Inteligencia Artificial: la intrahistoria

Antonio Benítez

VIII Seminario Internacional e Interuniversitario
de Biomedicina y Derechos Humanos
Madrid, 2014

Índice

1. ¿Se puede investigar la mente?

Al final de la segunda mitad del siglo XIX la Psicología científica, experimental estaba ya en laboratorios, en universidades. Wundt, Fechner, Ebbinghaus, Külpe en Alemania; Sechenov y Pavlov en Rusia; James, Titchener, Morgan, Thorndike, Watson en USA, son algunos de los nombres de ilustres psicólogos.

Es común a las diversas tradiciones nacionales el convencimiento de que la mente tiene dos características esenciales: primera, la vida mental acontece en primera persona. Es decir, que cada acto, objeto y contenido, lo son del sujeto que vive esa vida. La vida mental es subjetiva; segundo, sólo el sujeto de la vida mental puede acceder a las experiencias mentales. Nadie puede observar lo que pasa en la mente de otro.

El conductismo terminó siendo la psicología predominante en Estados Unidos, aunque su influencia se extendió a otros países. John B. Watson definió las bases del conductismo.

Es propio del método científico que las experiencias en que se basan los experimentos sea: 1.º, repetibles; 2.º, intersubjetivas. Si se toma como base de la experimentación psicológica la experiencia de la propia mente —la introspección—, los experimentos no pueden ser intersubjetivos ni repetibles. Por tanto, no se podrá hacer una Psicología científica.

Desde el punto de vista de la metodología científica la mente es opaca a toda investigación directa. Podemos considerarla como una caja negra cuyo contenido es inaccesible. Ahora bien, como la actividad mental necesita siempre de un estímulo o *input* y siempre produce una respuesta o *output*, es posible estudiar las relaciones *input-output* que parecen ser consustanciales con determinados fenómenos psíquicos. Y como tanto los *inputs* como los *outputs* son algo observable por todos, las exigencias metodológicas parecen satisfechas al fijar como objetivo de la investigación las relaciones señaladas, relaciones que recibieron el nombre de conducta o comportamiento.

Esta es la posición metodológica y teórica del conductismo. Skinner y el conductismo era la escuela de Psicología predominante en el mundo académico cuando irrumpió la I. A.

2. ¿Puede pensar una máquina?

En 1950 apareció en Mind un artículo de Alan Turing de título *Computing Machinery and Intelligence*. El trabajo se abre con el siguiente párrafo:

> Propongo que se considere la cuestión "¿Pueden pensar las máquinas?". La discusión debería comenzar por las definiciones sobre el significado de los términos "máquina" y "pensar". Cabría construir tales definiciones de modo que reflejasen en lo posible el uso normal de las palabras, pero esta actitud es peligrosa. Si se ha de encontrar el significado de las palabras "máquina" y "pensar" examinando cómo se las usa habitualmente, es difícil escapar a la conclusión de

> que el significado y la respuesta a la pregunta "¿Pueden pensar las máquinas?", han de buscarse mediante una investigación estadística, como las encuestas Gallup. Pero esto es absurdo. En lugar de buscar una definición de ese tipo, reemplazaré la pregunta por otra que está muy relacionada con ella y se expresa mediante palabras relativamente inequívocas.[1]

Por máquina, nos dice Turing, hay que entender «máquina computadora», un artefacto de esos que podemos comprar en una tienda —aunque entonces no se pudiera—. Si diseñamos un programa que *al ser ejecutado* nos permita jugar al ajedrez contra la máquina, la máquina funcionando bajo el dominio de dicho programa remeda o simula o es un jugador de ajedrez. Análogamente, piensa Turing, si el programa remeda la capacidad de deducir o el habla de un humano, la máquina funcionando bajo el dominio del programa de deducción o bajo el dominio del programa que habla se convierte así en un «deductor» o en un hablante. Si fuéramos capaces de sintetizar en un solo programa todas las funciones de pensamiento y de habla propias de un humano, entonces un hombre que no viera la máquina pero se comunicara con ella no sería capaz de advertir que interacciona con una máquina y no con otro como él.

Turing aventuró[2] que un siglo más tarde se habría logrado una máquina–programa inteligente. Lograr una máquina–programa inteligente fue el *gran reto* que Turing nos legó.

Pero para llegar a concebir que tamaño objetivo se puede conseguir, Turing tuvo que *suponer* que es posible conocer *cómo es el dinamismo* de cada una de las clases de fenómenos mentales —por ejemplo, cuál es el dinamismo del pensamiento cuando deduce—, es decir, que es posible conocer cada una de las llamadas *funciones* mentales. Si esto no fuera posible —como sostuvo el conductismo—, entonces no se podría escribir ningún programa que remedara una función mental.

3. 1956. La Conferencia de Darmouth

«The Dartmouth Summer Research Project on Artificial Intelligence» fue organizada por John MacCarthy a propuesta del propio McCarthy, Marvin Minsky

[1] A. M. Turing «¿Puede pensar una máquina?». Traducción de M. Garrido y A. Antón. Cuadernos Teorema, Valencia, 1974. Pág. 11
B. Jack Copeland (ed.) (2004): *The Essential Turing*. Oxford University Press, Oxford. Pág. 441: «I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.»

[2] «Can automatic calculating machines be said to think?» *in* B. Jack Copeland (ed.) (2004): *The Essential Turing*. Oxford University Press, Oxford. Pág. 495.

(Harvard U.), Nathaniel Rochester (IBM) y Claude Shannon (Bell Telephone Lab.). La idea original consistía en reunir a varios investigadores que aceptaran la conjetura de que la inteligencia podría ser explicada gracias a programas que simularan la conducta inteligente. A los cuatro anteriormente mencionados se sumaron Ray Solomonoff, Oliver Selfridge, Trenchard More, Arthur Samuel, Herbert Simon y Allen Newell.

El objetivo de las sesiones científicas era el estudio «sobre la base de l*a conjetura de que cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, ser descrito con tanta precisión que puede fabricarse una máquina para simularlo.* Se intentará averiguar cómo fabricar máquinas que utilicen el lenguaje, formen abstracciones y conceptos, resuelvan las clases de problemas hasta ahora reservados a los seres humanos, y se mejoren a sí mismas. Creemos que puede llevarse a cabo un avance significativo en uno o más de estos problemas si un grupo de científicos cuidadosamente seleccionados trabajan en ello conjuntamente durante un verano»[3].

Lo que a propósito de Turing señalaba yo, a saber: que la idea de hacer un programa que remede una clase de actos mentales requiere conocer en qué consiste un acto de esa clase, conocer su dinamismo o función, es afirmado a las claras por McCarthy, Minsky, Rochester y Shannon.

4. 1960. Putnam: Minds and Machines

Cuando se habla de mente o de conciencia en cuanto objeto de investigación de una ciencia como la Psicología, se considera la mente como una entidad autónoma e irreductible al cerebro. No se niega que la actividad mental sea actividad cerebral, pero como no toda actividad cerebral es mental, parece que: 1.º, podemos reducir el estudio de la mente a la concreta actividad cerebral que la produce; 2.º, como no contamos con una buena explicación neurológica de esa concreta actividad cerebral que produce la mente, se puede aceptar que hay un λόγος τής ψιχῆς, una psicología.

De la mente, así entendida, se afirma que consiste en actividad incesante: la mente no descansa. Pero no sólo no descansa sino que esa incesante actividad es un continuo [Bergson]. Sin embargo, a fin de estudiarla, hemos convenido en que la actividad mental es divisible en momentos discretos. A cada uno de estos momentos se ha llamado «acto» [Akt, Brentano] y también «estado».

Una vez dividido el continuo mental en actos discretos, se ha caído en la cuenta de que existen clases de actos o estados; por ejemplo, por mucha que sea la diferencia entre estados de cólera —quien lea el Quijote podrá familiarizarse

[3]Las cursivas son mías. J. McCarthy, M. Minsky, N. Rochester y C. Shannon: *A Proposal for the Darmouth Summer Research Project on Artificial Intelligence*, August 31, 1955: «The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.»

rápidamente con ellos—: por su objeto, por su intensidad, etc., todos nos parecen tener algo común: son una emoción intensa que nubla el ánimo y tiende a desatar una conducta violenta —verbal o física o ambas—. Los estados de cólera son muy distintos de aquellos que conforman un diálogo entre dos. Aquí los actos de habla van trenzando una comunicación entre los hablantes.

Brentano propuso una clasificación muy general en dos grandes tipos: primera clase, los actos de representación. Clase que alberga por igual percepciones sensoriales, recuerdos, imaginaciones e intelecciones. Esta clase se caracteriza por el carácter cognitivo de cada acto. En cada acto de representación se nos hace presente algo que, por ello, conocemos o sabemos; segunda clase, los actos que no son de representación. Y no son de representación estos actos porque el objeto al que apunta el acto no es apuntado en cuanto conocido sino en cuanto deseado, creído, juzgado, etc.

Pero Brentano mismo, al establecer su clasificación, ha de hacer uso de términos como «percepción sensorial» (vista, oído, tacto, etc.), «imaginación», «memoria», «entendimiento», «juicio», «deseo», etc. Términos todos usados por la Psicología para referirse a facultades. «Facultad» es un término que entronca casi directamente con la δύναμις aristotélica, esto es, un elemento de la dote natural del hombre consistente en una capacidad, no siempre activa, que cuando actúa produce una actividad continua que concluye con algo otro: ἐντελέχια. Por ejemplo, cuando cerramos los ojos no vemos, pero seguimos poseyendo la capacidad, facultad o δύναμις de ver: basta con volverlos a abrir.

Mediante las facultades la Psicología ha hecho una clasificación de los actos o estados mentales. Así hablamos de actos de visión, estados emocionales, actos de juicio y de pensamiento, etc.

Este ensayo de clasificación sirve, como cualquier otro, para intentar establecer qué es *lo que tienen de común* cada uno de los actos que caen bajo una clase. Ahora bien, eso común a actos o estados de una clase puede ser: *ciertas propiedades o atributos que cada acto parece poseer*. En este sentido, la clasificación de Brentano destaca ciertos rasgos o propiedades; los actos de representación comparten el ser actos de saber, cognitivos por los que nos apropiamos de la información ínsita en el objeto del acto; *o bien, cierto modo de producirse, de llevarse a cabo*. Tenemos una buena descripción de cómo se produce la visión hasta el momento de la integración, aunque aún no comprendamos el fenómeno de integración de la información que proviene de la retina con la que el cerebro pone y completa aquella de la retina, siempre parcial. Gracias a los programas realizados para automatizar el razonamiento deductivo, sabemos que este consiste en transformar expresiones de acuerdo con reglas que no admiten excepciones. *La actividad de una facultad* puede ser contemplada, por tanto, como *la producción de algo*, un producto específico, a partir de ciertos datos o *inputs* no menos específicos. En esta perspectiva, lo que interesa es el dinamismo mismo de la facultad y lo que la Psicología intenta es definirlo en términos de procesos o de relaciones procesuales entre elementos que componen la estructura de la facultad.

En 1960 Hilary Putnam publicó un artículo de título "Minds and Machines" en que estableció la analogía entre estados mentales y estados de una máquina de Turing, de un lado, y entre estados físicos y estados cerebrales, de otro.[4]

Conviene, no obstante, no olvidar que una Máquina de Turing es un *ens rationis*, una abstracción matemática. Lo que define y diferencia a una máquina de otra es lo que se llama *tabla de la máquina*. Una tabla de máquina es una manera fácil y sencilla de definir la función de transición de estado y la función de respuesta. Por ejemplo, la conocida máquina de «paso a la derecha» se define con la siguiente tabla:

ϵ_t	*input*	respuesta	ϵ_{t+1}
0	□	mov→R	1
0	|	mov→R	1
1	□	*stop*	1
1	|	*stop*	1

La tabla define el proceso, el dinamismo procesual de una facultad o máquina: la fila primera establece que si el *input* es un cuadradito vacío y el estado interno de la máquina es 0, entonces el estado interno cambia a 1; además, establece que con los mismos valores en el estado interno y en el *input* la respuesta de la máquina es moverse un cuadradito a la derecha. Obsérvese que el dinamismo de la máquina consiste en establecer relaciones entre elementos de la estructura de la máquina: estado, *input*, respuesta.

Al igual que hay distintas maneras de programar una máquina de *paso a la derecha*, hay distintas maneras de encarnar las facultades una vez entendidas como máquinas de Turing. En el hombre, las facultades mentales han sido encarnadas en el cerebro y una buena parte de la actividad de este consiste en llevarlas a cabo y, con ello, *hacer mente*[5].

[4]Hilary Putnam: "Minds and Machines". En *Philosophical Papers*, vol. 2. Cambridge : N. York, Cambridge University Press, 1975. Pp. 362–385.

«The analogy which has been presented between logical states of a Turing machine and mental states of a human being, on the one hand, and structural states of a Turing machine and physical states of a human being, on the other, is one that I find very suggestive. In particular, further exploration of this analogy may make it possible to further clarify the notion of a "mental state" that we have been discussing. This "further exploration" has not yet been undertaken, at any rate by me, but I should like to put down, for those who may be interested, a few of the features that seem to distinguish logical and mental states respectively from structural and physical ones:», pág. 373.

«La analogía presentada entre los estados lógicos de una máquina de Turing y los estados mentales de un ser humano, por un lado, y entre los estados estructurales de una máquina de Turing y los estados físicos de un ser humano, por otro, me parece muy sugestiva. En particular, la exploración ulterior de esta analogía puede hacer posible aclarar más adelante la noción de "estado mental" que hemos estado discutiendo. Esta "exploración ulterior" no ha sido todavía tomada en cuenta por mi parte, pero me gustaría indicar a los que puedan estar interesados, algunos de los rasgos que parecen distinguir a los estados lógicos y mentales de los estructurales y físicos respectivamente:». "Mentes y Máquinas". Trad. de Purificación Navarro. En A. M. Turing, H. Putnam y D. Davidson: *Mentes y máquinas*. Madrid, Tecnos, 1985, pág. 81.».

[5]Steven Pinker: *Cómo funciona la mente*. Trad. Ferran Meler-Orti. Barcelona, Ediciones Destino, 2007. Pág. 43.

Retengamos, pues, que una cosa es una descripción abstracta de una función o facultad mental y otra la estructura física que la soporta:

> En particular, la "descripción lógica" de una máquina de Turing no incluye ninguna especificación de la *naturaleza física* de estos "estados" ni desde luego de la naturaleza física de la totalidad de la máquina. (Constará de relés electrónicos, de cartón, de empleados humanos sentados frente a escritorios, o de lo que sea). En otras palabras, una "máquina de Turing" dada es una máquina abstracta que puede realizarse físicamente de un casi infinito número de formas diferentes.[6]

Esta distinción, más la analogía señalada entre máquinas de Turing y el dinamismo de una facultad mental, son las aportaciones básicas de este trabajo de Putnam.

5. 1956–1960: Automatización de la lógica

En 1956, en el Darmouth College, se presentó un programa bautizado «The Logic Theorist». Sus autores fueron Newell, Shaw y Simon. Este programa encontró demostración para 38 de los 52 teoremas de *lógica de enunciados* existentes en los *Principia Mathematica*. Incluso, la demostración de uno de ellos es bastante más elegante que la que se encuentra en los *Principia Mathematica*, como reconoció Russell.

En 1960 Hao Wang[7] presentó un sistema de cálculo para la *lógica de enunciados* basado en un cálculo de secuentes —siguiendo ideas de Gentzen—. Al no tener forma axiomática, un sistema de deducción natural puede entenderse como un conjunto de reglas que quintaesencian el conjunto de inferencias legítimas que una mente puede llevar a cabo, al razonar lógicamente. Como es sabido, la lógica de enunciados es un fragmento de la lógica de primer orden con tres propiedades muy importantes: es correcta, completa y decidible. La primera significa que en toda deducción legítima se preserva la verdad. Es decir, que si las premisas son verdaderas la conclusión deductiva no puede no ser verdadera. Y al contrario, por lo que es completa: si existe relación de consecuencia lógica entre un conjunto de premisas y una conclusión, ha de existir una deducción legítima con dichas premisas y dicha conclusión. Por último es decidible. Lo que quiere decir que existe un método capaz de establecer si una hipótesis, consistente en afirmar que de un conjunto de premisas se sigue una conclusión, es verdadera o no. Un ejemplo de esto último es el método de tablas de verdad.

[6] Trad. española pág. 78. *Philosophical Papers*, vol. 2, pág. 371: «In particular, the "logical description" of a Turing machine does not include any specification of the physical nature of these "states" —or indeed, of the physical nature of the whole machine—. (Shall it consist of electronic relays, of cardboard, of human clerks sitting at desks, or what?) In other words, a given "Turing machine" is an abstract machine which may be physically realized in an almost infinite number of different ways».

[7] Hao Wang: "Toward Mechanical Mathematics", IBM Journal of Research and Development 4 (1), 1960: pp. 2–22.

Pero algunos cálculos también son métodos de decisión: el cálculo de árboles analíticos (también llamado «tableaux» semánticos) o el cálculo de secuentes. Una propiedad característica de este tipo de cálculos es que existe una forma inequívoca de determinar que el *proceso de transformación de expresiones* está acabado.

El programa de Wang tiene moraleja, enseñanza profunda. Si nos preguntamos qué hacemos cuando llevamos a cabo una deducción, la respuesta que uno obtiene —meditando sobre el programa— es inquietante: transformar unas oraciones en otras siguiendo reglas cuya aplicación no admite excepciones. De esta meditación emerge una pregunta tan inquietante como todas las que hasta ahora hemos visto. Esta: ¿es posible que los programas nos puedan servir para investigar, conocer, en qué consisten las funciones mentales que remedan? Repárese que hemos invertido la dirección del vector de la investigación, que ya no apunta de la Psicología hacia la I. A. sino justamente al revés, de la I. A. hacia la Psicología.

6. La I. A. en cuanto ciencia

Pasados los primeros años, y creados ya Laboratorios de I. A. en las principales universidades estadounidenses, empezaron a aparecer los primeros ensayos cuyo objetivo era entender y fijar cuál ha de ser el objeto de estudio de la I. A. No hubo duda de que la I. A. es una disciplina dentro de las ciencias de la computación. En este sentido, la I. A. es una disciplina que desarrolla programas y metodología de programas. Pero lo característico es que esos programas tienen el objetivo explícito de ser *modelos* de algunos aspectos de la mente humana.

> Tengo mucho más clara la naturaleza de la inteligencia artificial que los problemas que deben plantearse sobre el concepto de mente. Para determinar esto último, he decidido adoptar un punto de vista ortodoxo de la psicología científica: el de comprender a la mente por medio de la construcción de teorías (si se prefiere, modelos) sobre la conducta de la mente humana, comprobándolas por cualquier método empírico que parezca apropiado, y manteniendo después como "concepto actual de mente" la teoría que mejor consiga superar las pruebas. No se puede conseguir mucho más.[8]

¿Qué se quiere decir con *modelo*? ¿Cómo puede un programa ser un modelo de la mente o de una parte de esta?

En el apartado 4 hemos visto que Putnam estableció un paralelismo entre máquina de Turing y clases de estados mentales, de manera que un algoritmo podría representar el dinamismo de una clase de estados mentales. Al final del apartado 5 hemos visto que el trabajo de Hao Wang obliga a cambiar el vector

[8] Allen Newell: "Artificial Intelligence and the Concept of Mind". En Schank & Colby (eds.): *Computer Model of Thought and Language*, S. Francisco, W. H. Freeman, 1973, pp. 1-60. Traducido por Seoane, J. y Ibáñez, E.: *Inteligencia Artificial y el concepto de mente*, Valencia, Teorema, 1980.

de la investigación que ahora quedaría así: I. A.$\Longrightarrow$Psicología. Esto mismo es parte de la tesis de Newell cuando afirma que los programas de I. A. pueden ser modelos de alguna parte de la mente.

Formulemos una pregunta clara: ¿qué es razonar deductivamente? Es decir, en qué consiste el dinamismo que hay en cualquier razonamiento en que de unas premisas dadas alcanzamos una conclusión mediante sucesivas inferencias deductivas. Y atendamos a dos tipos de programas: la automatización de Wang de la lógica de conectivas y a un sistema experto que transforma números naturales de notación arábiga a la romana[9].

En la figura 1 hay un ejemplo del programa de Wang. Una vez introducidas las fórmulas: $(p \to q), \neg q$ (premisas) y $\neg p$ (conclusión), el programa las escribe en forma de un secuente:

1. $\{(p \to q)(\neg q)\} \Rightarrow \{(\neg p)\}$

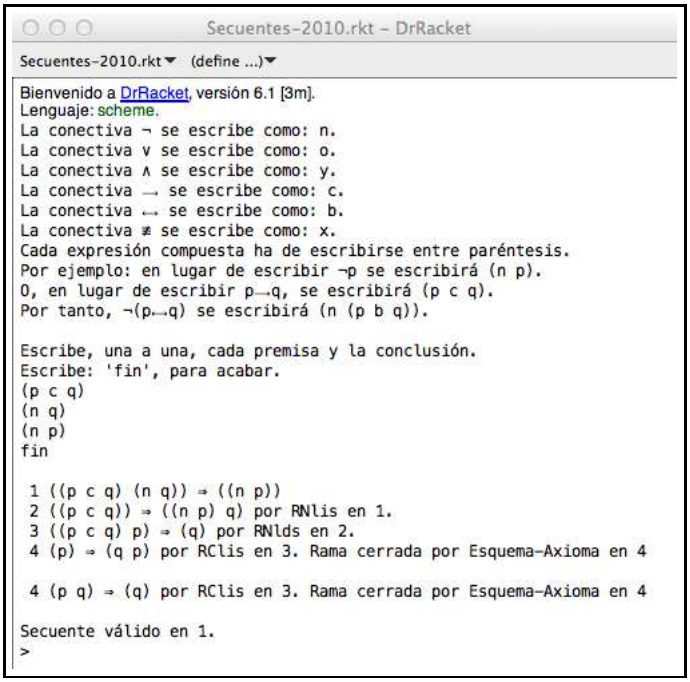

Figura 1: Interfaz del programa de Wang

A continuación, el programa ensaya la aplicación de alguna de las reglas-α, y encuentra que la regla de la negación es aplicable en el lado izquierdo del secuente. Esta regla permite re-escribir el secuente como en la línea 2. La línea 3 resulta de aplicar la regla de la negación en el lado derecho del secuente:

[9]Supuesta una modificación de la antigua notación romana en la que las rayas superiores de los miles se sustituye por un punto de separación entre unidades de mil y centenas.

2. $\{(p \to q)\} \Rightarrow \{(\neg p)\ q\}$ por RN_lis en 1.

3. $\{(p \to q)\ p\} \Rightarrow \{q\}$ por RN_lds en 2.

Las líneas 4 y 5 resultan de una ramificación obligada por aplicación de la regla del condicional en el lado izquierdo del secuente.

4. $\{p\} \Rightarrow \{q\ p\}$ por RC_lis en 3.
   Rama cerrada por esquema de axioma en 4.

5. $\{p\ q\} \Rightarrow \{q\}$ por RC_lis en 3.
   Rama cerrada por esquema de axioma en 5.
   Todas las ramas cerradas. Secuente válido en 1.

La deducción, en este caso, ha consistido en transformar unas expresiones en otras por aplicación de una regla de transformación (o deducción). «Aplicar» significa realizar una transformación de acuerdo con el esquema impuesto por una regla. *Deducir consiste en transformar una expresión en otra de acuerdo con una regla que no admite excepciones.*

Ahora bien, si el programa de Wang es un *modelo del razonamiento deductivo*, entonces será verdad, *ha de ser verdad*, que un hombre también llevará a cabo la deducción de $\neg p$ (conclusión) a partir de $(p \to q), \neg q$ (premisas) transformando unas expresiones en otras por aplicación de unas reglas dadas o esquemas de transformación.

Cualquiera que haya practicado con deducciones de la lógica de conectivas podrá aducir un buen número de experiencias deductivas que, en efecto, consisten en transformar unas expresiones en otras según reglas. Cierto que estas experiencias no constituyen una prueba irrefutable de que deducir sea siempre transformar expresiones, pero ratifican el modelo propuesto.

El siguiente ejemplo que quiero considerar es el de un Sistema Experto. Un S. E. requiere contar con un cierto número de conocimientos sobre un tema. En el ejemplo elegido, sobre cómo traducir o transformar la notación arábiga de un número natural en una notación romana. Ejemplo de este conocimiento es: caso de que el número sea mayor o igual que 900, modificar el valor de la variable de estado «respuesta» añadiendo la ristra «CM» y modificar también el valor de la variable de estado «input» restando 900 del valor que tuviera. Hay un conocimiento específico de este programa consistente en definir la variable de estado «meta»: la meta u objetivo del programa se alcanzará cuando las sucesivas transformaciones del número dado como *input* hagan *input* igual a cero. Es decir, este S. E. cuenta con dos tipos de conocimientos: el de las reglas de la Base de Conocimientos y con ciertas variables de estado. Las variables de estado sirven para controlar el proceso iterativo en que la transformación consiste.

La Base de Conocimientos está formada por *reglas de producción*. Una regla de producción es semejante a una expresión condicional. El antecedente de una regla de producción ha de ser una expresión cuya evaluación produzca un valor

Figura 2: Interfaz del S. E. natural→romano

de verdad. El consecuente ha de ser un *hacer* o *producción*[10]. Una regla de producción de nuestro ejemplo de S. E. es:

$\forall x[(Natural(x) \wedge ((x > 900) \vee (x = 900))) \Rightarrow$
Haz!-respuesta = *respuesta* + *"CM" y Haz!-input* = *input* – *900*]

Los conocimientos que conforman el antecedente son representaciones universales: el concepto de número natural (prescindible puesto que el dominio del que se habla es justamente el de los números naturales) y las relaciones *mayor que* e *igual que*. El consecuente dice lo que se ha de hacer en casos o situaciones en que el antecedente evalúe *verdadero*.

El sistema hace una y otra vez lo mismo hasta alcanzar la meta, que está definida como *input* = *0*. La dinámica del sistema es un bucle iterativo. Las iteraciones acaban cuando *meta* = *verdadero* (*input* = *0*). En cada iteración se transforman los valores de las variables de estado comprobando si el antecedente de una regla se cumple o no. «Comprobar» consiste en una inferencia mínima por aplicación de la regla del *modus ponens*, es decir:

1. $\forall x[((x > 900) \vee (x = 900)) \Rightarrow$ *Haz!-respuesta* = *respuesta* + *"CM" y Haz!-input* = *input* – *900*]
2. $((987 > 900) \vee (987 = 900)) \Rightarrow$ *Haz!-respuesta* = *" "* + *"CM" y Haz!-input* = *980* – *900*] elim$\forall$ en 1, x/input
3. $(987 > 900) \vee (987 = 900)$
4. *Haz!-respuesta* = *" "* + *"CM" y Haz!-input* = *980* – *900* MP 2 y 3

La dinámica del programa es distinta de la dinámica del programa de Wang. Pero hay un elemento o momento del programa en que son iguales: lo que permite aplicar una *producción* en un S. E. es la aplicación de la regla de inferencia del *Modus ponens*, que permite transformar dos expresiones dadas en otra nueva.

[10] En nuestro ejemplo, modificaciones del valor de variables de estado. Pero esto puede cambiar según el tema del sistema.

Este es *el momento deductivo* del proceso dinámico del S. E. Por tanto, también estos programas, los sistemas expertos, son un *modelo* del que se puede extraer la conjetura de que *razonar deductivamente* es sencillamente *transformar expresiones*.

Creo que Newell estaría de acuerdo con este análisis y quizá añadiría: *razonar deductivamente* es *transformar expresiones* con independencia de quien sea el sujeto del razonamiento, hombre o máquina. Cualquiera de esos dos programas, y mejor teniendo ambos en cuenta, es un modelo que explica en qué consiste deducir, con independencia de quién sea el agente de la acción.

Hasta bien entrada la década de 1980 la I. A. estuvo dedicada a los siguientes campos de estudio:

- Automatización de la lógica
- Sistemas expertos
- Solución de problemas en un espacio de estados
- Búsqueda heurística y otros sistemas de búsqueda
- Representación del conocimiento mediante redes semánticas
- Aprendizaje
- Sistemas de creencias y ontologías
- Robótica
- Procesamiento del lenguaje natural
- Representación del conocimiento de sentido común (Marcos y Guiones)

7. La década de 1980: el conexionismo

En 1943 publicaron McCulloch y Pitts un artículo de título «Un cálculo lógico de las ideas inmanentes en la actividad nerviosa». Por primera vez se propone la idea de *neurona artificial*, esto es, de una máquina o algoritmo que representa lo que una neurona biológica tiene de dispositivo de todo o nada al dejar pasar o no una señal eléctrica. Las neuronas artificiales no actúan solas, antes bien actúan en redes de neuronas. A la idea de neurona artificial hay que añadir lo propio de una red: conexiones. Más aún, como sucede con las neuronas biológicas en la sinapsis, la intensidad de la señal que recorre una conexión se debe a la presencia de un factor de multiplicación al que pronto se dio el nombre de «peso». Así que estructura de la red (disposición de las conexiones y de las neuronas), las mismas conexiones, los pesos de las conexiones y las neuronas son las nuevas ideas que se introdujeron progresivamente desde 1943.

Al trabajo de McCulloch y Pitts hay que sumar otro del psicólogo Donald Hebb quien en 1949 publicó *The Organization of Behavior*. Hebb entendía la

Psicología como una ciencia biológica uno de cuyos objetos de estudio había de ser las redes de neuronas que pueblan el cerebro.

En 1957 Frank Rosenblatt desarrolló un prototipo de red bicapa de neuronas artificiales en una computadora IBM 704 y en 1958 publicó "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain".

Marvin Minsky[11] estudió las ideas de McCulloch y Pitts como autómatas de estado finito y junto con Seymour Papert publicó en 1969 otro libro[12] en que sometió a crítica la idea de perceptrón bicapa de Rosenblatt. Mas Minsky y Papert no sólo hicieron una crítica del perceptrón bicapa sino que indicaron cómo llevar a cabo una mejora de la idea de Rosenblatt. Si se intenta crear una red perceptrón de neuronas cuya tarea consista en computar la conectiva lógica xor o disyunción exclusiva, el perceptrón bicapa se muestra incapaz de computar dicha función. Sin embargo, si se añade una capa —desde entonces llamada capa oculta— a las capas de neuronas de entrada y de salida, entonces es posible configurar los parámetros (pesos de las conexiones y umbrales) de forma que la red compute correctamente dicha función veritativa.

La crítica de Minsky y Papert resultó demoledora para Rosenblatt y cualquiera de sus seguidores: los fondos de investigación necesarios para haber seguido con el desarrollo de las redes perceptrón, y de otras formas de redes de neuronas artificiales, se negaron casi sistemáticamente a este grupo de investigadores.[13]

El conexionismo consiguió emerger de nuevo en la segunda mitad de la década de 1980. Apareció entonces lo que se ha considerado «la Biblia» del conexionismo, el libro *Parallel distributed processing*[14]. Libro en que se puede encontrar: 1.º, un marco teórico general para el conexionismo; 2.º, el desarrollo de la idea de redes multicapa de neuronas artificiales, llamadas «perceptrón multicapa»; 3.º, el desarrollo de un algoritmo de entrenamiento supervisado para el perceptrón, conocido como «backpropagation».

Desde entonces hasta hoy el conexionismo disfruta de los fondos de investigación tanto si no más que la I. A. simbolista, la que hemos estudiado hasta este apartado. En los últimos años se ha gestado un ambicioso proyecto de investigación de título "Proyecto Cerebro Humano" (HBP son sus sigla en inglés) en el participan unas 130 universidades de todo el mundo[15] cuyo objetivo es replicar el cerebro humano en una super-computadora. Como dice Markram «se espera que, para 2020, los cerebros digitales puedan representar el funcionamiento interno de una célula del cerebro o incluso del órgano en su conjunto».

[11] *Computation: Finite and Infinite Machines*. Englewood Cliffs (N. J.), Prentice-Hall, 1967.

[12] *Perceptrons: An introduction to Computational Geometry*. Cambridge (MA), MIT Press, 1969.

[13] Especialmente en los Estados Unidos.

[14] David E. Rumelhart, James L. McClelland and the PDP Research Group. *Parallel distributed processing*. Vol. 1: *explorations in the microstructure of cognition*. Vol. 2: *explorations in the microstructure of cognition*. Cambridge, Massachusetts ; London : The MIT Press, 1989.

[15] Cf. Henry Markram: "El proyecto cerebro humano". Investigación y ciencia, Agosto 2012 n.º 431, pp. 50–55.

8. La década de 1990: robots sin representaciones

En las décadas anteriores a 1990 es verdad que la I. A. había intentado desarrollar robots de distinta clase y con distintos propósitos. El planteamiento seguido en la construcción de estos robots partía del principio de que cualquier acción en el mundo requiere de un plan de acción en el agente y que, para elaborarlo, el agente ha de contar con conocimientos sobre el mundo. Según este principio, si queremos construir un robot capaz de moverse incesantemente por los caminos de un jardín como el de la figura 3, hay que proporcionarle una representación mental del jardín —se entienda esto como sea— y de un programa capaz de elaborar un plan de acción en forma de sub-acciones que determinan caminos a seguir.

Figura 3: Caminos en un jardín

Ahora bien, ese planteamiento no sólo tiene un coste computacional muy alto, sino que también suele proporcionar robots bastante ineficientes en la práctica. ¿Es posible cambiar aquel principio, es decir, el principio que dicta elaborar representaciones, conocimientos, antes de actuar?

Rodney A. Brooks en varios trabajos publicados en torno a 1990[16] sostuvo que sí. ¿Cómo?

Primera observación. Juguemos al siguiente juego en el jardín de la figura 3: nos tapamos los ojos y nos proveemos de un bastón tipo el de los ciegos. Aunque alguna vez tendríamos problemas, la mayoría de las veces avanzaríamos por el camino entre dos obstáculos detectados. En ese juego, cada uno de nosotros no necesitaría tener una representación mental del jardín, bastaría con que siguiera

[16]"Intelligence without representation", Artificial Intelligence 47 (1991), 139–159. "Intelligence Without Reason". A.I. Memo No. 1293, April 1991. "Elephants Don't Play Chess". Robotics and Autonomous Systems 6 (1990) 3-15.

una regla sencilla: avanzar hacia aquella posición (izquierda, centro o derecha) en la que el bastón no ha detectado ningún obstáculo; y si acaso en las tres posiciones se detecta un obstáculo, hacer un giro de 90° indistintamente a izquierda o a derechas. ¿Qué conclusión se puede sacar de ese juego? Que la acción en un entorno debe plegarse al entorno.

Segunda observación. El contacto con el mundo necesita de «receptores sensoriales» no sofisticados. En el juego anterior podemos sustituir al hombre por un robot móvil. Para detectar obstáculos no se necesita de grandes aparatos, de cámaras de visión, ni nada por el estilo. Basta con poner al frente una serie de varillas-sensores cuya salida puede agruparse en tres grupos, más o menos correspondientes a izquierda–centro–derecha. Cada sensor dará una u otra respuesta dependiendo de la configuración del entorno.

Tercera observación. La salida de los receptores sensoriales puede ir a parar a los distintos efectores: avanzar-un-paso-a-la-izquierda, avanzar-un-paso-a-la-derecha, avanzar-un-paso-al-centro y girar 90° a izquierdas o derechas. ¿Cuál actuará? Justamente aquel que como una llave encaje en la cerradura que es el esquema o forma de la respuesta de los receptores. Por ejemplo, si la salida de los receptores fuera: 11 00 00 (supuestos seis sensores agrupados en tres de dos), y los efectores respondieran a los siguientes esquemas

efector	esquema
avanzar-un-paso-a-la-izquierda	11 ## ##
avanzar-un-paso-a-la-derecha	## ## 11
avanzar-un-paso-al-centro	## 11 ##
girar	## ## ##

entonces se activaría el efector avanzar-un-paso-a-la-izquierda, mientras que los demás no harían nada. Brooks dice que el efector avanzar-un-paso-a-la-izquierda predominaría, en ese caso, sobre los demás efectores.

El sistema formado por receptores y efectores no necesita de ningún control central ni tampoco de grandes masas de conocimiento acerca del jardín y sus caminos. Por otro lado, mientras que hay un receptor no simple pero único, hay más de un efector. En cada acción predomina un efector sobre el resto. A un sistema formado por varios subsistemas en el que alguno de los subsistemas de salida predomina sobre los demás y cuál sea el que predomine depende de lo percibido en el entorno, es un sistema a cuya estructura llamó Brooks «arquitectura de subsunción».

La arquitectura de subsunción admite que el robot sea más complejo tanto en los receptores como en los efectores e incluso que entre ambos sistemas se interponga otro cuyo objetivo dinámico consista en procesar los distintos *inputs* de forma que al final se tenga una única orden motora.

9. La inspiración biológica

En lo que llevamos de siglo la inspiración predominante, no única desde luego, es la inspiración biológica. La Inteligencia Artificial ha cambiado el foco

de su atención predominante, ya no se fija tanto en cómo se llevan a cabo las distintas funciones mentales y sí presta atención a comportamientos de distintas especies de animales. Por ejemplo, se han desarrollado algoritmos que intentan reproducir el dinamismo de los hormigueros para proveerse de comida.

Lo curioso de este estudio es que ha permitido comprender varias cosas: 1.ª, no hay un plan en ningún rincón del hormiguero al que obedezca que las hormigas salgan en busca de comida y formen hileras de individuos que van y vienen; 2.ª, tampoco hay un control central que sirva para coordinar el trabajo de las hormigas; 3.ª, cada hormiga sigue unos rastros de feromona cuando va en busca de comida y, cuando regresa cargada de comida, deja un rastro de feromona; 4.ª, cada hormiga tiene una dotación natural para seguir rastros de feromonas y para dejar rastro de feromonas, así como la capacidad de portar trozos de comida.

La organización de la provisión de comida en los hormigueros es una característica que: 1.º, es *resultante* de la actividad de múltiples agentes; 2.º, «resultante» quiere decir no planeada; 3.º, si imaginariamente quitamos de cada hormiga la capacidad de seguir un rastro de feromona, entonces desaparecen las hileras de hormigas y, con esto, la posibilidad de proveerse de comida del hormiguero. El hormiguero, respecto a la provisión de comida, se presenta como un sistema complejo cuyas partes, las hormigas, son inseparables —aunque esta o aquella hormiga, si se quiere un buen número de ellas, pueda desaparecer—; y la provisión de comida como un logro resultante de la actividad individual de cada hormiga.

Quizá la idea más básica de esta inspiración sea la siguiente: no sabemos con precisión y exactitud qué es la inteligencia, por qué decimos que la acción de alguien ha sido más o menos inteligente. Aunque no lo sabemos, sí es cierto que estamos convencidos que sin inteligencia, sin acciones inteligentes ni la especie humana ni cada uno de nosotros conseguiría sobrevivir. Por tanto, sea lo que sea la inteligencia hay que suponer que ha de cumplir una función en la economía adaptativa de nuestra especie.

Ahora bien, es sumamente difícil que seamos capaces de desarrollar programas que puedan ofrecer claves para entender la vida del hombre desde la perspectiva de su adaptación. Por ello, se propone el estudio de otras especies menos complejas, incluso de especies de animales artificiales.

La Etología, la disciplina biológica que se ocupa de la conducta animal, ha sido una fuente de inspiración constante para la I. A. bioinspirada. Los etólogos cuentan con la observación de los estímulos y de la respuesta o conducta final de un animal o grupo de animales —en parte o en su totalidad—. Suponen que existe en el organismo del animal bajo estudio la actividad de un órgano o grupos de órganos a los que se debe la unión entre estímulo y respuesta. La conducta animal aparece así como una tripla formada por estímulos, órgano corporal y respuesta. En muchos de los animales bajo estudio el órgano corporal es alguna estructura del sistema nervioso.

En 2008 se publicó en *Nature* un estudio[17] de las larvas de un gusano marino, larvas que forman parte del zooplancton formado por invertebrados marinos. Esquemáticamente dicho, cada larva presenta los siguientes elementos:

- Dos ojos. Cada uno formado por:
  - una neurona fotorreceptora
  - una célula pigmentaria
- Cilios alrededor del cuerpo
- Cada neurona conecta con los cilios más próximos

Cuando la luz estimula uno de los fotorreceptores, se produce un pulso axónico que induce el movimiento de los cilios conectados. El batir de los cilios produce dos movimientos: uno, de rotación axial; otro, de avance helicoidal.

Es posible establecer con máxima precisión cada elemento de la tripla: estímulo, estructura del sistema nervioso y conducta observable. A mi juicio, este trabajo representa el ideal de la metodología de la Etología.

En la I. A. bioinspirada, especialmente en la disciplina de nombre Vida Artificial, el objetivo es construir modelos de sistemas complejos cuyas partes pueden ser, a su vez, sistemas complejos, cuya dinámica resulta de las relaciones definidas entre las dinámicas de las partes. Como ejemplo voy a exponer un modelo, cuyo origen está en un trabajo de Stewart W. Wilson, al que éste bautizó con el nombre *animat*.

10. *Animat*, un animal artificial

Este animal artificial habita en un entorno. El entorno está compuesto, excepción hecha de *animat*, por objetos de distintos colores, entre ellos un blanco que intenta representar lo que vulgarmente entendemos por espacio vacío. El entorno puede ser estático o dinámico: la figura 4 (el rojo es *animat*) pretende dar una idea de esto.

El modelo más sencillo que se puede programar es el de un *animat* cuyo objetivo único es moverse entre los objetos existentes en el entorno. Para construir ese modelo, se dota a *animat* de un órgano sensorial o de contacto con el entorno; de una capacidad de moverse compuesta por dos efectores: uno de cambio de lugar y otro de giro sobre el propio eje; finalmente de una red perceptrón multicapa que toma por entrada la salida producida por el órgano sensorial y produce una orden motora con la que se activa uno u otro de los efectores. Para hacerse cargo del modelo es necesario ver una ejecución del programa.

Las partes de *animat* no tienen una disposición espacial, excepto los receptores sensitivos: «los fotorreceptores, la retina» de *animat* se ubica en la

[17] Gáspár Jékely, Julien Colombelli, Harald Hausen, Keren Guy, Ernst Stelzer, François Nédélec & Detlev Arendt, "Mechanism of phototaxis in marine zooplankton". Nature, Vol. 456, 20 November 2008, pp. 395-399.

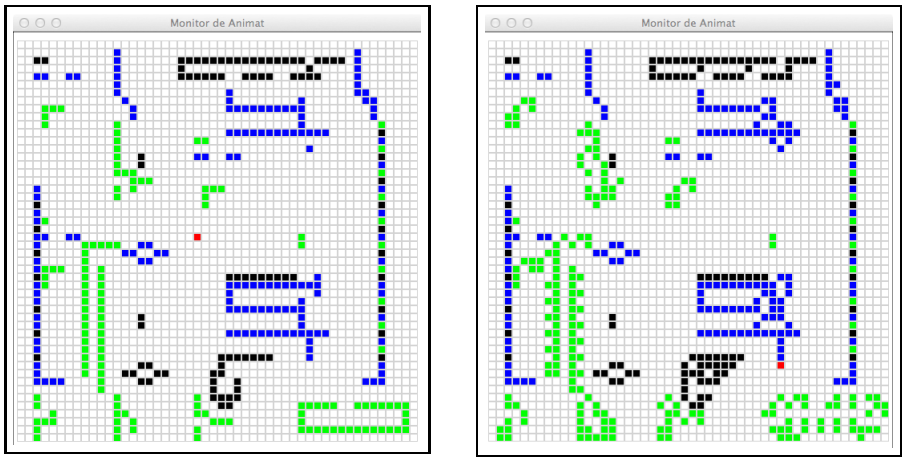

Figura 4: Entornos de *animat*

parte frontal. Sin embargo, es claro que existe una cierta disposición entre todas las partes, puesto que en su funcionamiento *animat* exige que una parte —receptores— actúe antes de que la que la red perceptrón pueda hacerlo. Lo mismo cabe decir de los efectores respecto a la red perceptrón. Desde el punto de vista del funcionamiento interno de *animat*, las partes tienen una disposición funcional, están ordenadas funcionalmente. Por ello, cada parte que compone *animat* es una parte inseparable.

Esta disposición funcional permite determinar cada acción de *animat* como una tripla formada por los receptores, la red perceptrón y los efectores —siempre en este orden—, según el esquema siguiente:

(estímulos, receptores–red perceptrón–efectores, acto efector)

que se puede descomponer en los tres siguientes:

1. (estímulos→receptores sensoriales→salida–sensorial de 15 elementos (1 o 0))
2. (salida–sensorial de 15 elementos (1 o 0)→red perceptrón→salida: orden motora de 4 elementos (1 o 0))
3. (salida: orden motora de 4 elementos (1 o 0)→efectores→acción efectora)

De donde se sigue que cada acción de *animat* responde a la idea de *esquema de conducta* de la Etología. En efecto, el acto de avanzar es relativo a la presencia de un blanco entre los estímulos, mientras que el giro se produce cuando no existe un blanco entre los estímulos. Por esto *animat* es, pues, un objeto compuesto; un sistema funcional de partes inseparables; y un sistema dinámico.

Al igual que en la sección 6 concluiré ahora con una enumeración de algunos campos de estudios dentro de la I. A. bioinspirada:

- Autómatas celulares
- Algoritmos genéticos
- Vida artificial
- Sistemas de aprendizaje
- Sistemas de Lindenmayer
- Sistemas colectivos
- Modelización de animales artificiales
- Modelización de especies animales
- Robótica