Algunas reflexiones sobre el autoconocimiento *

Dr. José E. Burgos**

Resumen

El problema del autoconocimiento es un tema central de la filosofía de la mente. Tanto en una posición cartesiana como en una conductista, el autoconocimiento completo es imposible en principio, por cuanto lleva a una regresión al infinito. Sin embargo, ambas posiciones se concentran sobre el autoconocimiento explícito (o declarativo, o "saber que"). En el presente trabajo, planteo la pregunta de si el autoconocimiento implícito (o ejecutivo o "saber cómo") completo es posible en principio. Para ello, describo un experimento imaginario con redes neurales artificiales del tipo McCulloch-Pitts (MP). En este experimento, un sistema MP recurrente enfrenta la tarea de conocer su propia conducta, donde "conducta" significa "universo conductual", definido como el conjunto de relaciones entrada-salida que el sistema puede realizar, y "conocimiento" significa "capacidad de clasificar algún universo de relaciones entrada-salida". El análisis detallado de un sistema MP con una conexión recurrente muestra que, al menos, el autoconocimiento implícito máximamente fino (y, muy posiblemente, el autoconocimiento implícito completo) es imposible en principio.

Palabras clave: autoconocimiento explícito, autoconocimiento implícito, redes neurales McCulloch-Pitts, experimento imaginario.

* Una versión preliminar de este artículo fue presentada en la 29na. Convención Anual de la Asociación de Análisis de la Conducta, San Francisco, Mayo del 2003. Comentarios y preguntas pueden enviarse al autor, a la siguiente dirección: Av. 12 de Diciembre, #204, Colonia Chapalita (A.P. 5-374), Jalisco, Código postal 45030, México. Correos electrónicos: jburgos@cucba.udg.mx, jeburgos@iteso.mx. Página electrónica: www.ceic.cucba.udg.mx. Agradezco al Dr. Gustavo Peña por haberme invitado a escribir este artículo para *Analogías del comportamiento*.

** Centro de Estudios e Investigaciones en Comportamiento. Universidad de Guadalajara.

Emplazado en la ladera sur del Monte Parnaso, en lo que fuera la antigua ciudad griega de Phocis, solía hallarse el Oráculo de Delfos, cuyo pórtico, donde se erigía una escultura de Homero, estaba adornado con inscripciones de los preceptos de los Siete Sabios de Grecia. Uno de los más famosos preceptos reza "Conócete a ti mismo", atribuido al gran jurista ateniense Solón (ca. 638-558 A.C.). Tal precepto, tan misterioso y ambiguo como los consejos del oráculo, es cimiento de la posibilidad misma de la autoayuda, tema que, por demás, constituye una porción nada despreciable de la Literatura psicológica contemporánea. El presente artículo trata sobre el autoconocimiento, visto desde una perspectiva muy distinta de la oficial. Permítaseme, entonces, examinar dicha doctrina, a manera de introducción.

El autoconocimiento es un tema central en la Filosofía de la mente. En ese contexto, "autoconocimiento" usualmente se refiere al conocimiento de los estados mentales (creencias, deseos, sensaciones) propios (es decir, de uno mismo). Según la doctrina oficial, el autoconocimiento es un tipo fundamentalmente distinto y especial de conocimiento, tanto en su objeto como en su método. Esta concepción adquirió su forma más explícita en la obra de Descartes (1596-1650). Por ejemplo, en sus *Meditaciones sobre la Filosofía primera* (publicado originalmente en latín en 1641, y mejor conocidas como *Las meditaciones metafísicas*), encontramos su famoso análisis:

> *... ¿cómo sé que no hay algo más que no permita la más mínima oportunidad para la duda? ¿Es que no hay un Dios, o como sea que pueda llamarlo, que coloca dentro de mí los pensamientos que estoy teniendo en este momento? Pero ¿por qué pienso esto, si yo mismo puedo ser el autor de estos pensamientos? En ese caso, ¿no soy, al menos, algo? Pero acabo de decir que no tengo ni sentidos ni cuerpo. Este es el punto problemático: ¿qué se sigue de esto? ¿No estoy tan ligado a mi cuerpo y mis sentidos que no puedo existir sin ellos? Pero me he convencido a mi mismo de que no hay absolutamente nada en el mundo, ni cielo ni tierra, ni mentes, ni cuerpos. ¿No se sigue de ello que yo tampoco existo? No: si me convencí a mí mismo de algo entonces ciertamente existo. Pero hay alguien, de un poder y astucia supremos, que me engaña deliberada y constantemente. En ese caso, también resulta indudable que existo, si en efecto me*

> *está engañando; y así me engañe tanto como pueda, jamás podrá hacer que yo sea nada, siempre y cuando yo piense que soy algo. Así que, luego de considerar todo muy cuidadosamente, debo finalmente concluir que esta proposición,* "soy, existo", *es necesariamente verdadera cada vez que es afirmada por mí o concebida en mi mente*[1].

El argumento cartesiano parece intuitivamente claro: todo lo que piensa, existe (aunque, por supuesto, no todo lo que existe piensa); yo pienso; por lo tanto, yo existo. Para Descartes, pensar es, fundamentalmente, dudar. El acto mismo de la duda es mental. Lo que no existe no puede dudar, pero yo dudo; por lo tanto, necesariamente existo[2].

El párrafo aparece en una sección titulada *Segunda meditación: La naturaleza de la mente humana, y cómo es mejor conocida que el cuerpo*. Tal título resume la propuesta de que el conocimiento de nuestra propia mente (o "alma" o "espíritu" o como quiera llamársele) es en algún sentido más autoritario que el conocimiento de nuestro propio cuerpo. ¿Pero en qué sentido? La respuesta de Descartes es igualmente explícita: Él no puede dudar que está dudando. Bajo esta doctrina, tenemos un *acceso privilegiado* a nuestra mente, un contacto tan íntimo con ella que el autoconocimiento se hace absolutamente *certero*, es decir, epistémicamente *infalible*. Bajo esta concepción, entonces, nadie puede equivocarse acerca de sus propias creencias, deseos y sensaciones, al menos respecto a que *está teniendo* tales creencias, deseos o sensaciones[3].

---

1 Traducción libre mía de la traducción inglesa de Cottingham, Stoothuff y Murdoch (1985), reimpresa en Rosenthal (1991, p. 21).

2 Russell (1912, p. 19) señaló que este argumento debe ser leído con mucho cuidado, ya que supone más de lo debido. El problema reside en el uso del término "Yo", el cual implica la existencia de un ser que permanece en el tiempo. Obviamente, ni siquiera Descartes estaba dudando a cada instante de su vida (e.g., en su infancia, cuando dormía, etc.). Si la duda es lo que garantiza la existencia, Descartes no existía cuando no estaba dudando. El cartesiano sólo puede decir que en este instante una substancia mental está dudando y, por tanto, existiendo.

3 Por supuesto, uno puede estar equivocado o desconocer las *causas* de los estados mentales propios, pero ése es otro problema. También, es posible estar inseguro acerca de si uno cree, desea o siente algo, pero en ese caso no habría estado mental alguno acerca del cual estar equivocado. La posibilidad misma de equivocarse acerca de un estado mental propio implica que uno se encuentra en un estado particular.

Desde esta perspectiva, no puede haber tal cosa como una creencia falsa acerca de los propios estados mentales. Si creo firmemente que estoy sintiendo un dolor, es imposible que no lo esté sintiendo (independientemente de sus causas). Si creo que creo firmemente en Dios, es imposible que no crea en Dios. Si creo que realmente deseo comerme un chocolate, es imposible no tener el deseo en cuestión, y así sucesivamente. En suma, creer que uno se encuentra en un cierto estado mental es saber que uno se encuentra en ese estado[4]. En el autoconocimiento, y a diferencia de cualquier otro tipo de conocimiento, creer es saber[5]. Puede haber creencias falsas acerca de todo lo que existe y ocurre en el mundo externo, pero no acerca de los estados mentales propios.

La doctrina cartesiana de la infalibilidad epistémica del autoconocimiento implica que uno es la autoridad máxima respecto al conocimiento de sus propias creencias, deseos y sensaciones. Tal implicación se encuentra profundamente enraizada en el lenguaje cotidiano y el sentido común. Estamos tan convencidos de ello, que tendemos a ofendernos cuando se nos dice que estamos mintiendo al respecto. 'Tu no sabes lo que siento' es una expresión muy frecuente en discusiones acerca de los problemas personales, y lo máximo que pueden hacer los psicólogos clínicos es *inferir* qué sienten (o creen o desean) sus clientes, a través del lenguaje.

---

4 Siguiendo a Descartes, yo puedo dudar que está lloviendo, pero no que creo que está lloviendo. Mi creencia acerca de que está lloviendo puede ser falsa, pero no mi creencia acerca de que yo creo que está lloviendo.

5 Esta doctrina adopta la noción filosófica tradicional de conocimiento como *creencia verdadera justificada*. Si aplicamos esta noción al autoconocimiento, creer que se está en un cierto estado mental es suficiente (aunque quizás no necesario) para realmente estar en ese estado (i.e., para que dicha creencia sea verdadera). Por supuesto, queda abierto el problema de la justificación. Si yo creo que me duele la cabeza, no puedo estar equivocado al respecto (i.e., es necesariamente verdadero que me duele la cabeza). En principio, ante la pregunta de por qué creo que me duele la cabeza, uno simplemente respondería "porque lo siento". Tal respuesta ofrece el estado mismo como justificación de la creencia. Entonces, pareciera que los estados mentales constituyen creencias fundamentales, que no requieren justificación más allá de sí mismas, precisamente el tipo de creencias que los dogmáticos (empiristas y racionalistas, y, entre estos últimos, Descartes) buscaron para detener la regresión al infinito planteada por los escépticos, la cual amenazaba la posibilidad misma del conocimiento, visto precisamente como creencia verdadera justificada. Más adelante, colocaré estas consideraciones en un contexto más amplio que llevará a la pregunta central del presente análisis.

La doctrina de que el autoconocimiento es fundamentalmente distinto del conocimiento del mundo externo es una consecuencia inevitable del dualismo cartesiano. Como es bien sabido, tal dualismo, a diferencia de aquellos propuestos por Leibniz (1714) y Malebranche (1688), es *interaccionista* y, a diferencia del dualismo de Spinoza (1677), es un dualismo de *substancias*[6]. Bajo el dualismo cartesiano, las substancias se dividen exhaustivamente en dos tipos fundamentalmente distintos: substancias físicas y substancias mentales. Las substancias físicas son inherentemente públicas, mientras que las mentales son inherentemente *privadas*. Los eventos que involucra cualquier substancia mental, pues, pueden ser conocidos directamente sólo por un sujeto: precisamente, aquel cuyo cuerpo interactúa (en la glándula pineal) con la substancia mental en cuestión. Es el carácter privado de lo mental lo que lo hace accesible por parte de uno y sólo un sujeto[7]. De allí que lo

---

6 Para entender las diferencias básicas entre los tres tipos de dualismo, hay que entender a qué se refieren los términos "interaccionista" y "substancia". Por interaccionista se quiere decir que la mente y el cuerpo se afectan mutuamente (los eventos mentales son tanto efecto como causa de los corporales). Aclarar el significado de "substancia" es más difícil, debido al carácter fuertemente equívoco del término en metafísica. En aras de la brevedad, tomaré el término en su sentido aristotélico de "particular no analizable", donde "particular" es sinónimo de "individuo", en el sentido metafísico de "entidad o ser único e irrepetible". Los particulares constituyen una categoría ontológica fundamental. Otra categoría está constituida por las *propiedades*, las cuales pueden ser entendidas como universales (entidades abstractas, ejemplificadas de manera múltiple) o como tropos (propiedades particulares). Entonces, el dualismo cartesiano es una partición del ámbito de los particulares en mentales y físicos. Los dualismos de Leibniz (conocido como "teoría de la harmonía preestablecida") y Malebranche (conocido como "ocasionalismo") también son dualismos de substancia, pero no interaccionistas. Por su parte, el dualismo de Spinoza (conocido como "teoría del aspecto dual") es interaccionista de propiedades. Hoy en día, virtualmente ningún filósofo de la mente es dualista de substancias. Sin embargo, la mayoría de los filósofos contemporáneos de la mente son dualistas de propiedades.

7 Ello plantea otro tema de la filosofía de la mente, a saber, el problema de otras mentes: ¿Cómo sé que existen otras mentes diferentes de la mía? Una solución extrema a este problema es el solipsismo, que consiste en negar la existencia de otras mentes diferentes de la mía. Hume (1748) admitió que la razón inevitablemente nos lleva al solipsismo como la posición más racional. Sin embargo, Hume también arguyó que el solipsismo es una doctrina vacía, en el sentido de que ningún ser humano es capaz de ser solipsista, ni siquiera por un instante. Según Hume, nos es imposible, por naturaleza, dejar de creer en lo que la razón nos dicta que no creamos (en la existencia de una realidad independiente de nosotros, en que el futuro será como el pasado, etc.). En efecto, en manos de Hume, los humanos nos volvemos seres esencialmente irracionales.

que se conoce en el autoconocimiento sea fundamentalmente distinto de lo que se conoce en el conocimiento del mundo físico, y que requiera un método especial. Tal método consiste, metafóricamente, en "mirar hacia adentro", es decir, llevar a cabo una *introspección*[8].

En resumen, bajo la doctrina cartesiana, el autoconocimiento es especial, en el sentido de que, a diferencia del conocimiento del mundo externo, el primero es epistémicamente seguro y completo, su objeto (estados mentales) y método (la introspección) son fundamentalmente distintos, y la persona es la máxima autoridad respecto al conocimiento de sus propios estados mentales. Por supuesto, el carácter oficial de esta doctrina no significa que sea unánimemente aceptada.

En oposición a ella encontramos, por ejemplo, a Ryle, quien, en su obra clásica *The concept of mind* (1949), le dedica todo un capítulo (el número VI, titulado, precisamente, "Auto-conocimiento") al problema. En este capítulo, Ryle arguye que el autoconocimiento no es un tipo fundamentalmente distinto de conocimiento, ni en su objeto ni en su método[9]. El objeto del autoconocimiento es tan físico como los objetos del conocimiento del mundo externo. Específicamente, autoconocerse, según Ryle, es conocer lo que uno mismo ha hecho y tiende a hacer en situaciones de distintos tipos. En terminología ryleana, autoconocerse es conocer las ocurrencias y disposiciones conductuales de uno mismo. De hecho, autoconocerse también es conducta, en el sentido de que es la ocurrencia o tendencia a describir lingüísticamente lo que uno ha hecho y las condiciones bajo las cuales lo ha hecho. Según Ryle, entonces, el autoconocimiento es *retrodicción en primera persona*.

Bajo el esquema ryleano, pues, autoconocerse es conducta, lo que conocemos en el autoconocimiento es conducta, y la manera en la cual nos conocemos a nosotros mismos es mediante la observación de nuestra

---

8 Etimológicamente, 'introspección' se deriva del latín *introspic_re* ("mirar adentro"). Se cree que el término fue utilizado por primera vez en la segunda mitad del siglo XVII, luego de la muerte de Descartes, aunque se desconoce exactamente quién lo acuñó. Es muy probable que el término haya sido acuñado para referirse justamente al proceso involucrado en la duda cartesiana.

9 Skinner (1974, pp. 30-2 y 168-171) repite estas ideas, pero sin citar a Ryle.

propia conducta, observación que no es fundamentalmente distinta de la observación de la conducta de otros o de cualquier otro evento u objeto del mundo externo. Hay varias formas de observar la propia conducta. Por ejemplo, uno podría mantener un registro del número de ocurrencias de cierto tipo de actividad bajo cierto tipo de condiciones ambientales. Uno puede saber si uno es un fumador registrando el número de veces que uno fuma al día y bajo qué condiciones ambientales. Esta noción se encuentra en el análisis conductual, descrita como técnica para controlar la propia conducta (e.g., Skinner, 1953, pp. 260-282).

Un aspecto importante de la crítica de Ryle a la doctrina cartesiana es su argumento en contra de la introspección. Según Ryle, la introspección desencadena una regresión al infinito, ya que ¿cómo sabe uno que está "mirando adentro"? Mirar adentro es, en sí mismo, un acto mental, cuya ocurrencia debería también poder ser autoconocida. Pero si el método del autoconocimiento es la introspección, entonces podemos saber que estamos "mirando" adentro sólo mediante la introspección. Saber que uno se está "mirando" adentro (que obviamente es parte de conocerse a uno mismo), pues, requiere de un segundo acto de introspección. Sin embargo, ¿cómo sabemos que tal acto está ocurriendo? Pues sólo podemos saberlo mediante un tercer acto de introspección, y así sucesivamente, *ad infinitum*. *En principio*, pues, es imposible autoconocerse *completamente* por medio de la introspección[10]. Siempre habrá un acto mental *consciente* incognoscible[11].

---

10 Enfatizo "en principio" para señalar que la imposibilidad en cuestión es estrictamente lógica, no circunstancial. Por supuesto, hay innumerables circunstancias que imposibilitan el autoconocimiento completo. Por ejemplo, resulta difícil autoconocerse durante los primeros meses de vida, mientras se duerme, o, en casos más trágicos, si se tiene un daño cortical severo. A ello podemos añadir el ámbito de los actos inconscientes (ver siguiente nota). Sin embargo, todos estos casos son circunstanciales en el sentido de que si no existieran, aun sería imposible autoconocerse completamente mediante la introspección.

11 Enfatizo "consciente" para excluir los estados mentales inconscientes, que también son (supuestamente) incognoscibles. La introspección es un acto estrictamente consciente, así como lo son todos aquellos estados mentales que son objeto de la introspección. El presente análisis, pues, cae fuera del ámbito de lo inconsciente.

Resulta tentador suponer que la crítica de Ryle a la introspección no se aplica a su propia concepción del autoconocimiento. Sin embargo, tal suposición sería errónea. En efecto, describir la propia conducta también es conducta, cuyo conocimiento requeriría de una descripción de segundo orden (la descripción de la descripción), y así *ad infinitum*. La crítica de Ryle se aplica igualmente al autoconocimiento conductual. El autoconocimiento *completo*, entonces, es imposible en principio, independientemente de si adoptamos una posición cartesiana o una ryleana. La máxima "Conócete a ti mismo" no puede ser cumplida de manera final y definitiva. Sólo podemos conocernos a nosotros mismos de manera incompleta (de nuevo, aun si no existieran las circunstancias que impiden un autoconocimiento completo, en especial los estados y actos inconscientes; ver Notas 10 y 11).

No obstante, cabe preguntar cuán general es la imposibilidad en principio del autoconocimiento completo. Esta cuestión se hace importante si tomamos en cuenta la distinción entre "saber que" y "saber cómo", o entre conocimiento *explícito* (o declarativo) y conocimiento *implícito* (o ejecutivo), respectivamente. Bajo la distinción estándar en ciencia cognitiva, el conocimiento explícito involucra de manera crítica el uso declarativo del lenguaje. El conocimiento implícito, por su parte, es estrictamente ejecutivo en el sentido de que no requiere al lenguaje como condición necesaria.

Saber *que* $2 + 2 = 4$, *que* la Tierra es (aproximadamente) esférica, *que* $E = mc^2$, *que* está lloviendo, *que* Guadalajara es la capital de Jalisco, etc., son ejemplos de conocimiento declarativo. Por su parte, saber *cómo* tocar el bongó, *cómo* jugar ajedrez, *cómo* armar un rompecabezas, *cómo* presionar un botón o una palanca, *cómo* fumar, *cómo* comer, *cómo* caminar, *cómo* manejar una bicicleta, *cómo* nadar, etc., son ejemplos de conocimiento implícito.

Ejemplos provenientes de la conducta humana son problemáticos, ya que, en su mayoría, no permiten una separación precisa entre lo declarativo y lo implícito. La conducta humana está tan envuelta en lenguaje, que resulta difícil encontrar ejemplos puros de conocimiento

implícito en humanos[12]. Quizás, los ejemplos más puros se encuentran en la conducta refleja, y aun allí, no es fácil deslindarse completamente del lenguaje. El conocimiento implícito es ejemplificado de manera más pura por los organismos no lingüísticos. Los peces saben cómo nadar y las aves cómo volar, pero (supuestamente), no saben que están nadando o volando (mucho menos que $2 + 2 = 4$, $E = mc^2$, que Guadalajara es la capital de Jalisco, o que la aceleración es la segunda derivada del cambio en la distancia como función del tiempo). Nadar en peces y volar en aves, pues, califican como conocimiento implícito puro, en el sentido de que (supuestamente) no involucran lenguaje en lo absoluto.

Tal y como lo anticipé en la Nota 5, el autoconocimiento, visto desde la doctrina cartesiana, es estrictamente declarativo, ya que adopta la noción de conocimiento como creencia verdadera justificada. Esta noción se restringe al conocimiento declarativo, por cuanto plantearse cuestiones de verdad y de justificación adquiere sentido sólo en referencia tal conocimiento[13]. El conocimiento implícito no se presta fácilmente a un análisis en términos de verdad y justificación. En efecto, en la doctrina cartesiana, uno sabe *que* le está doliendo la cabeza, *que* cree en Dios, *que* desea comer chocolate, etc.

Lo mismo puede decirse sobre la doctrina ryleana: uno sabe *que* uno tiende a comportarse de tal o cual forma bajo tales o cuales circunstancias. Si bien, el objeto del autoconocimiento en esta doctrina

---

12 En efecto, podría argüirse que hablar y escribir también califican como conocimiento implícito, ya que involucran saber *cómo* realizar ciertos actos (*cómo* mover la boca y lengua, y *cómo* mover la mano de cierta forma, etc.). En este sentido, conocimiento explícito pareciera ser reducible al implícito. Sin embargo, también podría argüirse que califican como conocimiento explícito, por cuanto involucran (al menos en el español) saber, por ejemplo, *que* los artículos van al principio de ciertas frases, *que* tal o cual verbo se conjuga de tal o cual manera, *que* en ciertas palabras se acentúa tal o cual letra, *que* el sujeto viene antes del predicado en ciertas frases, etc. Es decir, hablar y escribir involucran el conocimiento de las reglas gramaticales del lenguaje en el cual se pretende hablar o escribir, y tal conocimiento es explícito. Resulta claro que la presencia del lenguaje complica enormemente el análisis.

13 De nuevo, el trabajo de Descartes se encuentra claramente enmarcado en los esfuerzos por identificar un conjunto de creencias fundamentales que permitieran detener la regresión al infinito planteada por los escépticos, la cual amenazaba la posibilidad misma del conocimiento, justamente entendido como creencia verdadera justificada.

involucra en gran medida conocimiento implícito (e.g., yo sé *que* sé *cómo* tocar la guitarra y sé *que* sé *cómo* jugar al ajedrez), el conocimiento de ese conocimiento implícito es declarativo.

Cabe, entonces, plantear la siguiente pregunta: ¿Es la regresión al infinito desencadenada por el autoconocimiento consecuencia del carácter declarativo del mismo? Es decir, ¿está la imposibilidad en principio del autoconocimiento circunscrita al autoconocimiento explícito, o también se da en el autoconocimiento implícito? Más brevemente, ¿es posible en principio el autoconocimiento implícito completo? Esta es la pregunta que deseo responder en el presente artículo.

Para responderla, echaré mano del recurso analítico conocido como "experimento imaginario". Los experimentos imaginarios son frecuentemente utilizados en Ciencia y Filosofía. Según Kuhn (1964), varios experimentos imaginarios en ciencia han llevado a revoluciones científicas. A pesar de ello, lo imaginario de estos experimentos ha hecho que gocen de mala reputación. Su carácter imaginario tiende a ser considerado como arbitrario, irreal, meramente mental (en efecto, introspectivo). Sin embargo, tal y como lo arguye Sorensen (1992), la gran mayoría de las críticas en contra de los experimentos imaginarios son arbitrarias (véase también Brown, 1991).

El experimento en cuestión involucrará un sistema neural artificial del tipo propuesto por McCulloch y Pitts (1943). En dicho experimento, el sistema enfrenta la tarea de conocer su propia conducta. Estos sistemas resultan ideales, ya que son muy simples, lo cual permite un nivel considerable de claridad y precisión. Por supuesto, no pretendo generalizar el resultado de mi análisis más allá de este tipo de sistemas. Tal generalización se justificará en la medida en que análisis ulteriores con otros tipos de sistemas, tanto artificiales como naturales, lleven a resultados comparables.

El presente análisis es una continuación de otro en el cual examiné la relación entre la complejidad estructural relativa de dos sistemas McCulloch-Pitts (MP) y la posibilidad de que uno de ellos pudiera conocer, de manera completa y máximamente fina, la conducta del otro (Burgos,

2002). Específicamente, demostré que existía al menos un par de sistemas MP tales que poseían la misma complejidad estructural y ninguno de ellos podía conocer de manera completa y máximamente fina la conducta del otro. Un corolario de este resultado es que todo sistema MP es igualmente complejo a sí mismo y que, por lo tanto, tampoco podía conocer su propia conducta de manera completa y máximamente fina. Mi objetivo en éste artículo es examinar en mayor detalle esta posibilidad, como punto de partida hacia una determinación de si la regresión al infinito desencadenada por el autoconocimiento explícito también se da en el autoconocimiento implícito.

En la primera sección, describo la teoría MP y defino conceptos de conducta y conocimiento que sean adecuados para sistemas MP. En la segunda sección, describo el tipo de sistema MP que analizaré y la tarea que enfrentará. En la tercera sección, llevo a cabo un análisis detallado del escenario descrito en la segunda sección. En la cuarta sección discuto las implicaciones del análisis.

El análisis involucrará un grado considerable de simplificación, relativo al ser humano (de hecho, relativo al organismo natural más simple que pueda encontrarse). La simplificación es sólo un medio para aumentar la rigurosidad, característica que brilla por su ausencia en las discusiones sobre el autoconocimiento. El lector que no esté acostumbrado a tal grado de rigurosidad deberá armarse de paciencia, si desea sacarle provecho a este análisis. Sin embargo, la paciencia es una virtud. Los psicólogos nos hemos malacostumbrado a buscar soluciones rápidas a problemas que muy bien pueden no admitirlas. Estoy convencido de que el problema del autoconocimiento (y el resto de los problemas en psicología) no admite una solución ligera. Parte de lo que busco en este artículo, entonces, es causarle al lector una especie de choque cultural que le haga darse cuenta de cuán inesperadamente complejo puede llegar a ser un problema, si tenemos la suficiente paciencia para tratarlo con rigurosidad y elaborar sus detalles.

El Modelo MP

El modelo MP fue el primer modelo de conducta adaptativa (o, más ambiguamente, "inteligencia") inspirado en la estructura y funcionamiento del sistema nervioso. El modelo describe la estructura y funcionamiento de un elemento procesador que representa una especie de neurona artificial. La Figura 1 muestra un elemento MP genérico. Informalmente, el elemento consiste de un número finito n de unidades de entrada (sensores, receptores, o, como los denominaron McCulloch y

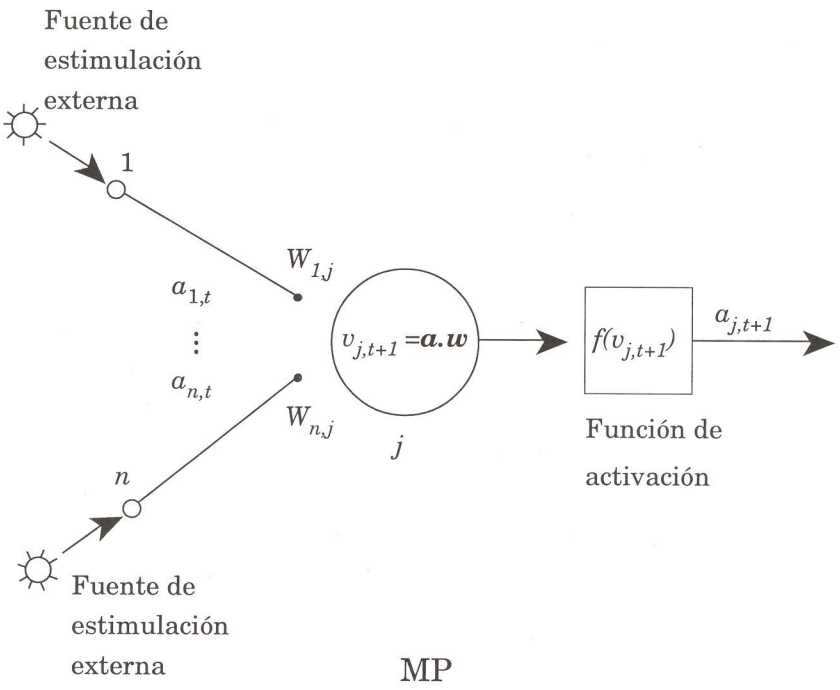

Figura 1. Elemento genérico MP. Los sensores, círculos pequeños blancos marcados como 1, ..., n, reciben señales de fuentes de estimulación externas. Estas señales de entrada ($a_{1,t}$,..., $a_{n,t}$) son transmitidas al elemento procesador (j) a través de conexiones (círculos pequeños negros) con pesos $w_{i,j}$, ..., $w_{n,j}$. En el siguiente momento, el elemento procesador calcula el producto interno ($v_{j,t+1}$) entre el vector de activaciones de entrada (**a**) y el vector de pesos (**w**). El resultado de este cómputo es pasado como argumento a una función de activación (f), la cual retorna la salida del sistema ($a_{j,t+1}$).

Pitts, "aferentes periféricos"). Los sensores (círculos pequeños blancos etiquetados como 1,..., *n*) son activados por cambios en el medio ambiente local del elemento. En la jerga de la psicología del aprendizaje, las activaciones de los sensores representan *estímulos*. Las activaciones de los sensores en un momento *t* ($a_{1,t}$,..., $a_{n,t}$), son transmitidas a una unidad de cómputo (círculo grande etiquetado como '*j*'). Esta unidad realiza ciertos cómputos (ver más adelante) y emite una señal de salida $a_{j,t+1}$, la cual representa el estado de activación del elemento *j* en *t* + 1 (el siguiente momento; ver más adelante). En la jerga de la psicología del aprendizaje, la señal de salida representa una *respuesta*.

Siguiendo la Ley del Todo o Nada, el modelo asume que toda señal (tanto de entrada como de salida) *es binaria*, es decir, que en un cierto momento puede adquirir uno y sólo un valor, de dos valores posibles (o 1 o 0). El estado de activación de *j* en *t* + 1 depende de cuántos de sus sensores estén activados en *t*, cuán fuertemente están conectados al elemento, y el umbral de *j* (q). En la figura, las conexiones están representadas por los círculos pequeños negros y representan sinapsis. La fuerza de una conexión viene dada por un valor numérico que representa el *peso* de la contribución de una señal de entrada $a_{i,t}$ a $a_{j,t+1}$. El peso de la conexión que va del elemento *i* al elemento *j* se denota como $w_{i,j}$. Esta notación no incluye un subíndice temporal debido a que, en sistemas MP, cualquier peso dado permanece constante en el tiempo (i.e., no hay "aprendizaje"), aunque diferentes conexiones pueden tener diferentes pesos y éstos pueden ser negativos, cero o positivos. El valor de θ permanece constante, no sólo en el tiempo, sino también de un elemento a otro, y también puede ser negativo, cero o positivo.

El funcionamiento de un elemento MP se describe matemáticamente en términos de una regla para calcular su estado en *t* + 1, como función de las señales de entrada que recibe y los pesos de las conexiones respectivas. La regla es una función de umbral que especifica las condiciones bajo las cuales el elemento está activado o desactivado. De acuerdo con las descripciones estándar, el elemento se activará si la combinación lineal de sus señales de entrada con sus pesos corres-

pondientes ($v_{j,t+1}$) alcanza o excede θ. De lo contrario, el elemento se considerará como desactivado. Es decir:

$$\text{Si } v_{j,t+1} \geq \theta, \text{ entonces } a_{j,t+1} = 1; \text{ si no, } a_{j,t+1} = 0,$$

donde $v_{j,t+1}$ denota la combinación lineal de señales de entrada y sus pesos correspondientes. En la Figura 1, $v_{j,t+1}$ es definida como el producto interno entre el vector **a** de activaciones de entrada y el vector **w** de pesos. Algebraicamente, $v_{j,t+1}$ es una suma ponderada de activaciones (i.e., una suma de productos entre activaciones y pesos):

$$v_{j,t+1} = \sum_{i=1}^{n} a_{i,t} w_{i,j}$$

La diferencia entre los subíndices $t + 1$ y t significa que tanto el producto interno entre el vector de activaciones de entrada y el vector de pesos, como la activación misma del elemento procesador, tienen lugar un momento después de que los sensores han sido activados. Es decir, hay una *demora* (o, en la jerga de la psicología del aprendizaje, una *latencia*) de un momento entre las entradas y la salida. Ello se debe a que el modelo asume que las conexiones son la única fuente de demora en un sistema MP, tal que cada conexión por la cual tenga que pasar una señal introduce una demora de un momento.

Un aspecto central del presente análisis es la definición de un concepto de conducta que, por una parte, sea razonablemente intuitiva y que, por otra, sea aplicable a sistemas MP. Propongo, entonces, definir la conducta de un sistema MP en términos de su *universo conductual*, donde un sistema MP es un sistema constituido por uno o más elementos MP, con uno o más sensores. En general, el universo conductual de un sistema MP es el conjunto de relaciones entrada-salida (o estímulo-respuesta) que son *lógicamente* (en contraposición a teóricamente) posibles para el sistema. De este modo, el concepto de conducta de un

sistema MP es estrictamente relacional, en el sentido de que se refiere a relaciones entre el ambiente (entradas, estímulos) y el responder (salidas) del sistema. Para ejemplificar este concepto, tomemos el sistema MP mostrado en la Figura 2, al cual llamaré MP_C ('C' de 'conducta' y de 'saber cómo'). MP_C consta de dos sensores (círculos blancos pequeños marcados como '1' y '2') y un elemento procesador (círculo blanco grande marcado como '3').

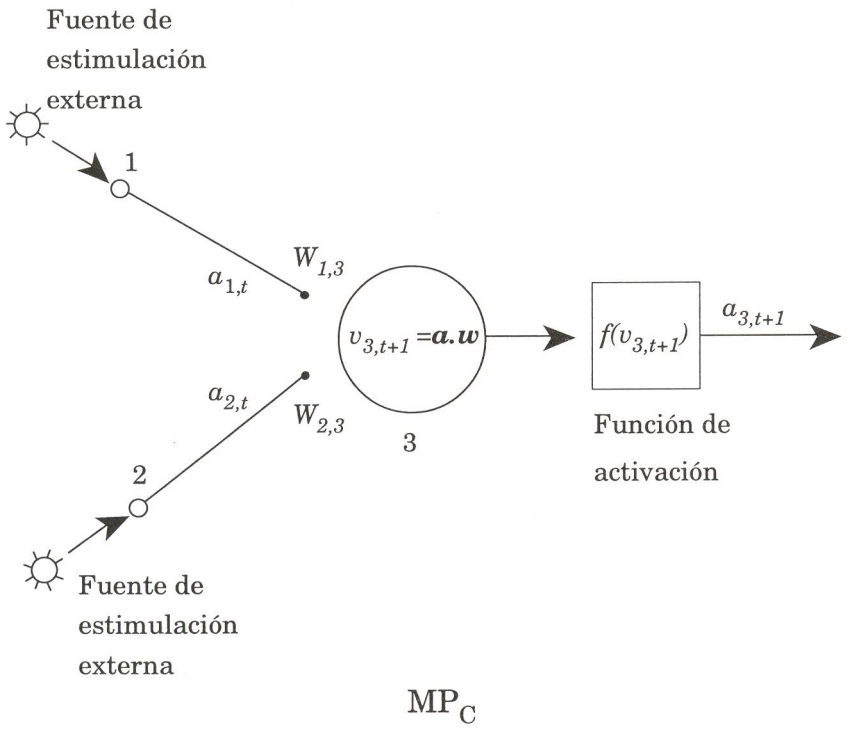

Figura 2. Sistema MP_C, el cual consta de dos sensores (círculos pequeños blancos marcados como 1 y 2) y un elemento procesador (círculo grande blanco marcado como 3).

El universo conductual de MP_C tiene dos componentes: entrada y salida. El componente relativo a la entrada se refiere al número de *formas*

distintas posibles en las cuales pueden ser activados sus sensores. Puesto que los estados de activación de los sensores de estos sistemas son binarios, MP_C puede recibir $2^n = 4$ *posibles patrones de entrada distintos*, donde $n = 2$ es el número de sensores del sistema. Cada patrón de entrada es un par de estados de activación aplicados a los sensores del sistema en un momento t. En principio, el orden temporal en el cual se apliquen los posibles patrones de entrada a MP_C no hace diferencia alguna.

Por ejemplo, una posible secuencia temporal de patrones posibles de entrada sería

$a_{1,1} = 1$	$a_{2,1} = 1$
$a_{1,2} = 1$	$a_{2,2} = 0$
$a_{1,3} = 0$	$a_{2,3} = 1$
$a_{1,4} = 0$	$a_{2,4} = 0$

donde una columna representa el estado de activación de un sensor (primer subíndice de cada a) y una fila representa un momento temporal (segundo subíndice de cada a). En cada momento temporal, *ambos* sensores adquieren su estado de activación simultáneamente. De nuevo, la secuencia temporal particular de pares de estados de activación de entrada es irrelevante.

En el componente relativo a la salida, el estado de activación de MP_C puede a su vez adquirir dos valores posibles. Es decir, MP_C puede responder de dos maneras diferentes (mutuamente excluyentes) a cualquier patrón de entrada. Ello permite definir una relación entrada-salida para MP_C. Si adoptamos el concepto conjuntista de relación, una relación entrada-salida puede definirse como un conjunto de tríos ordenados, donde los dos primeros miembros de un trío representan un patrón de entrada (las activaciones de los dos sensores) en t y el tercer miembro representa la salida (la activación del elemento procesador) del sistema en $t + 1$. Por ejemplo, una posible relación R_1 de entrada-salida para MP_C, o $R_1(MP_C)$, sería:

$$R_1(\text{MP}_\text{C}) = \{(1,1,1), (1,0,0), (0,1,0), (0,0,0)\},$$

donde cada trío de números entre paréntesis es un trío ordenado que representa una *ocurrencia particular* de entrada-salida[14]. Los primeros dos números de cada par representan un patrón de entrada posible (aplicado en t), mientras que el tercer número representa una posible activación de salida (la cual ocurre en el momento $t + 1$, debido a la presencia de una conexión para cada entrada).

Debido a que la señal de salida siempre se demorará un momento en ocurrir, respecto a cualquier patrón de entrada, la forma genérica de una ocurrencia entrada-salida para MP_C será:

$$(a_{1,t}, a_{2,t}, a_{3,t+1})$$

Ello significa que a MP_C le tomará un mínimo de ocho momentos realizar una relación entrada-salida, tal y como se muestra a continuación para $R_1(\text{MP}_\text{C})$:

$$\{(a_{1,1}=1, a_{2,1}=1, a_{3,2}=1), (a_{1,3}=1, a_{2,3}=0, a_{3,4}=0), (a_{1,5}=0, a_{2,5}=1, a_{3,6}=0), (a_{1,7}=0, a_{2,7}=0, a_{3,8}=0)\}.$$

Otra posible relación sería:

$$R_2(\text{MP}_\text{C}) = \{(1,1,1), (1,0,1), (0,1,1), (0,0,0)\}.$$

Otras serían:

$$R_3(\text{MP}_\text{C}) = \{(1,1,1), (1,0,0), (0,1,1), (0,0,1)\}.$$
$$R_4(\text{MP}_\text{C}) = \{(1,1,0), (1,0,1), (0,1,1), (0,0,0)\}.$$
$$R_5(\text{MP}_\text{C}) = \{(1,1,1), (1,0,0), (0,1,0), (0,0,1)\}.$$

---

14 Estrictamente, una relación es definida en teoría de conjuntos como un conjunto de *pares* ordenados, mientras que aquí he definido una relación entrada-salida como un conjunto de tríos. Sin embargo, todo conjunto ordenado puede expresarse como un par ordenado. En el presente caso, el trío (1,1,0), por ejemplo, puede expresarse como el par ordenado ((1,1),0), donde el primer miembro es, as su vez, un par ordenado. Mi definición, pues, es consistente con el concepto general de relación en teoría de conjuntos. Sin embargo, para simplificar la notación, seguiré expresando los tríos como tríos, y no como pares.

Y así sucesivamente. Nótese que todas estas relaciones difieren sólo en la secuencia de estados de activación de salida. En total, hay 2^{2n} = 16 relaciones entrada-salida lógicamente posibles para MP_C, donde, de nuevo, n = 2 [15]. El universo conductual de MP_C, entonces, es, precisamente, ese conjunto de todas las relaciones entrada-salida lógicamente posibles para MP_C. Si una realización de una relación entrada-salida le toma MP_C un mínimo de ocho momentos realizar, entonces una realización de su universo conductual le tomará un mínimo de $8 \times 16 = 128$ momentos.

Otro aspecto importante del presente análisis tiene que ver con la noción de conocimiento. La dificultad principal al respecto es el fuerte carácter multívoco del término. En el caso del conocimiento explícito, de nuevo, la noción tradicional que encontramos en filosofía define "conocimiento" como "creencia verdadera justificada". Tal noción no puede ser aplicada al conocimiento implícito sin que se adopten suposiciones por demás arriesgadas. Puesto que resulta difícil suponer que un sistema MP es capaz de tener conocimiento explícito, hablar de conocimiento en referencia a estos sistemas debe circunscribirse al conocimiento implícito. Afirmar que un sistema MP sabe *cómo* hacer tal o cual cosa (conocimiento implícito), es menos problemático que afirmar que sabe *que* tal o cual cosa (conocimiento explícito). Pero, ¿qué significa que un sistema MP sabe cómo hacer algo?

La respuesta más sencilla proviene del hecho de que los sistemas MP son sistemas *clasificadores*, es decir, sistemas que pueden responder diferencialmente a (i.e., discriminar entre) *tipos* distintos de patrones

---

15 Debido a la forma en la cual funcionan los sistemas MP, MPC no puede realizar teóricamente ni *R*4(MPC) ni *R*5(MPC), debido a que son tareas linealmente inseparables. Sin embargo, obviaré este detalle técnico en mi análisis. Asimismo, nótese que hay una relación estrecha entre este tipo de sistemas y la lógica proposicional. En efecto, el número máximo de relaciones entrada-salida lógicamente posibles para MPC coincide con el número total de tablas de verdad que pueden construirse con dos proposiciones. Algunas de dichas tablas definen los operadores lógicos estándar [viz., conjunción, disyunción inclusiva, condicional material, disyunción exclusiva y bicondicional, los cuales corresponden a *R*1(MPC), *R*2(MPC), *R*3(MPC), *R*4(MPC) y *R*5(MPC), respectivamente].

de entrada. Entonces, conocimiento implícito en estos sistemas consiste, precisamente, en *clasificar* (en el sentido técnico de dividir tajantemente, o 'partir', exhaustivamente un cierto dominio; el concepto técnico aquí es el de *partición*). Sobre esta base, se puede decir que un sistema MP sabe *cómo* responder diferencialmente a distintos tipos de patrones de entrada, sí, en efecto, responde diferencialmente a tales tipos (del mismo modo en que se dice que un ave sabe cómo volar, sí, en efecto, vuela).

Un tipo de patrón de entrada viene dado por un conjunto de patrones que tienen algo en común. Por ejemplo, en referencia de nuevo a MP_C, los siguientes seis patrones, presentados en secuencia temporal:

$a_{1,1} = 1$	$a_{2,1} = 0$
$a_{1,2} = 1$	$a_{2,2} = 0$
$a_{1,3} = 1$	$a_{2,3} = 0$
$a_{1,4} = 1$	$a_{2,4} = 0$
$a_{1,5} = 1$	$a_{2,5} = 0$
$a_{1,6} = 1$	$a_{2,6} = 0$

tienen en común que, en todos ellos, la activación del primer sensor es siempre 1 y la del segundo sensor es siempre 0. En este caso, se han presentado siete instancias de un mismo tipo de patrón de entrada, a saber, el patrón (1,0). Otro tipo de patrón de entrada sería, por ejemplo:

$a_{1,1} = 1$	$a_{2,1} = 1$
$a_{1,2} = 1$	$a_{2,2} = 1$
$a_{1,3} = 1$	$a_{2,3} = 1$
$a_{1,4} = 1$	$a_{2,4} = 1$
$a_{1,5} = 1$	$a_{2,5} = 1$
$a_{1,6} = 1$	$a_{2,6} = 1$
$a_{1,7} = 1$	$a_{2,7} = 1$

En este caso, se han presentado siete instancias del tipo (1,1), y así sucesivamente.

La presentación repetida de un tipo de patrón se hace necesaria para asegurarnos que el sistema efectivamente responde a dicho patrón *de la misma manera*. Se dice, entonces, que MP_C discrimina entre los dos patrones anteriores si, en efecto, responde de la misma manera a instancias de un mismo tipo y de manera diferente a instancias de tipos diferentes [e.g., con una activación de 1 ante instancias del tipo (1,0) y con una activación de 0 ante instancias del tipo (1,1), o viceversa][16]. En tal caso, se dice que MP_C sabe cómo realizar la tarea en cuestión.

Como otro ejemplo, tómese la tarea $R_3(MP_C)$, o {(1,1,1), (1,0,0), (0,1,1), (0,0,1)}, que en la Nota 14 llamé "condicional material". Se dice que MP_C sabe cómo realizar esta tarea si responde (en el siguiente momento) con 1 a instancias de los patrones de entrada (1,1), (0,1) y (0,0), y con 0 a instancias del patrón (1,0) (o viceversa). Los dos tipos de patrones de entrada, entonces, serían {(1,1),(0,1),(0,0)} versus {(1,0)}. Se dice que MP_C sabe cómo realizar esta tarea sí, en efecto, ante instancias del primer tipo de patrón responde con 1 y ante instancias del segundo tipo responde con 0[17].

---

16 Puesto que MPC consiste de un elemento procesador, y todo elemento procesador en el modelo MP es binario, este sistema sólo puede realizar tareas que involucren dos tipos de patrones de entrada. Para realizar tareas que involucren más de dos tipos de patrones, un MP debe tener más de un elemento procesador. Estas tareas y los sistemas MP que pueden realizarlas son considerablemente más complejos y no los examinaré en el presente análisis.

17 El que un sistema MP sepa o no cómo realizar una cierta tarea, depende de si hay al menos un conjunto de pesos, para un valor de θ, que le permita al sistema responder de la manera especificada por la tarea. Por ejemplo, $w1,j$ = -0.25 y $w2,j$ = 0.4 le permitirán a MPC realizar el condicional material, si θ = -0.1. De hecho, hay un número infinito de conjuntos de pesos para θ = -0.1 que le permitirán a MPC realizar esta misma tarea. También, un mismo conjunto de pesos puede servirle a MPC para realizar otras tareas.

Autoconocimiento Implícito en Sistemas MP

Para poder hablar de autoconocimiento en referencia a sistemas MP, es necesario buscar el sistema mínimo necesario que pueda conocer su propia conducta. La mejor estrategia al respecto es buscar el sistema más simple posible, como punto de partida. Para ello, propongo el sistema que se muestra en la Figura 3, al cual llamaré MP_A ("A" de "autoconocimiento").

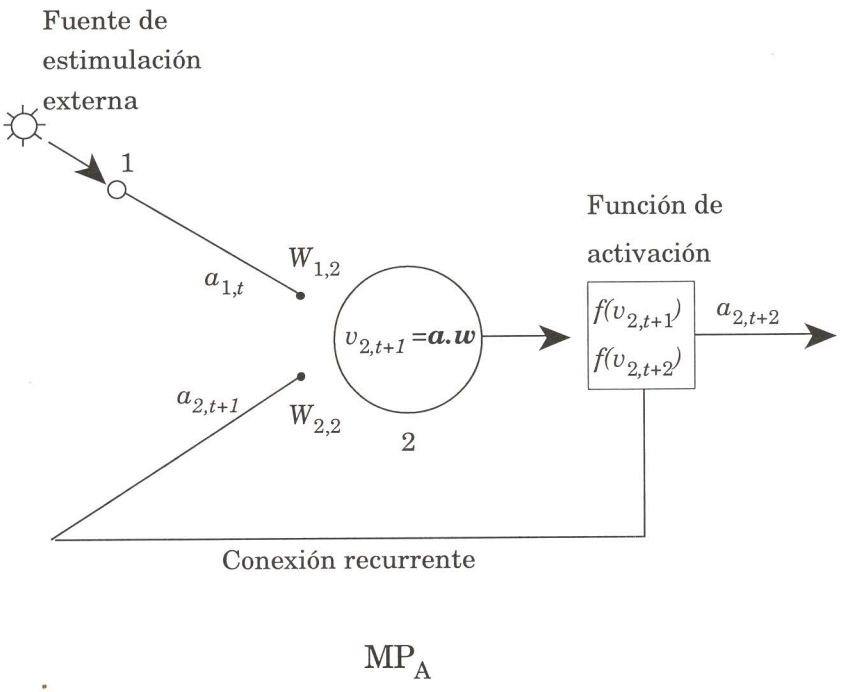

Figura 3. Sistema MP_A, el cual consta de un sensor (círculo pequeño blanco marcado como 1), un elemento procesador (círculo grande blanco marcado como 2), y una conexión recurrente que va del elemento procesador a sí mismo. Se requieren por lo menos tres momentos para calcular la salida del sistema. En el momento t, se aplica el estímulo externo. En el momento $t+1$, se calcula el estado de activación del elemento procesador, que servirá como segunda señal de entrada. En el momento $t+2$, se vuelve a calcular el estado de activación del elemento procesador, que servirá como salida del sistema.

MP_A tiene un sensor (círculo pequeño blanco marcado como "1") especializado en detectar estimulación externa y una conexión *recurrente* (o de retroalimentación) que va del elemento procesador (círculo grande blanco etiquetado como "2") hacia sí mismo, cuyo peso es denotado por $w_{2,2}$.

La clave para entender el funcionamiento de MP_A es la latencia. Siempre habrá una demora entre la activación del sensor y la del elemento procesador, debido a la conexión del primero con el segundo. En efecto, en $t = 1$, el estado de activación del elemento procesador estará determinado sólo por el estado de activación del sensor ($a_{1,1}$) y $w_{1,2}$, ya que $a_{2,1}$ es indeterminado. El estado de activación del elemento procesador podrá ser determinado sólo a partir de $t = 2$. Sin embargo, esta activación será la segunda señal de entrada para MP_A, por lo cual, la forma genérica de un patrón de entrada para MP_A será ($a_{1,t}$, $a_{2,t+1}$). La respuesta MP_A a este patrón de entrada podrá ocurrir sólo en el siguiente momento ($t = 3$). Por lo tanto, la forma genérica de una ocurrencia entrada-salida para MP_A es:

$$(a_{1,t}, a_{2,t+1}, a_{2,t+2})$$

Esta forma difiere de ($a_{1,t}$, $a_{2,t}$, $a_{3,t+1}$), la forma genérica de una ocurrencia entrada-salida de un sistema como MP_C, el cual carece de la estructura mínima necesaria para poder conocer su propia conducta. Resulta claro, entonces, que una realización de una relación entrada-salida requiere más tiempo para MP_A que para MP_C. Más precisamente, a MP_A le tomará un mínimo de 12 momentos (1/3 más del tiempo que a MP_C) realizar cualquier relación entrada-salida que le sea lógicamente posible, tal y como se muestra a continuación para {(1,1,1), (1,0,0), (0,1,0), (0,0,0)}:

$$\{(a_{1,1}=1,a_{2,2}=1,a_{2,3}=1), (a_{1,4}=1,a_{2,5}=0,a_{2,6}=0), (a_{1,7}=0,a_{2,8}=1,a_{2,9}=0),$$
$$(a_{1,10}=0,a_{2,11}=0, a_{2,12}=0)\}$$

Cualquier intento de ahorrarle tiempo a MP_A (de acelerar su proceso de autoconocimiento) hará imposible la realización de cualquier

relación entrada-salida. Por lo tanto, una realización de su universo conductual le tomará a MP_A un mínimo de 192 momentos.

Finalmente, llegamos a la pregunta central del análisis: ¿Puede MP_A clasificar su propio universo conductual? Aquí, el dominio por clasificar consta de *relaciones*. Clasificar un universo conductual, pues, significa responder diferencialmente a relaciones, no a ocurrencias entrada-salida. Para poder decir que MP_A conoce su propia conducta (sabe que se comporta de tal o cual forma), debe responder de la misma manera a realizaciones de un mismo tipo de relación entrada-salida y de diferentes maneras a realizaciones de diferentes tipos de relaciones entrada-salida. Si el universo conductual de MP_A consiste de 16 relaciones entrada-salida, entonces MP_A debería ser capaz de responder de 16 maneras diferentes. Sin embargo, MP_A tiene sólo un elemento de salida, lo cual le permite responder de sólo dos maneras diferentes. Por consiguiente, MP_A no puede clasificar su propia conducta. Un sistema que pueda responder de 16 maneras diferentes requeriría, como mínimo, cuatro elementos procesadores de salida, pero tal sistema incrementaría su universo conductual más allá de su propia capacidad de salida. Ningún sistema como MP_A jamás podrá conocer su propia conducta de manera final y definitiva. El autoconocimiento implícito completo es imposible en principio, al menos en sistemas como MP_A.

Conclusiones

El análisis precedente sugiere que la regresión al infinito desencadenada por el autoconocimiento explícito también se da en el autoconocimiento implícito, al menos en un tipo de sistema MP y por razones semejantes. En el autoconocimiento explícito, la regresión al infinito es desencadenada por el hecho de que siempre quedará un acto de introspección sin ser objeto de introspección alguna, en el caso de la doctrina cartesiana, o una descripción de la propia conducta sin ser descrita por descripción alguna, en el caso de la doctrina ryleana. En el autoconocimiento implícito, el universo conductual de un sistema MP

trasciende la capacidad de salida del sistema mismo, haciendo imposible que el sistema realice una clasificación de su propia conducta. Un aumento en la capacidad de salida del sistema automáticamente aumentará el tamaño de su universo conductual y, en esa medida, la capacidad de salida necesaria para realizar una clasificación del nuevo universo, y así *ad infinitum*. De manera análoga al autoconocimiento explícito, siempre quedará una porción substancial del universo conductual de MP_A sin ser conocida por MP_A.

Es posible que un sistema como MP_A pueda conocer parcialmente su propia conducta, en el sentido de que pueda responder diferencialmente a algunas de las relaciones entrada-salida que constituyen su universo conductual. En este caso, el análisis muestra que un sistema que posea la estructura mínima necesaria para conocer su propia conducta (i.e., un sistema que posea al menos una conexión recurrente), requerirá más tiempo para ser expuesto a su objeto de conocimiento que un sistema que no posea tal estructura. La implicación es que el autoconocimiento implícito, además de ser inherentemente incompleto, también es inherentemente más lento, al menos en los sistemas MP que he examinado.

El análisis presentado, por supuesto, es tentativo y preliminar, por cuanto lleva a más interrogantes de las que planteé al inicio. Por ejemplo, es posible que MP_A sea incapaz, no tanto de no realizar una clasificación, sino, más bien, una clasificación *máximamente fina* de su propio universo conductual. Para poder afirmar que MP_A puede autoconocerse, aunque sea gruesamente, habría que demostrar que es capaz de responder diferencialmente a por lo menos una partición de su universo conductual en dos subconjuntos. Sin embargo, ello requeriría un análisis mucho más extenso que el que puedo realizar aquí. El presente análisis sólo demuestra que el autoconocimiento implícito máximamente fino es imposible en principio para MP_A.

También se plantea el problema de la generalidad del resultado obtenido. Para determinar tal generalidad, el presente análisis deberá realizarse no sólo con diversos tipos de sistemas MP, sino también con

diversos tipos de sistemas en general, tanto subsimbólicos (redes neurales) como simbólicos (máquinas de Turing). La hipótesis de que el autoconocimiento implícito es imposible en principio para cualquier tipo de sistema, ganará apoyo en la medida en que tales análisis lleven a resultados comparables.

No obstante, la rigurosidad del análisis disminuirá en la medida en que aumente la complejidad de los sistemas analizados. El grado de rigurosidad alcanzado en el presente análisis fue posibilitado por la extrema simplicidad del tipo de sistema analizado. Si la rigurosidad de un análisis es inversamente proporcional a la complejidad del sistema analizado, entonces el grado de rigurosidad alcanzado con sistemas mucho más complejos (como, por ejemplo, los seres humanos) será mucho menor. En esa medida, las conclusiones a las que arribemos serán considerablemente más dudosas.

Bibliografía

Brown, J. R. (1991). *The laboratory of the mind: Thought experiments in the natural sciences*. Nueva York: Routledge.

Burgos, J. E. (2002). Behavioral knowledge and structural complexity in McCulloch-Pitts systems. Psycholoquy, **13**, #26 (http://psycprint.ecs.soton.ac.uk/archive/00000218/).

Descartes, R. (1985). The philosophical writings of Descartes (J. Cottingham, R. Stoothoff y D. Murdoch, trads.). Reimpreso parcialmente en D. M. Rosenthal (Ed., 1991), *The nature of mind* (pp. 21-32). Cambridge University Press.

Hume, D. (1748). *An inquiry concerning human understanding*. Reimpreso en The empiricists (pp. 307-430). Nueva York: Doubleday.

Kuhn, T. (1964). La función de los experimentos imaginarios. Reimpreso en T. Kuhn (Ed., 1977), *La tensión esencial: Estudios selectos sobre la tradición y el cambio en el ámbito de la ciencia* (pp.

248-262). Versión en español de Roberto Helier. México: Fondo de Cultura Económica.

Leibniz, G. W. (1714). The monadology. Reimpreso en *The rationalists* (pp. 455-471). NuevaYork: Doubleday.

McCulloch, W. S. y Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, **5**, 115-133. Reimpreso en W. S. McCulloch (Ed., 1965), *Embodiments of mind* (pp. 19-39). Cambridge, MA: MIT Press.

Malebranche, N. (1688). *Dialogues on metaphysics and religion* (N. Jolley, Ed.; D. Scott, Trads.). Cambridge University Press.

Russell, B. (1912). *The problems of philosophy*. Nueva York: Oxford University Press.

Ryle, G. (1949). *The concept of mind*. Chicago: The University of Chicago Press.

Skinner, B. F. (1953). *Science and human behavior*. Nueva York: The Free Press.

Skinner, B. F. (1974). *About behaviorism*. Nueva York: Alfred A. Knopf.

Sorensen, R. A. (1992). *Thought experiments*. Nueva York: Oxford University Press.

Spinoza, B. (1677). The ethics. Reimpreso en *The rationalists* (pp. 179-406). Nueva York: Doubleday.