinente se conoce como *problema de la pertinencia* (Copeland, 1993, p. 143), o *problema de la selección del marco*. Fue advertido por Minsky en las últimas páginas del citado artículo, pero lo dejó abierto, sin solución, y sin solución permanece hasta nuestros días y permanecerá siempre, pues es un problema derivado de la forma proposicional de representación del conocimiento inherente a las computadoras electrónicas. Dreyfus lo formula así: «Para reconocer un contexto uno debe haber seleccionado previamente las características relevantes de entre el número indefinido de características, pero esa selección sólo puede ser realizada una vez el contexto ha sido reconocido como similar a otro ya analizado» (Dreyfus, 1979, p. 200).

Aplicándolo al ejemplo de Schank, esto quiere decir que para reconocer que algo es un restaurante de comida rápida uno debe haber seleccionado previamente las características relevantes que distinguen a ese objeto precisamente como un restaurante de comida rápida, pero esa selección sólo puede ser realizada una vez uno ha identificado que, en efecto, se trata de un restaurante de ese tipo, pues de lo contrario, sin un marco que sirva como esquema interpretativo para discriminar lo esencial de lo inesencial, la cantidad de características de cualquier situación es potencialmente infinita, y además se presentan todas como igualmente relevantes. Entre la totalidad, que es el marco, y las partes, que son los conocimientos de los que se compone, hay por tanto una relación circular de determinación recíproca. El problema para la IA simbólica es encontrar una manera de penetrar en el círculo.

Hay dos formas de hacerlo: utilizando información previa como criterio para dirigir la búsqueda del marco pertinente en la situación actual, o bien buscando hechos no interpretados, previos al acto comprensivo, que sean inequívocamente exclusivos de cada marco. A la primera la denominaremos estrategia temporal, mientras que a la segunda nos referiremos como estrategia de correspondencia.

Gadamer, que se ocupó prolijamente del asunto de la estructura circular de la comprensión, se refiere a la *estrategia temporal* al hablar de la traducción de un texto. Dice Gadamer: «La anticipación de sentido que hace referencia al todo sólo llega a una comprensión explícita a través del hecho de que las partes que se determinan desde el todo determinan a su vez a este todo. Este hecho nos es familiar por el aprendizaje de las lenguas antiguas. Aprendemos que es necesario "construir" una frase antes de intentar comprender el significado lingüístico de cada parte de dicha frase. Este proceso de construcción está sin embargo ya dirigido por una expectativa de sentido procedente del contexto de lo que le precedía. Por supuesto que esta expectativa habrá de corregirse si el texto lo exige. Esto significa entonces que la expectativa cambia y que el texto se recoge en la unidad de una referencia bajo una expectativa de sentido distinta. El movimiento de la comprensión va constantemente del todo a la parte y de ésta al todo. [...] El criterio para la corrección de la comprensión es siempre la congruencia de cada detalle con el todo. Cuando no hay tal congruencia, la comprensión ha fracasado» (Gadamer, 1960, p. 360).

En este fragmento, la sentencia que más nos interesa es la que dice que el proceso de construcción de las frases para entender su significado está dirigido por una expectativa de sentido procedente del contexto anterior. En el ejemplo de Schank, si suponemos que alguien nos ha invitado a comer, lo más seguro es que el lugar al que nos lleve sea un restaurante, y no una ferretería o una tintorería, una suposición ésta, basada en información previa, que permitiría acotar el rango de la búsqueda del marco pertinente al subconjunto de los marcos señalados con la etiqueta indicativa de lugares de restauración. A partir de ahí, comenzaría un proceso de prueba y error hasta encontrar el marco más adecuado, que será aquel que permita, como dice Gadamer, la comprensión más congruente entre el todo y las partes.

El defecto de esta estrategia, la temporal, que la hace insuficiente para resolver el problema de la selección del marco, es que supone una continuidad de sentido inquebrantable. En la lectura de ciertos textos puede concederse que la comprensión de lo anterior guía la de lo posterior, pero en la experiencia vital ordinaria nos asaltan continuamente estados de cosas imprevistos que no podían ser anticipados y, sin embargo, gracias a nuestro sentido de la situación, enseguida se nos hace presente un marco interpretativo que nos abre la comprensión de lo dado (Dreyfus, 1992, p. 218). Ser capaz de encontrar un marco no relacionado con lo anterior es imprescindible.

En cuanto a la *estrategia de correspondencia*, consiste, como decimos, en suponer que existen hechos brutos, no interpretados, previos al acto comprensivo, que corresponden siempre al mismo marco y sólo a él, y que por tanto servirían para identificarlo de manera inequívoca. Ésta es la solución adoptada por la gente de las computadoras, dice Dreyfus, para abordar el problema de la selección del marco: «En tanto que la computadora no se encuentra en situación [...] la solución de los

teóricos de la computación es construir la máquina para que responda a bits últimos descontextualizados, datos completamente determinados que no requieran interpretación ulterior para ser entendidos» (Ibíd., p. 204). Entendiendo por *universo* la totalidad de la realidad física, y por *mundo* la totalidad de las relaciones de sentido que actúan como el trasfondo a partir del cual las cosas son interpretadas (Rodríguez, 1987, p. 214), lo que desean los investigadores de la IA simbólica es encontrar puntos fijos de correspondencia entre ambas esferas: el universo y el mundo.

El problema de este enfoque es que tales puntos no existen. Muy al contrario, la misma situación mundana puede darse sobre estados de cosas físicos distintos, y un mismo estado de cosas físico puede ser interpretado como situaciones del mundo diferentes en función de las metas y los intereses de los seres humanos que participan en él (Dreyfus, 1992, p. 213). Un ejemplo lo encontramos en el ajedrez: el maestro puede ver la misma jugada en dos disposiciones físicas distintas de las piezas, y a la inversa, la misma disposición física de las piezas puede ser interpretada por el mismo maestro como dos jugadas diferentes en función del estilo del adversario.

En definitiva, es imposible que la IA simbólica consiga reproducir el sentido de la situación que tenemos los seres humanos. Nosotros nos encontramos en situación gracias a un proceso cerebral del que no emerge ningún correlato mental, es decir, un proceso puramente subsimbólico, y por tanto fuera del alcance por definición de la IA simbólica. El ingeniero y empresario informático Jeff Hawkins, interesado en la IA subsimbólica, describe así brevemente este proceso que ocurre en la corteza cerebral: «La intersección de estas dos sinapsis (de arriba-abajo y de abajo-arriba) nos proporciona lo necesario. [...] Este mecanismo de correspondencia de abajo-arriba/arriba-abajo nos permite decidir entre dos o más interpretaciones. [...] Donde los dos conjuntos se intersecan es lo que percibimos. [...] Así es como decidimos si la foto es de un jarrón o de dos caras» (se refiere a la ilusión óptica del jarrón de Rubin) (Hawkins & Blakeslee, 2004, p. 180). El problema de la selección del marco, así pues, no se resuelve, sino que se disuelve al abandonar el enfoque de la IA simbólica.

3.2. Sentido común

El *sentido común* es, a juicio de muchos, uno de los obstáculos principales de la IA (Kaku, 2011, p. 113). Al tratarse de un concepto muy amplio, que comprende habilidades necesarias para una gran cantidad de tareas ordinarias, aquí nos vemos en la necesidad de acotarlo y examinar sólo una de sus aplicaciones: la interpretación de reglas generales. Dada una regla general, los seres humanos sabemos gracias a nuestro sentido común cuándo es aplicable y cuándo no. El problema para la IA simbólica radica en que una computadora, al carecer de sentido común, aplicará la regla de manera indiscriminada, a ciegas (Weizenbaum, 1976, p. 43), a menos

que se le proporcione un listado de las condiciones de validez de la misma, lo cual suele ser imposible debido a que la cantidad de excepciones que puede haber a una regla es potencialmente infinita, y por tanto no hay forma de listarlas todas.

Este problema, conocido como el *problema de la cualificación*, fue identificado por John McCarthy y Patrick Hayes en un artículo de 1969 que lleva por título *Some philosophical problems from the standpoint of artificial intelligence*. El término "cualificación" se refiere en inglés a las *exceptions* o *qualifications* (Crevier, 1993, p. 120) de una regla, es decir, sus excepciones. McCarthy y Hayes lo formulan así: «El problema que se pretende abordar, a saber, la imposibilidad de nombrar todo lo que puede ir mal, es importante para la inteligencia artificial, y algún tipo de formalismo tiene que ser desarrollado para tratar con él» (McCarthy & Hayes, 1969, p. 34). Ellos ponen el ejemplo de un conjunto de reglas para realizar una llamada telefónica. Se supone que para realizar la operación con éxito basta con buscar en el listín telefónico el número de la persona a la que se desea llamar y marcarlo. Sin embargo, son muchas las cosas que pueden ir mal y hacer que estas reglas generales no sean suficientes: la página con el número de esa persona puede haber sido arrancada, el que realiza la llamada puede ser ciego, puede que la persona no figure en el listín porque acaba de darse de alta en la compañía telefónica, y así hasta el infinito.

El problema de la cualificación es un problema generado por el supuesto ontológico, según el cual, recordemos, todo lo que es relevante para la conducta inteligente puede ser formalizado en una descripción estructurada (Dreyfus, 1979, p. 200). El hecho es que los seres humanos somos capaces de discernir cuándo una regla general es aplicable y cuándo no, sin necesidad para ello de tener almacenado en la memoria un listado de proposiciones con sus condiciones de validez. El proceso mediante el cual sabemos cuándo es válida una regla se encuentra, de nuevo, por debajo del nivel de la mente, es decir, que es sólo cerebral, subsimbólico. Hay muchas cosas que sabemos sin saberlas explícitamente en forma oracional ni en ninguna otra forma de representación, y sin embargo, las sabemos y las aplicamos. Tal es el caso de las reglas generales: sin necesidad de pensar un conjunto de proposiciones que indiquen sus condiciones de validez, sabemos cuándo son aplicables.

Por supuesto, se puede sostener la hipótesis de que sí tenemos listados de condiciones de validez en algún lugar ignoto de la mente al cual no podemos acceder. De hecho, eso es lo que sostienen los defensores del supuesto psicológico. O bien, adquiriendo un menor compromiso, se puede defender, como hacen los partidarios del supuesto epistemológico, que esos listados se pueden obtener, y aunque no sean necesarios para producir la conducta inteligente, sí que son suficientes. Esto es lo que defiende Marvin Minsky en la introducción de *Semantic information processing*. Refiriéndose al conocimiento de sentido común que una persona emplea en la vida ordinaria, Minsky calcula la cantidad de cosas que una computadora electróni-

ca necesitará saber: «Pienso que una máquina necesitará como mínimo adquirir alrededor de cien mil elementos de conocimiento para comportarse de una manera razonablemente sensata en situaciones ordinarias. Si mi argumento no les convence, multipliquen las cifras por diez» (Minsky, 1968, p. 26).

Minsky se quedó bastante corto en opinión de Douglas Lenat, un investigador de la IA, convencido también del supuesto ontológico, que en 1984 puso en marcha uno de los más ambiciosos proyectos de la Historia de la IA simbólica: el CYC (Lenat & Guha, 1990). La máquina de Lenat requería que un grupo de operarios elaborase una base de datos con los enunciados que representan todos los conocimientos de sentido común necesarios para comportarse de manera sensata (Crevier, 1993, p. 241). Tras los seis primeros años de trabajo, el número de enunciados codificados en el lenguaje formal del CYC ascendía a más de un millón, lo cual representaba según Lenat el 0,1% de la realidad consensuada en Occidente (Copeland, 1993, p. 160). Lenat pensaba que el número de enunciados que la máquina necesitaría para ser operativa era de unos cien millones, es decir, el 10% de la realidad consensuada en Occidente, y también creía, erróneamente, que el proyecto se consumaría en 2007 (Kaku, 2011, p. 114).

La ontología y la epistemología implicadas en este planteamiento, el del supuesto ontológico, las encontramos presentes en la Historia de la filosofía en el *atomismo lógico*. Uno de los máximos exponentes de esta corriente positivista de la primera mitad del siglo XX fue Wittgenstein, quien en su *Tractatus logico-philosophicus* dejó plasmadas unas cuantas sentencias que reflejan bien las convicciones más profundas de algunos investigadores de la IA simbólica: «El signo proposicional usado, pensado, es el pensamiento» (Wittgenstein, 1921, §3.5). «El pensamiento es la proposición con sentido» (Ibíd., §4). «La proposición es una figura de la realidad. La proposición es un modelo de la realidad tal como nos la pensamos» (Ibíd., §4.01). «La especificación de todas las proposiciones elementales verdaderas describe el mundo completamente. El mundo queda completamente descrito por la especificación de todas las proposiciones elementales más la especificación de las que de ellas son verdaderas y de las que de ellas son falsas» (Ibíd., §4.26).

Según creía Wittgenstein en la época del *Tractatus*, existen proposiciones absolutamente simples, no descomponibles en otras. Él las denomina *proposiciones elementales* (*Elementarsätze*) o *completamente analizadas*, un concepto equivalente al de *proposición atómica* en la teoría de Bertrand Russell, maestro de Wittgenstein y fundador del atomismo lógico. Esas proposiciones, como dice Russell, expresan hechos atómicos (Russell, 1918, p. 27). La cuestión es que, según esta teoría, el mundo entero es descriptible mediante proposiciones atómicas, y a la inversa, la totalidad de las proposiciones atómicas verdaderas describe el mundo de manera total, exhaustiva, sin dejar nada fuera de la descripción. Esto es justo lo que acabamos de ver que sostienen Minsky y Lenat: que el mundo se agota en proposiciones.

Observemos la ineficacia de este planteamiento en un caso particular. En el artículo antes citado *A framework for representing knowledge*, Minsky analiza el trabajo de un alumno suyo, Eugene Charniak, acerca de los marcos de memorias necesarios para entender la siguiente historia infantil: «Jane fue invitada a la fiesta de cumpleaños de Jack. Ella se preguntó si le gustaría una cometa. Fue a su habitación y agitó su cerdito-hucha, pero no sonó» (Minsky, 1975, p. 103). Para entender la historia hace falta saber, entre otras muchas cosas, lo que es un cerdito-hucha, así que Charniak se lanza a intentar capturar en una descripción lingüística todo lo que hay que saber sobre los cerditos-hucha para entender las historias en las que aparezcan. La definición es muy extensa, y no es necesario transcribirla íntegra, sino que bastan unas cuantas oraciones para señalar su debilidad. Dice Charniak en su tesis doctoral: «Los cerditos-hucha (CH en adelante) los hay de todos los tamaños y formas, aunque la forma preferida es la de cerdito. *Generalmente* el tamaño oscila entre el del pomo de una puerta y una tartera. *Generalmente* en los CH se guarda dinero, por lo que cuando un niño necesita dinero a menudo irá a mirar en su CH. *Usualmente* para conseguir el dinero necesitas sostenerlo y agitarlo (de arriba a abajo). *Generalmente* ponerlo boca abajo facilita las cosas» (Crevier, 1993, p. 113). Lo importante aquí son los adverbios de frecuencia, que hemos resaltado con cursivas: *generalmente* y *usualmente*. Como se puede apreciar, la definición se compone de reglas generales que no siempre se cumplen sobre los atributos del objeto y las prácticas mundanas de las que participa. Para reproducir computacionalmente nuestra conducta, las definiciones de este tipo deberían ser complementadas con listados exhaustivos de las condiciones de validez que indiquen cuándo son aplicables, pero obtener tal cosa es imposible, porque nadie sabe explícitamente cuáles son esas condiciones.

Si los seres humanos operásemos a nivel inconsciente con definiciones, quizás podríamos formalizarlas en un programa informático. Pero la verdad es que no lo hacemos, y no lo hacemos sencillamente porque no se puede hacer, dado que no existe nada semejante a las definiciones reales de las cosas en el habla ordinaria. Tras renegar del *Tractatus*, Wittgenstein se dio cuenta de que: «Somos incapaces de delimitar claramente los conceptos que usamos; no porque no sepamos su definición real, sino porque no existe tal cosa como la "definición" real. Suponer que *debe* haberla es como suponer que cuando los niños juegan con una pelota están jugando a un juego con reglas estrictas» (Wittgenstein, 1958, p. 25). Toda definición es insuficiente para capturar la realidad de manera perfecta. Su impotencia es revelada por las cláusulas de renuncia en forma de adverbios de frecuencia.

El motivo por el cual los investigadores de la IA simbólica se empeñan, no obstante, en encontrar definiciones reales suficientes para la acción inteligente es que son víctimas de la obsesión logocéntrica de la cultura occidental. Como dice John Haugeland, los cognitivistas, entre los que se incluye la mayoría de los defensores

de la IA simbólica, son herederos de la tradición racionalista, según la cual ser inteligente es ser capaz de manipular representaciones claras y distintas mediante reglas racionales (Haugeland, 1978, p. 276). Dreyfus apunta en la misma dirección cuando señala a Platón como el precursor de la IA simbólica, en tanto que su objetivo, dice, era capturar todo el conocimiento en definiciones explícitas (Dreyfus, 1992, p. 67), es decir, en las representaciones claras y distintas a las que alude Haugeland. Y, por citar un tercer testimonio relevante en contra de las pretensiones de la IA simbólica, el ingeniero informático del MIT Joseph Weizenbaum denuncia que la verosimilitud de esta disciplina se debe a una corriente de pensamiento, en referencia al positivismo, que ha determinado hace tiempo que «la vida es aquello que puede someterse a cálculo y nada más» (Weizenbaum, 1976, p. 167).

La única razón que hay para creer en la posibilidad de formalizar todo el sistema de creencias del sentido común es el prejuicio cientifista sin base científica de que el cálculo operado en un sistema formal puede dominar la realidad al completo. Ésta es la convicción de McCarthy, según la cita Weizenbaum: «La única razón por la cual no hemos tenido éxito en la simulación de todos los aspectos del mundo real es que nos ha faltado un cálculo lógico lo bastante poderoso. En la actualidad, me encuentro trabajando en ese problema» (Ibíd., p. 167). McCarthy murió en 2011 y, por supuesto, se fue sin haber resuelto ese problema.

3.3. Habilidades procedimentales

Las *habilidades procedimentales* son las relativas a procesos o procedimientos para llevar a cabo alguna acción (García, 1996, p. 304). En este artículo, por razones de espacio, acotaremos el significado y nos referiremos sólo a las habilidades procedimentales más complejas, que son las que permiten la ejecución de acciones como bailar, conducir o subir escaleras. Frente al conocimiento declarativo o explícito del "saber qué" (*know what*), las habilidades procedimentales o implícitas son las del "saber cómo" (*know how*) (Kandel, Schwartz & Jessell, 1995, p. 656). El problema de la IA simbólica al respecto es que, al ser los sistemas formales conjuntos de expresiones cuasi-lingüísticas, necesita conceptualizar las conductas de las habilidades procedimentales como si fueran el resultado de un discurso mental.

En su relato *Instrucciones para subir una escalera*, el escritor argentino Julio Cortázar hace patente el absurdo resultante de codificar una habilidad procedimental, como es el subir escaleras, en una secuencia de órdenes explícitas: «Nadie habrá dejado de observar que con frecuencia el suelo se pliega de manera tal que una parte sube en ángulo recto con el plano del suelo, y luego la parte siguiente se coloca paralela a este plano, para dar paso a una nueva perpendicular, conducta que se repite [...] hasta alturas sumamente variables. [...] Para subir una escalera se comienza por levantar esa parte del cuerpo situada a la derecha abajo, envuelta casi siempre

en cuero o gamuza, y que salvo excepciones cabe exactamente en el escalón. Puesta en el primer peldaño dicha parte, que para abreviar llamaremos pie, se recoge la parte equivalente de la izquierda (también llamada pie, pero que no ha de confundirse con el pie antes citado), y llevándola a la altura del pie, se le hace seguir hasta colocarla en el segundo peldaño, con lo cual en éste descansará el pie, y en el primero descansará el pie. (Los primeros peldaños son siempre los más difíciles, hasta adquirir la coordinación necesaria. [...] Cuídese especialmente de no levantar al mismo tiempo el pie y el pie). Llegando en esta forma al segundo peldaño, basta repetir alternadamente los movimientos hasta encontrarse con el final de la escalera» (Cortázar, 1962, p. 11).

Los partidarios de la IA simbólica fuerte están obligados a suponer que, aunque no sea así como opera la mente humana, una secuencia de instrucciones semejante tiene que ser suficiente para que una computadora suba escaleras. Obviamente la secuencia estará escrita en un lenguaje formal, y será mucho más compleja, pero en esencia seguirá siendo una secuencia como la de Cortázar, en tanto que consistirá en un conjunto de enunciados que, de acuerdo a unas reglas de cálculo, se transforman unos en otros. Por ejemplo, cuando Cortázar advierte que los dos pies no deben confundirse, el programador lo tendrá en cuenta creando dos objetos distintos de la clase "pie", así como también programará una función supervisora que se ocupe de que ambos pies no se intenten levantar al mismo tiempo. Es evidente que a este *algoritmo*, que es el nombre técnico de las secuencias de instrucciones ejecutadas por las computadoras, se le puede objetar también el problema de la cualificación, pues sus instrucciones generales sólo son aplicables si se cumple una lista indefinidamente extensa de condiciones de validez. El problema de la cualificación, al estar derivado de la falta de sentido común y al ser el sentido común algo necesario para casi todas las actividades de la vida ordinaria, afecta ampliamente a la IA simbólica. No obstante, como ya hemos hablado del problema de la cualificación, centrémonos aquí en la problemática específica de las habilidades procedimentales.

El problema de las habilidades procedimentales para la IA simbólica suele ser planteado señalando la diferencia entre el novato y el experto que ejecutan la misma acción (Franklin, 1995, p. 103). Según esta diferencia, el novato ejecutaría una acción como conducir un coche aplicando conscientemente una serie de reglas, mientras que el experto lo haría sin pensar en ellas. La cuestión es que, aunque tal diferencia existe, a nuestro juicio no va a la raíz del problema: y es que ni siquiera el novato realiza la acción *solamente* ejecutando reglas. El novato tiene presentes las instrucciones, tales como que cuando el motor se revoluciona hay que subir de marcha o que en los pasos de peatones hay que disminuir la velocidad, pero ésas son órdenes de alta jerarquía, muy generales, y por tanto según la IA simbólica deberían concretarse en una miriada de órdenes cuasi-lingüísticas de bajo nivel tales como tensionar en tal medida éstos y aquellos músculos de la última falange del dedo meñique de la mano derecha.

Es curioso que los cognitivistas, que como hemos señalado son los psicólogos del paradigma multidisciplinar al que pertenece la IA simbólica, reclamen la posibilidad de investigar las operaciones mentales más refinadas al margen de su correlato fisiológico cerebral (Damasio, 1994, p. 286), y sin embargo no reclamen lo mismo respecto de las menos refinadas, como la de mover el dedo meñique. La realidad es que las habilidades procedimentales, aún cuando se ejecutan de manera consciente siguiendo un algoritmo, requieren de la participación de mecanismos puramente cerebrales, esto es, de los que no emerge representación mental alguna.

Se sabe, por ejemplo, que la locomoción, a pesar de que es una acción voluntaria y por tanto dirigida mentalmente, no requiere de dirección consciente en condiciones normales después de haberse iniciado (Kandel, Schwartz & Jessell, 1995, p. 523). Si a una rata se le secciona la médula espinal a la altura de las vértebras lumbares, cuando se la ponga sobre una cinta andadora realizará de todas maneras la acción locomotriz de alternancia de las patas traseras (Ibíd., p. 524), un experimento real que pone en evidencia el absurdo del experimento mental de Cortázar, pues buena parte de lo que hacemos al subir una escalera, que no es más que una variante del caminar, se decide neuronalmente a nivel de la médula espinal, sin que la mente intervenga. Si los dos pies no se levantan al mismo tiempo no es porque una orden mental subconsciente lo dicte, sino porque un circuito espinal inhibitorio lo impide.

En resumen, las habilidades procedimentales suponen uno de los mayores obstáculos para la IA simbólica en sus dos vertientes. La duplicación de manera realista es imposible porque se trata de habilidades basadas en procesos cerebrales, o incluso neuronales a nivel de la médula espinal, de los que no emerge ningún correlato mental. Y la duplicación de forma instrumental es imposible debido a que las proposiciones que deberían sustituir funcionalmente a dichos procesos cerebrales inconscientes son inoperantes a causa del problema de la cualificación.

3.4. Abducción de teorías

Citando a Habermas, las ciencias naturales, denominadas por él como ciencias empírico-analíticas, no son más que «la continuación sistemática de un proceso de aprendizaje acumulativo que se realiza de forma precientífica en el ámbito funcional de la actividad instrumental» (Habermas, 1968, p. 194). Esto implica, entre otras muchas cosas, que las estrategias de producción de teorías a nivel científico son las mismas que a nivel precientífico. Charles Peirce distinguió, en un esquema clásico, tres formas de inferencia: inducción, abducción y deducción. A su juicio, la única que verdaderamente sirve para descubrir nuevas teorías, y por tanto la única del contexto de descubrimiento, es la abducción (Rivadulla, 2010, p. 120). No obstante, la inducción, así como la deducción en forma de preducción, también son consi-

deradas a menudo como estrategias de producción de teorías (Rivadulla, 2009, p. 244). Por ser la preducción exclusiva de las ciencias teóricas, su consideración como una estrategia de producción de teorías a nivel precientífico es dudosa, y por tanto aquí sólo examinaremos la inducción y, sobre todo, la abducción.

Lo que nos proponemos mostrar es que, para crear una auténtica IA simbólica fuerte, habría que dotarla de algoritmos para producir teorías tal y como lo hacemos los seres humanos constantemente en la vida ordinaria, pero el caso es que ni la inducción ni la abducción son formalizables, es decir, que no existe un camino lógico para producir teorías, ni inductiva ni abductivamente. Primero observemos el hecho: no existen máquinas capaces de inducir o abducir de manera genuina, es decir, sin una determinación previa por parte de sus creadores humanos de cuáles son los objetos y las características sobre los que deben centrarse. La prueba de esta afirmación es que no existen computadoras capaces de automatizar el proceso de producción de teorías científicas. He aquí una relación interesante entre la IA simbólica y el mito del método científico. Si el método científico, entendido como un conjunto de procedimientos cuya aplicación garantiza el logro de conocimiento, existiera, entonces la IA simbólica fuerte sería una realidad. Y, a la inversa, si la IA simbólica fuerte fuera posible, entonces sus algoritmos de inducción y abducción de teorías constituirían el método científico en el señalado sentido de procedimiento seguro.

La cuestión es que tal método no existe. Albert Einstein, que de crear teorías científicas algo sabía, afirmó lo siguiente en referencia a las leyes fundamentales del universo: «No hay camino lógico que lleve a estas leyes fundamentales. Debemos dejarnos conducir por la intuición, que se basa en una sensación de la experiencia. [...] Nadie que haya profundizado de veras en esto podrá negar que el sistema teórico ha sido prácticamente determinado por el mundo de las suposiciones, pese a que no existe camino lógico alguno que conduzca desde éstas hasta las leyes fundamentales» (Einstein, 1953, p. 131). Cuando dice "intución", Einstein emplea el término en su sentido vulgar de "corazonada". Para el físico más importante del siglo XX, por tanto, la ciencia comienza su tarea de descubrimiento de teorías con algo tan vago y difuso como las corazonadas. Son las corazonadas las que mueven al científico a realizar una cierta inducción o abducción, y no otra cualquiera.

Respecto a las razones por las cuales la inducción y la abducción no podrán ser formalizadas nunca, tienen que ver con el problema de la selección del marco. Empezando por la inducción, ésta se puede definir como un modo de inferencia que consiste en elevarse de enunciados observacionales particulares a normas generales. Por ejemplo, habiendo visto siempre que los cuervos son negros, se concluye por inducción que todos los cuervos son negros. El problema de la *inducción* para la IA simbólica es que en cualquier observación el número de variables que concurren es, en potencia, infinito, y sin un marco que discrimine cuáles son las esencia-

les, la inducción es imposible. En el ejemplo del cuervo hay que ser capaz de darse cuenta de que lo esencial para que el animal sea siempre negro es que pertenezca a la especie de los cuervos, y no otros factores que también estaban presentes en el momento de las observaciones, tales como la hora, la temperatura o la humedad relativa. Respecto a la *abducción*, sucede algo parecido. La abducción, en palabras de Peirce, «consiste en estudiar hechos e inventar una teoría que los explique» (Rivadulla, 2009, p. 239). Por ejemplo, dado el fenómeno de la fuerza, Newton abdujo que se trataba de una magnitud directamente proporcional a la masa y la aceleración, formulando así su segunda ley de la dinámica: F=ma. El problema para la IA simbólica es que para seleccionar justo esos factores entre un número potencialmente infinito hace falta conocer de antemano aquello que se busca: la propia teoría.

Ciertamente, las teorías que producimos a nivel ordinario, y que una IA simbólica fuerte necesitaría producir, son sobre otro tipo de cosas: por qué alguien se comporta de determinada manera, cuándo pasará el autobús o qué es lo que pasará cuando un político gane las elecciones. Sin embargo, la esencia del problema es la misma, y es que la inducción y la abducción requieren un marco previo que dirija la mirada a unos hechos y no a otros. Se trata, en definitiva, de una reedición moderna de la paradoja del *Menón*. Newell y Simon, a quienes ya hemos mencionado por ser dos de los fundadores de la IA, creyeron haberla resuelto: «La paradoja del *Menón* es resuelta por la observación de que la información puede ser recordada, y también la información nueva puede ser extraída del dominio designado por los símbolos» (Newell & Simon, 1976, p. 65). El problema de este enfoque es que la revisión de la experiencia anterior no puede ser indiscriminada, sino que debe estar orientada a aquellos datos útiles para formular una teoría, pero tal orientación sólo puede ser proporcionada precisamente por aquello que se busca: la teoría. De lo contrario, por mucho que se revise la experiencia anterior, jamás se encontrará en ella nada útil.

4. La obsesión positivista por el dominio

A través del examen de cuatro facultades del intelecto humano imprescindibles para la conducta inteligente hemos demostrado la tesis inicial de que la IA simbólica es imposible. Imposible de manera realista, en tanto que las citadas facultades dependen de procesos cerebrales de los que no emerge ningún correlato mental, e imposible también de modo instrumental, pues el enfoque simbólico de la manipulación de símbolos referidos al nivel de las representaciones mentales se fundamenta en un supuesto, el ontológico, que da lugar a varios problemas irresolubles. La inteligencia entendida como el discurso mental silencioso de formación y transfor-

mación de las representaciones cuasi-lingüísticas de un sistema formal no es capaz de reproducir ni tan siquiera conductualmente el sentido de la situación, el sentido común, las habilidades procedimentales y la abducción de teorías. Los problemas de fondo son sólo dos, el de la selección del marco y el de la cualificación, pero hemos optado por exponerlos concretando cuatro de sus consecuencias.

Lo único que resta es explicar por qué, a pesar de la imposibilidad de la IA simbólica fuerte, la IA se presenta de continuo como una técnica inminente. Podría pensarse que la razón por la cual la IA se presenta como una realidad en ciernes es la IA subsimbólica. De hecho, la IA subsimbólica sí es posible en principio, en tanto que el objeto que intenta reproducir, el cerebro, no es más que materia y los sistemas formales, que es lo que son las computadoras electrónicas, se han mostrado capaces hasta el momento de capturar los rasgos esenciales de cualquier parcela del universo físico mediante la simulación de modelos. Sin embargo, esta explicación no sería correcta por dos razones: una técnica y la otra histórica.

La histórica es que la posibilidad de la IA fuerte fue proclamada por los científicos desde el momento mismo en que esta disciplina se constituyó como tal, en 1956, una fecha que es muy anterior al primer desarrollo de la IA subsimbólica, que se produjo en la década de 1980. Es verdad que la IA subsimbólica fue planteada mucho antes, pues Turing ya describió el funcionamiento de una red de neuronas artificiales en 1948 (Turing, 1948, p. 418), Wesley Clark y Belmont Farley hicieron la primera simulación neuronal computerizada de la Historia en 1954 (Copeland, 2004, p. 406), y Frank Rosenblatt presentó su red de neuronas, el Perceptrón, en 1958. Sin embargo, debido a diversos conflictos sociales en el seno de la IA (Crevier, 1993, p. 107), el programa de investigación subsimbólico permaneció prácticamente inexplorado hasta principios de los 80, cuando fue retomado por John Hopfield y el conocido como Grupo PDP. Por tanto, históricamente no ha podido ser la IA subsimbólica la que ha dado crédito a la convicción popular de que las computadoras electrónicas son el soporte adecuado para producir máquinas inteligentes.

Y técnicamente, tampoco, ya que las simulaciones de neuronas a gran escala pertenecen todas al ámbito de la IA débil, que como ya hemos señalado es aquella vertiente de la IA que se conforma con realizar simulaciones de ciertos fenómenos mentales o cerebrales para hacer avanzar el conocimiento científico. Hay redes de neuronas artificiales que tienen aplicaciones prácticas más allá de la investigación científica, por supuesto, pero están todavía muy lejos de exhibir las competencias intelectuales de un ser humano, o incluso de algún otro mamífero. Un ejemplo serían los programas de reconocimiento facial. Las redes de neuronas son eficaces para realizar tareas de reconocimiento de patrones (Rumelhart, 1989, p. 216), tales como rostros, voces o grafemas. Sin embargo, los programas de ese tipo están todavía muy lejos de exhibir una flexibilidad como la nuestra. Si esto es todo lo que ha

logrado la IA subsimbólica en tres décadas, entonces su estado de desarrollo técnico tampoco da respaldo científico a la creencia de que las computadoras electrónicas reproducirán nuestras facultades intelectuales de aquí a poco tiempo.

La verdadera razón por la cual la IA se presenta como una técnica inminente es la antes mencionada obsesión de la sociedad actual por el control. Las explicaciones que ofrecen la forma de control más precisa son las de estilo nomológico-deductivo, es decir, las del tipo de la física (Horkheimer, 1947, p. 102). Suponer que la mente es calculable mediante un sistema formal es tanto como suponer que es calculable de manera nomológico-deductiva y, en consecuencia, que se puede ejercer sobre ella un control tan exhaustivo como el que las ciencias empíricas ejercen sobre la naturaleza. Ésta es la clave: la obsesión positivista por el dominio, por hacerlo todo calculable para manipularlo y pronosticarlo. Dijimos antes que la única razón que hay para creer en la posibilidad de formalizar todo el sistema de creencias del sentido común es el prejuicio cientifista sin base científica de que el cálculo operado en un sistema formal puede dominar la realidad al completo. Pues bien, ese prejuicio cientifista tiene su razón de ser en el ciego afán de dominio en el que ha caído nuestra civilización como resultado del desastre de la Ilustración (Horkheimer & Adorno, 1944, p. 131). Lo que pretendía ser un proyecto emancipatorio ha resultado en una cosmovisión, la positivista, que lo único a lo que aspira es a aumentar el dominio, tanto sobre la naturaleza como sobre los demás seres humanos, sin más fin que el propio aumento tumoral.

Creer en la posibilidad técnica de la IA simbólica fuerte es creer que existe un discurso racional capaz de producir cualquier conducta inteligente de manera perfecta y autosuficiente, es decir, sin depender de la acción irracional en tanto que subsimbólica de ciertos procesos cerebrales de los que no emerge correlato mental alguno. Decimos que la IA subsimbólica es irracional porque, aunque la actividad de las redes de neuronas es en principio calculable, su forma no es intencional, esto es, que los símbolos manipulados en una simulación de ese tipo no refieren a nada externo al sujeto, y ni tan siquiera al sujeto, sino sólo a una parte de su sustrato material, que es el cerebro. Incluso en el caso de que se llegara a calcular la actividad neuronal de todo el sistema nervioso, ese cálculo sería de algo tan extraño al hombre como lo es el cálculo del funcionamiento del corazón o de cualquier otro órgano, porque el hombre se siente a sí mismo como fenómeno psíquico, y no como fenómeno físico.

En 1997 la IA jugadora de ajedrez Deep Blue, creada por IBM, derrotó al campeón del mundo Gary Kaspárov. Éste, ante la sospecha de que una jugada bastante extraña le hubiera sido indicada a la máquina por un operario humano, pidió que se le facilitase una copia de los procesos de Deep Blue. Si Kaspárov pidió ver esos procesos es porque eran simbólicos, pues de lo contrario, si hubieran sido subsimbólicos, no habría sido capaz de entender nada. Mirar los registros de procesos sub-

simbólicos es como mirar un cerebro al microscopio, y un cerebro al microscopio no es una teoría que explique la visión o cualesquiera otras operaciones cognitivas que realice (Davidson, 1973, p. 353). En el mejor de los casos, la IA subsimbólica construirá algún día una red de neuronas capaz de hacer lo mismo que nuestro cerebro y de la misma manera, pero esa red sólo sería una copia (Benítez, 2011b, p. 277). Decir que esa red es una teoría explicativa de la cognición humana es como decir que tirar una maceta desde un segundo piso es una teoría sobre la caída de los graves.

Por eso se denomina sub-simbólico al paradigma conexionista de la IA: porque, aún a pesar de que sus modelos, en tanto que programas informáticos en ejecución, siempre consistirán en manipulaciones de símbolos de un lenguaje formal, esos símbolos no simbolizan contenidos mentales. Símbolo es una palabra que viene del griego σύμθυλον, y significa «juntar, reunir lo que estaba disperso o separado» (Díaz, 1997, p. 451). Un símbolo, por tanto, es tal porque apunta hacia otra cosa que no es él mismo, sino que está separada. La IA simbólica se denomina así porque tiene un carácter intencional en el sentido en que Brentano habla de la intencionalidad (Benítez, 2011b, p. 275), es decir, que sus símbolos, como los pensamientos, refieren a cosas externas a ellos. Gracias a esta cualidad intencional compartida con nuestro psiquismo, la IA simbólica tiene respecto a la subsimbólica la particularidad de que pretende hacer transparente la mente ante la mente. Aunque sea una mente de tipo instrumental, pues a los científicos, como dice Steve Woolgar, no suele preocuparles el estatus epistemológico de sus teorías (Woolgar, 1988, p. 133). La IA simbólica pretende iluminar el fenómeno psíquico de la misma manera que la física ha iluminado el fenómeno físico: para dominarlo y someterlo al deseo ciego del poder.

Así como Galileo renunció a la pretensión de realismo del modelo heliocéntrico, los investigadores de la IA simbólica fuerte también están dispuestos a renunciar al realismo. Lo que les importa es, metafóricamente, tener un modelo eficaz para poner satélites en órbita y, deshaciendo la metáfora, tener un modelo de la mente calculable. Aquí está el gran interés filosófico de la IA simbólica: se trata de un intento por demostrar que la razón, en su máxima expresión ilustrada del cálculo de los sistemas formales, puede conocerse a sí misma perfectamente de forma instrumental, para aplicar ese conocimiento a la expansión del dominio. La consecuencia sería la automatización, por fin, de todas las tareas intelectuales, con lo que se acabaría el castigo bíblico al trabajo, y la humanidad quedaría totalmente emancipada. Al precio, eso sí, de la caída en el principio de inmanencia (Hokheimer & Adorno, 1944, p. 67) y de convertirnos en lotófagos (Ibíd., p. 114). La IA simbólica fuerte es, en definitiva, el penúltimo delirio de la razón degenerada en positivismo. Afortunadamente, su fracaso está garantizado por problemas irresolubles como los analizados en este artículo.

Referencias bibliográficas

Benítez, A. (2011a): *Fundamentos de inteligencia artificial. Libro primero: Programación en Scheme*, Madrid, Escolar y Mayo.

Benítez, A. (2011b): *Fundamentos de inteligencia artificial. Libro segundo: Inteligencia artificial clásica*, Madrid, Escolar y Mayo.

Benítez, A. (2013): *Fundamentos de inteligencia artificial. Libro tercero: Inteligencia artificial bioinspirada*, Madrid, Escolar y Mayo.

Carpintero, H. (1996): *Historia de las ideas psicológicas*, Madrid, Pirámide.

Copeland, B. J. (1993): *Artificial intelligence: A philosophical introduction*, Oxford, Blackwell. Seguimos la traducción de Julio César Armero San José: *Inteligencia artificial: Una introducción filosófica*, Madrid, Alianza, 1996.

Copeland, B. J. (ed.) (2004): *The essential Turing*, Oxford, Clarendon Press.

Cortázar, J. (1962): *Historias de cronopios y de famas*, Buenos Aires, Alfaguara.

Crevier, D. (1993): *The tumultuous history of the research for AI*, New York, Basic Books.

Damasio, A. (1994): *Descartes' error: Emotion, reason and the human brain*, New York, Macmillan. Seguimos la traducción de Joandomènec Ros: *El error de Descartes: La emoción, la razón y el cerebro humano*, Barcelona, Crítica, 2007.

Davidson, D. (1973): "The material mind", en P. Suppes et al. (eds.), *Logic, methodology and philosophy of Science IV*, Amsterdam, North-Holland. Seguimos la reimpresión en J. Haugeland (ed.), *Mind design*, Cambridge, The MIT Press, 1981, pp. 339-354.

Díaz, C. (1997): *Manual de Historia de las religiones*, Bilbao, Desclée de Brouwer.

Dreyfus, H. L. (1979): "From micro-worlds to knowledge representation: AI at an impasse", en J. Haugeland (ed.), *Mind design*, Cambridge, The MIT Press, 1981, pp. 161-204.

Dreyfus, H. L. (1992): *What computers still can't do*, Cambridge, The MIT Press.

Einstein, A. (1953): *Mein weltbild*, Zürich, Europa Verlach (ed. Carl Seelig). Seguimos la traducción de Sara Gallardo y Marianne Bübeck: *Mi visión del mundo*, Barcelona, Tusquets, 2002.

Falguera, J. L., & Martínez, C. (1999): *Lógica clásica de primer orden*, Madrid, Trotta.

Franklin, S. (1995): *Artificial minds*, Cambridge, The MIT Press.

Gadamer, H. G. (1960): *Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik*, Tübingen, JCB Mohr. Seguimos la traducción de Ana Agud Aparicio y Rafael de Agapito: *Verdad y método I*, Salamanca, Sígueme, 1977.

García, E. (1996): "Inteligencia y metaconducta", *Revista de psicología general y aplicada*, 50 (3), pp. 297-312.

García, E. (2001): *Mente y cerebro*, Madrid, Síntesis.

GARDNER, H. (1985): *The mind's new science: A history of the cognitive revolution*, New York, Basic Books. Seguimos la traducción de Leandro Wolfson: *La nueva ciencia de la mente: Historia de la revolución cognitiva*, Barcelona, Paidós, 1988.

HABERMAS, J. (1968): *Erkenntnis und Interesse*, Frankfurt am Main, Suhrkamp Verlag. Seguimos la traducción de Manuel Jiménez, José F. Ivars y Luis Martín Santos: *Conocimiento e interés*, Madrid, Taurus, 1990.

HAUGELAND, J. (1978): "The nature and plausibility of cognitivism", *The Behavioral and brain sciences*, 1, pp. 417-424, Cambridge, Cambridge University Press. Seguimos la reimpresión en J. Haugeland (ed.), *Mind design*, Cambridge, The MIT Press, 1981, pp. 243-281.

HAUGELAND, J. (1981): "Semantic engines", en J. Haugeland (ed.), *Mind design*, Cambridge, The MIT Press, pp. 1-34.

HAUGELAND, J. (1996): "What is mind design?", en J. Haugeland (ed.), *Mind design II*, Cambridge, The MIT Press, 1997, pp. 1-28.

HAWKINS, J., & BLASKESLEE, S. (2004): *On intelligence*, New York, Times Books. Seguimos la traducción de Carmen Martínez Gimeno: *Sobre la inteligencia*, Madrid, Espasa, 2005.

HORKHEIMER, M. (1947): *Eclipse of reason*, New York, Oxford University Press. Seguimos la traducción de Jacobo Muñoz: *Crítica de la razón instrumental*, Madrid, Trotta, 2002.

HORKHEIMER, M., & ADORNO, TH. W. (1944): *Dialektik der Aufklärung*, New York, Social Studies Association. Seguimos la traducción de Juan José Sánchez: *Dialéctica de la Ilustración*, Madrid, Trotta, 2009.

KAKU, M. (2011): *Physics of the future: How science will shape human destiny and our daily lives by the year 2100*, New York, Random House. Seguimos la traducción de Mercedes García Garmilla: *La física del futuro*, Barcelona, Debate, 2011.

KANDEL, E. R., SCHWARTZ, J. H., & JESSELL, TH. M. (eds.) (1995): *Essentials of neural science and behavior*, Stamford, Prentice Hall.

LENAT, D. B., & GUHA, R. V. (1990): *Building large knowledge-based systems: Representation and inference in the CYC project*, Reading, Adisson-Wesley.

MCCARTHY, J., & HAYES, P. J. (1969): "Some philosophical problems from the standpoint of artificial intelligence", en B. Meltzer & D. Michie (eds.), *Machine intelligence*, vol. 4, Edinburgh, Edinburgh University Press. Seguimos la edición digital gratuita descargada de www-formal.stanford.edu/jmc.

MINSKY, M. L. (1975): "A framework for representing knowledge", *Memo 306 of the Artificial Intelligence Laboratory of the MIT*, Cambridge, The MIT Press. Seguimos la reimpresión en J. Haugeland (ed.), *Mind design*, Cambridge, The MIT Press, 1981, pp. 95-128.

Minsky, M. L. (ed.) (1968): *Semantic information processing*, Cambridge, The MIT Press.

Newell, A., & Simon, H. A. (1976): "Computer science as empirical enquiry: Symbols and search", *Communications of the Association for Computing Machinery*, 19 (March 1976), pp. 113-126. Seguimos la reimpresión en J. Haugeland (ed.), *Mind design*, Cambridge, The MIT Press, 1981, pp. 35-66.

Rivadulla, A. (2009): "El mito del método y las estrategias del descubrimiento científico: Inducción, abducción, preducción", en O. Pombo & A. Nepomuceno (eds.), *Lógica e Filosofía da Ciência*, Centro de Filosofía da Ciências da Universidade de Lisboa, Colecçao Documenta 2, Lisboa, pp. 231-246.

Rivadulla, A. (2010): "Estrategias del descubrimiento científico: Abducción y preducción", *Filosofía e História da Ciência no Cone Sul*, 6º encontro, pp. 120-129.

Robinet, A. (1973): *Le défi cybernétique*, Paris, Editions Gallimard. Seguimos la traducción de Carmen García-Trevijano: *Mitología, filosofía y cibernética*, Madrid, Tecnos, 1982.

Rodríguez, R. (1987): *Heidegger y la crisis de la época moderna*, Madrid, Síntesis.

Rumelhart, D. E. (1989): "The architecture of mind: A connectionist approach", en M. Posner (ed.), *Foundations of cognitive science*, Cambridge, The MIT Press. Seguimos la reimpresión en J. Haugeland (ed.), *Mind design II*, Cambridge, The MIT Press, 1997, pp. 205-232.

Russell, B. (1918): *The philosophy of logical atomism*, London, Routledge.

Schank, R. C. (1999): *Dynamic memory revisited*, Cambridge, Cambridge University Press.

Searle, J. R. (1980): "Minds, brains and programs", *The behavioral and brain sciences*, 3, Cambridge, Cambridge University Press, pp. 417-424. Seguimos la reimpresión en J. Haugeland (ed.), *Mind design*, Cambridge, The MIT Press, 1981, pp. 282-306.

Smolensky, P. (1989): "Connectionist modeling: Neural computation, mental connections", en L. Nadel, L. A. Cooper, P. Culicover & M. Harnish (eds.), *Neural connections, mental computation*, Cambridge, The MIT Press. Seguimos la reimpresión en J. Haugeland (ed.), *Mind design II*, Cambridge, The MIT Press, 1997, pp. 233-250.

Turing, A. M. (1948): "Intelligent machinery", *Turing Papers*, Cambridge, King's College Modern Archive Centre. Seguimos la reimpresión en J. Copeland (ed.), *The essential Turing*, Oxford, Clarendon Press, pp. 410-432.

Turing, A. M. (1950): "Computing machinery and intelligence", *Mind*, 59. Seguimos la reimpresión en J. Copeland (ed.), *The essential Turing*, Oxford, Clarendon Press, 2004, pp. 441-464. Hay edición en español con introducción de Ricardo Álvarez y traducción de Alejandro Bazán: *¿Puede pensar una máquina?*, Buenos Aires, Almagesto, 1990.

WEIZENBAUM, J. (1976): *Computer power and human reason*, San Francisco, W. H. Freeman and Company. Seguimos la traducción de Santiago Páez Fuentes: *La frontera entre el ordenador y la mente*, Madrid, Pirámide, 1978.

WITTGENSTEIN, L. (1921): "Logisch-philosophische Abhandlung", en W. Oswald (ed.), *Annalen der Naturphilosophie*, 14, Leipzig. Seguimos la traducción de Jacobo Muñoz e Isidoro Reguera: *Tractatus logico-philosophicus*, Madrid, Alianza, 2000.

WITTGENSTEIN, L. (1958): *The blue and brown books*, New York, Harper & Row.

WOOLGAR, S. (1988): *Science: The very idea*, London, Routledge. Seguimos la traducción de Eduardo Aibar: *Ciencia: Abriendo la caja negra*, Barcelona, Anthropos, 1991.

Manuel Carabantes López
manuel.carabantes@gmail.com