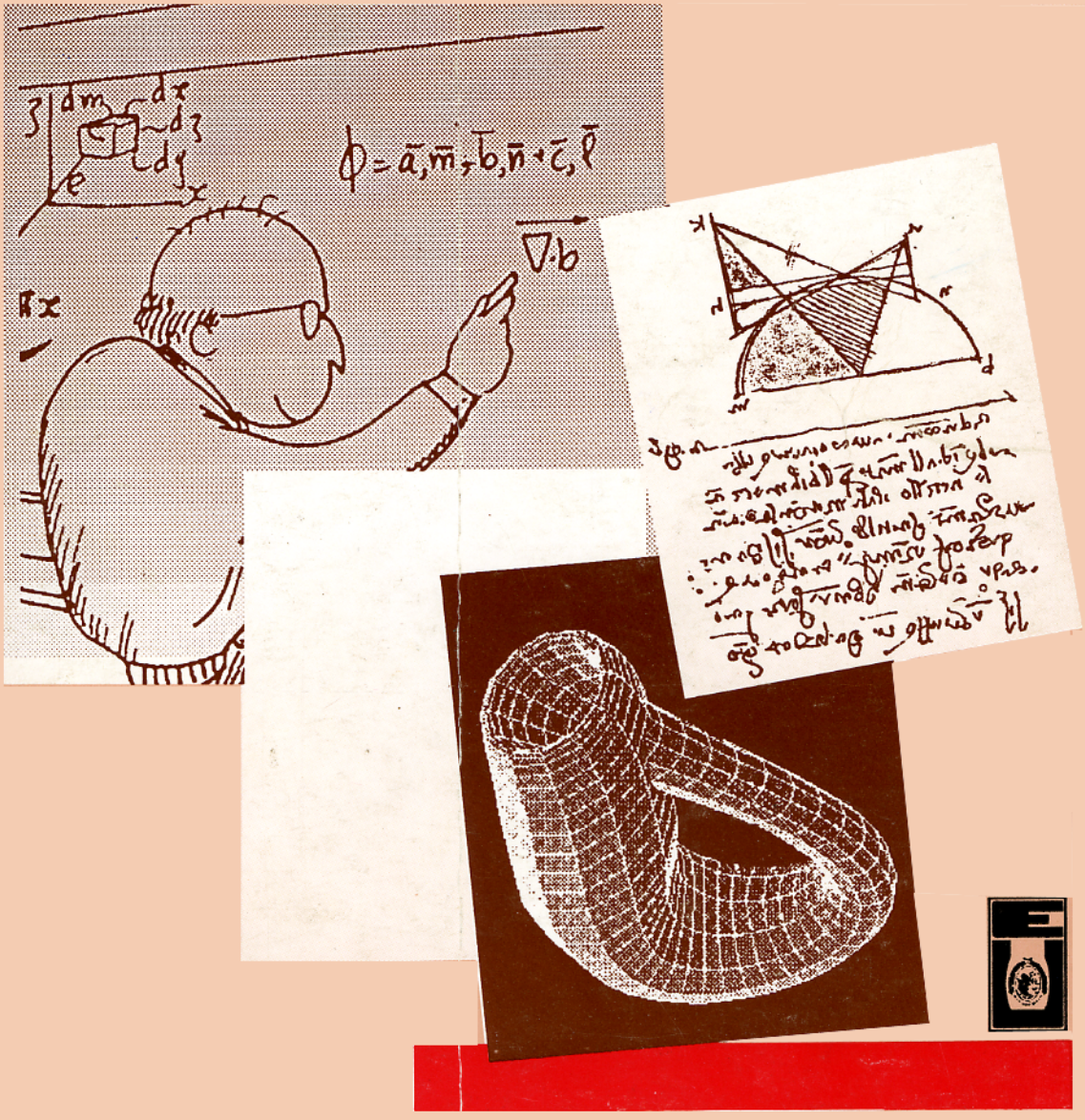
Ingeniería

Revist de la Universidad de Costa Rica
ENERO/JUNIO 1993 VOLUMEN 3 No. 1



Ingeniería 3 (1): 29-38, 1993 San José Costa Rica

FUNCIONALISMO ROBOTICO: ¿UNA RESPUESTA AL ARGUMENTO DE SEARLE?

*Manuel Arce Arenales**
*Max Freund Carvajal**

RESUMEN:

En este artículo se examinan las principales propuestas ofrecidas por Stevan Harnad para presentar una nueva versión del funcionalismo, llamada por él *funcionalismo robótico*, que pretende responder definitivamente a las objeciones planteadas por Searle en su ya clásico argumento del "cuarto chino". Después de un examen crítico de estas propuestas, se concluye que no logran su objetivo, y de camino se replantean algunos de los principales problemas que tiene que enfrentar cualquier teoría en Ciencias Cognoscitivas, particularmente aquéllas que asumen alguna variante de la *visión computacional de la mente*.

SUMMARY:

In this paper, the main proposals offered by Stevan Harnad to present a new version of functionalism, which he calls *robotic functionalism*, are examined. This new version purportedly escapes in a definitive fashion the objections presented by Searle in his now classic "Chinese Room Argument". After a critical examination of these proposals, the authors conclude that such proposals fail to meet their purpose, and in the process some of the main issues of Cognitive Science are reformulated, issues that are of fundamental importance for any theory in Cognitive Science, particularly for those which assume a *computational view of the mind*.

1. INTRODUCCIÓN

La conceptualización de la mente humana como una entidad funcionalmente idéntica a la implantación de un programa computacional, constituye una de las tendencias más importantes en la actual filosofía de la mente. De acuerdo con esta tendencia, conocida como *funcionalismo*, el cerebro constituye una especie de computador y la mente un programa computacional (o colección de procedimientos computacionales) ejecutado en este computador. En consecuencia, estados psicológicos tales como los deseos, las añoranzas, las creencias, los pensamientos y las imaginaciones, son considerados por esa teoría filosófica como estados computacionales del cerebro.

Uno de los argumentos más importantes en contra del funcionalismo ha sido el formulado por J. Searle, por ejemplo en Searle (1980), (1981) y (1984), y conocido como el *argumento del cuarto chino*. El argumento se fundamenta en la posibilidad de imaginar a Searle (quien no sabe chino, ni escrito ni hablado), encerrado en un cuarto con un algoritmo escrito en inglés (idioma que Searle sí conoce); este algoritmo le posibilita escribir caracteres chinos al recibir formulaciones escritas también con estos caracteres. Fuera del cuarto se encuentran varios hablantes nativos del chino, quienes le pasan, debajo de la puerta, pedazos de papel con preguntas escritas en caracteres de su lengua nativa. Estos símbolos no tienen ningún significado para Searle, pero él los usa como entrada y genera como salida, siguiendo solo el algoritmo, respuestas escritas en chino, las cuales son recibidas por los hablantes fuera del cuarto. Éstos encuentran las respuestas completamente indiferenciables de las que pudiera haber ofrecido cualquiera de ellos (en tanto hablantes del chino).

La intención de Searle con su argumento es mostrar que la implantación de un programa computacional para procesar un lenguaje natural dado no significa que haya comprensión de este lenguaje (y, por ende, presencia de su semántica) en el computador y, por otra parte, que no hay elemento alguno (ya sea en el programa o en el computador) en el proceso de implantación de cualquier programa para procesar un lenguaje natural dado, que implique la presencia de su semántica. Estas supuestas consecuencias del *Argumento del cuarto chino* han provocado diversas reacciones y publicaciones, en las cuales se ha cuestionado la fuerza y validez de dicho

*Profesores, Programa de Posgrado en Ciencias Cognoscitivas. ECCI, UCR

argumento. Sin embargo, parece claro que por ahora no se ha logrado formular un contraargumento que suponga elementos no cuestionables o que confronten directamente el punto esencial del razonamiento de Searle, esto es, el problema de cómo puede surgir la semántica de un programa computacional o de la interacción del programa con el computador (aun en la ejecución del programa). Este problema puede ser formulado en términos más generales como el problema de cómo puede surgir la intencionalidad, ya sea de un programa computacional o de la interacción de éste con el computador. Tómese en cuenta que la intencionalidad es un elemento cuya presencia es evidente en la mente humana y, por tanto, un problema importante para cualquier teoría de la naturaleza de la mente, en particular para el funcionalismo.

En Harnad (1989) se han expuesto varias objeciones al argumento de Searle, y se ha formulado una alternativa al funcionalismo clásico: el *funcionalismo robótico*. Es opinión de Harnad que esta alternativa escapa del *Argumento del cuarto chino*. En este artículo analizaremos críticamente algunos de los aspectos defendidos por Harnad. Mostraremos que varias de sus objeciones no están en lo correcto, y que su funcionalismo robótico no escapa del punto medular del argumento de Searle.

2. Searle y los presupuestos básicos del programa de inteligencia artificial clásico

El programa de la IA clásica nace con una pregunta y una respuesta afirmativa para esta pregunta: "¿Puede una computadora llegar a ser **inteligente**?". La pregunta puede también reformularse de la siguiente manera: "¿Puede una computadora manifestar **entendimiento**?" O, de manera más lata, "¿Puede una computadora adquirir los procesos cognoscitivos propios de los seres inteligentes (los seres humanos en particular)?". El fundamento para dar una respuesta afirmativa a estas preguntas es, por una parte, una caracterización (asumida como correcta) de la inteligencia, el entendimiento, y los procesos cognoscitivos; y por otra, la suposición de que la inteligencia, el entendimiento, y los procesos cognoscitivos en general, son reducibles a procedimientos computacionales de algún tipo u otro[1]. La caracterización de los procesos cognoscitivos en términos de procedimientos computacionales conlleva a su vez el supuesto de **autonomía funcional**, según el cual lo esencial de la función congnoscitiva (i.e., la función misma) es independiente de la infraestructura material en la cual se "encarna"[2].

---

[1] En las palabras de Paul Churchland, "...para cualesquiera procedimientos computacionales bien definidos, una máquina universal de Turing es capaz de simular una máquina que ejecute dichos procedimientos. Hace esto al reproducir exactamente el comportamiento de insumo/salida de la máquina por simular. Y lo interesante es que la computadora moderna es una máquina universal de Turing... La pregunta que confronta al programa de investigación en IA, por consiguiente, no es si computadoras adecuadamente programadas son capaces de simular el comportamiento continuo producido por los procedimientos computacionales que se encuentran en los animales naturales, incluidos aquéllos que se encuentran en los seres humanos... La pregunta importante es si las actividades que constituyen la inteligencia consciente son todas procedimientos computacionales de algún tipo u otro. La suposición guía en IA es que sí lo son, y el objetivo es construir programas reales que los simulen." [Churchland, Paul, 1990: pag. 105. Ésta, y todas las subsecuentes traducciones de citas en este artículo, son nuestras.]

[2] De nuevo en las palabras de Churchland: "...no tiene por qué haber diferencia entre los procedimientos computacionales [humanos] y los procedimientos computacionales de una simulación de máquina, ninguna diferencia más allá de la sustancia física particular que soporta dichas actividades. [En el ser humano], es material orgánico; en la computadora, sería metales y semiconductores. Pero esta diferencia no es más pertinente a la pregunta de la inteligencia consciente que una diferencia de tipo sanguíneo, de color de la piel, o de química metabólica... Si las máquinas llegan a simular todas nuestras actividades cognoscitivas internas, hasta el último detalle computacional, negarles el "status" de personas genuinas no sería otra cosa que un nuevo tipo de racismo..." [Churchland, Paul, 1990: pags. 119-120]. Como veremos más adelante, el funcionalismo robótico parte de los supuestos del funcionalismo clásico, pero los modifica de manera sustancial.

En la propia raíz del argumento de Searle (y del problema fundamental mismo) está el concepto de **entendimiento** ("understanding"). En la IA "clásica" se acepta explícita o implícitamente que alguien o algo **entiende** una cosa (concepto, operación, procedimiento) cuando no podemos distinguir entre el desempeño de un ser comprobadamente inteligente (e.g. un ser humano) al entender esa cosa, y el desempeño correspondiente de ese otro alguien o algo[3]. En IA, como hemos visto, la pregunta generalmente se ha reducido a si una computadora es capaz de **entender** en el sentido utilizado para los seres humanos.

Es claro que Searle niega estos presupuestos básicos y asume, más bien, las siguientes tesis (las cuales, de ahora en adelante, llamaremos "las tesis de Searle"):

I. la infraestructura física (en nuestro caso la infraestructura orgánica basada en la química del carbono) es un componente crucial de la inteligencia consciente;

II. las actividades que constituyen la inteligencia consciente no pueden ser caracterizadas todas ellas, ni cada una en forma exhaustiva, como procedimientos computacionales de alguno u otro tipo;

III. no es cierto que sea obligatorio concluir que dos entidades "hacen la misma cosa" cuando ya no podemos diferenciar entre sus respectivos desempeños.

Con respecto al punto III, por ejemplo, puede ser que para desempeñar una determinada función un organismo necesite inteligencia, mientras que para desempeñar una función idéntica otro organismo no la requiera del todo. O podría ser que ambos requieran inteligencia, pero de tipos radicalmente distintos. Nos parece que del argumento de Searle se desprende que una caracterización operacional de **entendimiento** no es suficiente para determinar su verdadera naturaleza.

3. El Funcionalismo Robótico

Harnad es de la opinión que el argumento de Searle es irrefutable si se asume una posición que él llama "funcionalismo simbólico", la cual caracteriza como:

> "...la hipótesis de que el funcionamiento mental consiste solamente en manipulación formal de símbolos..." [Harnad, 1989: pag. 15].

Contrapuesto al funcionalismo simbólico, Harnad propone lo que él denomina "funcionalismo robótico", esto es,

> "...la hipótesis de que las funciones no-simbólicas están involucradas de manera crítica en los estados mentales..." [Harnad, 1989: pag. 15].

Harnad cree que esta forma de funcionalismo sí escapa al argumento de Searle. Nos dice que:

> " (...) el argumento de Searle del Cuarto Chino falla por completo para la versión robótica de la prueba de Turing cuando la propiedad mental correspondiente en cuestión es la *percepción* de objetos, más que la la comprensión de símbolos. Para que se vea esto, nótese que los términos del Argumento requieren que Searle muestre (i) que él puede asumir todas las funciones del robot (...) y sin embargo (ii) que él no ha podido exhibir la propiedad mental en cuestión— en este caso, la percepción de objetos. Ahora bien, considérense dos casos posibles: (1) Si Searle simula sólo la manipulación de símbolos *entre* los transductores y los efectores, entonces no está llevando a cabo todas las funciones del robot (y de ahí que no sorprenda que él no perciba los objetos que el robot supuesta-

[3] "De acuerdo con...[la prueba de Turing], deberíamos dejar de negar que una máquina 'realmente' está haciendo la misma cosa que hace una persona cuando ya no podemos diferenciar entre los respectivos desempeños." [Harnad, 1989: pag. 6]. Aunque Harnad aparentemente niega que la mente humana sea equiparable a una máquina universal de Turing (cf. la siguiente nota al pie de página), de sus planteamientos creemos deducir que en lo mínimo no se pronuncia respecto de este otro problema (a iguales desempeños, iguales estados internos).

mente percibe) (2) Si, por otra parte, Searle juega el papel de homúnculo para el robot, él mismo viendo la escena o pantalla, entonces él es su transductor (y de ahí que no sorprenda que perciba de verdad lo que el robot supuestamente percibe). Un argumento similar se aplica a la actividad motora. La función robótica a diferencia de la función simbólica, es inmune al argumento de Searle del Cuarto Chino." [Harnad, 1989: pag. 19; el subrayado es nuestro].

El argumento fundamental del funcionalismo robótico es que no basta con una caracterización, por completa que ésta sea, de los procedimientos computacionales. Esta es una condición necesaria, pero puede no ser suficiente para replicar un fenómeno dado[4]. Para lograr esto último, debe haber una **implantación efectiva**, que incluya no solo las estructuras formales pertinentes, sino también los conectores con el medio, e.g. sensores y extensiones motoras. Para el funcionalismo robótico, el argumento de Searle es vacuo porque tan solo plantea una simulación de una simulación, no una implantación verdadera (la cual, en opinión de Harnad, sería imposible en los términos planteados por Searle).

4. Las tesis de Searle y el funcionalismo robótico

Consideremos ahora las tesis de Searle en su relación con el funcionalismo robótico. Es claro que con respecto a la primera tesis (esto es, la necesidad de una infraestructura biológica), esta forma de funcionalismo argumentaría que una implantación efectiva, por el solo hecho de existir, demostraría que una base biológica no es indispensable para alojar el comportamiento inteligente. En cuanto a la segunda tesis (negación de la equiparación entre inteligencia consciente y procedimientos computacionales), ésta de hecho es compartida por el funcionalismo robótico, aunque de alguna manera se mantiene que el comportamiento inteligente efectivo es reducible a principios formales o formalizables, si bien ahora éstos no estarían limitados a principios meramente simbólicos[5].

Ahora bien, no es fácil ver cómo la propuesta de Harnad podría dar cuenta de la tercera tesis (desempeños indiferenciables no implican necesariamente estados internos idénticos), pero ésta podría considerarse meramente académica, una vez logrado un mecanismo que se desempeñare, para todo propósito práctico, igual que un ser humano. Sin embargo, aun así no es tan fácil descartar el problema. No hay que olvidar que, para el mismo Harnad,

"si la simulación modela o formaliza los rasgos relevantes...de la implantación...entonces desde el punto de vista de nuestra comprensión funcional del mecanismo causal involucrado, las dos son

---

[4] Hasta aquí hay acuerdo parcial con la posición de Searle: la aprehensión de las estructuras formales determinantes (los rasgos 'causales' en su terminología) quizá sea suficiente para (ciertas) simulaciones, pero no para la replicación de un fenómeno dado. Nadie afirmaría fácilmente que la simulación de un fenómeno atmosférico (e.g. un huracán) sea el fenómeno atmosférico mismo. Hay que tener en mente la distinción entre implantación y simulación, que para Harnad es crucial. Además, Harnad al igual que Searle niega que la función mental sea "realmente solo función simbólica (e.g. función verbal, inferencial, computacional): que la mente manipule símbolos como lo hace una máquina de Turing, y que por consiguiente el cerebro únicamente provea el "hardware" necesario para computar..."[Harnad, 1989: pag.8] Para quienes mantienen todavía una posición dentro de la IA "clásica" (e.g. Patrick Hayes o Kenneth Ford, 1992), no es necesario ceder terreno a Searle en este punto, puesto que hay una confusión implícita de niveles. La simulación de un huracán ciertamente no es un huracán, pero la simulación de un procedimiento computacional ES un procedimiento computacional. Por supuesto, para dar plena cuenta del argumento de Searle habría que aceptar que todo proceso cognoscitivo es reducible a procedimientos computacionales.

[5] En este punto hay que tener cuidado, pues para un funcionalista robótico el término 'principio formal' no es equivalente (ni totalmente reducible) a 'funciones simbólicas' (quienes aceptan esta equivalencia son llamados por Harnad "funcionalistas simbólicos"). Para los funcionalistas robóticos, "las funciones no-simbólicas (e.g. las funciones sensoriales, motoras, analógicas, y asociativas) [son] potencialmente tan mentales o cognoscitivas como las funciones simbólicas, y pueden incluso ser primarias, con las funciones simbólicas 'enraizadas' en las [funciones] no-simbólicas." [Harnad, 1989: pag. 8].

teóricamente equivalentes" [Harnad, pag. 7: 1989][6].

Esto deja abierta la posibilidad de que dos organismos de desempeño aparentemente idéntico posean diferentes propiedades causales relevantes. Podríamos llegar a entender plenamente el modelo o formalización bajo el cual funcionare un androide, sin por esto llegar a tener garantía de que entendemos cómo funciona realmente un ser humano, en particular sin llegar a tener garantía de que el androide realmente **entiende** como entiende un ser humano[7].

5. Mecanismo y Operación

Otros dos conceptos básicos que involucran presupuestos teóricos de importancia, son los conceptos interrelacionados de **mecanismo** y **operación**. El propio Harnad nos dice: "La idea de **mecanismo** está realmente en el corazón del problema hombre/máquina (mente/programa)." [Harnad, 1989: pag. 7]. Según este autor,

> "...un mecanismo es un sistema físico que opera de acuerdo con leyes causales, físicas (incluyendo principios específicos de ingeniería)" [Harnad, 1989: pag. 7].

Esta caracterización tiene la virtud de incluir junto con molinos de viento, relojes, aviones, y computadoras (que corresponden a la noción intuitiva de mecanismo), entidades tales como una ameba, un roedor, y un hombre [Harnad, 1989: pag. 7].

Dicha virtud, sin embargo, se convierte en problema, puesto que al ser tan amplia la caracterización de mecanismo de igual manera tendríamos que incluir un ecosistema, un átomo, una ola, una mano, un zapato, una ciudad. De hecho, pocas entidades existentes en un sentido más o menos unitario (i.e., distinguibles como tales) quedarían por fuera de esta caracterización de **mecanismo**. Por otra parte, hay entidades que podrían corresponder intuitivamente a la noción de mecanismo, pero que no pueden ser tan fácilmente caracterizadas como mecanismos en los términos utilizados por Harnad. Tal es el caso de la ramita que utiliza un pajarillo para buscar y conseguir gusanos en un tronco, o del puente que erigen las hormigas con sus propios cuerpos para cruzar una corriente de agua. Por supuesto, no habría problema alguno si se adopta una posición determinista (lo cual es común en los proponentes de cualquiera de las variantes del funcionalismo). Sin embargo, en este caso también tendríamos que incluir como mecanismos una declaración de derechos humanos, una obra de teatro, un poema, una propuesta de amor, un partido político, una pareja enamorada.

Además, es claro que, relacionado con el problema de determinar mecanismo simplemente por medio de una operación acorde con leyes causales, está el problema de establecer los límites de lo que debemos o podemos considerar como mecanismo. Por ejemplo, hay seres humanos capaces de determinar la hora casi con tanta precisión como la obtenible con un reloj, atendiendo únicamente al sol y a otros fenómenos

[6] Esta misma "equivalencia teórica" no tiene por qué ser aceptada necesariamente: toda implantación está acompañada por "conocimientos del mundo real" que debe suplir el implantador, y que casi nunca (excepto en los casos más triviales) están todos considerados en el modelo (la simulación). Esto es particularmente importante para el funcionalismo robótico, donde deben considerarse no solo las funciones puramente simbólicas sino, incluso de manera primordial, aquéllas que son no-simbólicas. El que estos conocimientos tengan o no tengan la categoría de "rasgos causales" en uno u otro sentido es otra pregunta, y no es fácilmente descartable. Podría decirse que en principio no pueden tenerla, o la simulación sería "deficiente" por este solo hecho. Sin embargo, cabe preguntarse si una simulación efectiva debe incluir todos los rasgos causales, o si basta muchas veces con los más relevantes (i.e., ¿cómo puede determinarse en forma teórica el grado de relevancia de un rasgo?). A manera de ejemplo, no conocemos a nadie que siempre haya tenido la experiencia de que un texto salga impreso exactamente como se quería, vale decir idéntico a la simulación proporcionada por el procesador de textos: algún detalle tiene que corregirse siempre con un primer 'prototipo' en mano.

[7] Parte del problema es que toda simulación eficaz simula no solamente el mecanismo involucrado, sino también partes cruciales del medio en el cual opera u operaría dicho mecanismo.

celestes y atmosféricos. Este conjunto de indicadores naturales, ¿constituye un mecanismo?[8] El concepto de mecanismo que ofrece Harnad parece demasiado lato y muy poco específico como para ser útil.

Por otra parte, ¿cómo se sabe que se conocen los rasgos causales relevantes de un mecanismo? Para Harnad, si alguien es capaz de construir el mecanismo en cuestión, por este solo hecho conoce sus propiedades causales relevantes; también se sabe que las conoce si es capaz de proveer el plano formal o el programa para construirlo [Harnad, 1989: pag. 7]. ¿Será siempre posible pasar de un tipo de conocimiento al otro (i.e., de un conocimiento puramente operacional a un conocimiento teórico, y viceversa)? En otras palabras, ¿son estos tipos de conocimiento equivalentes? Si una edificación[9] puede ser considerada como un mecanismo, ¿conocen las termitas o las abejas las propiedades causales que subyacen sus edificaciones? Ciertamente son capaces de construirlas. Si la respuesta es afirmativa (algo que una aplicación consecuente de las definiciones de Harnad parece exigir), ¿quién conoce estas propiedades? ¿Cada una de las termitas? ¿Algún subgrupo relevante? ¿Todas ellas juntas? ¿Cuál es el número mínimo de termitas necesario para garantizar que tienen el conocimiento en cuestión? Pero no se necesita ir tan lejos: ¿hasta qué punto conoce un maestro de obras rural las propiedades causales relevantes que le permiten construir exitosamente una casa? El conocimiento que tenían los constructores medievales que erigieron las magníficas catedrales góticas que hoy admiramos todavía, ¿es equivalente al que posee un ingeniero moderno capaz (¡tal vez!) de dirigir la construcción de un edificio similar o idéntico?

En definitiva, los conceptos seminales de **mecanismo** y **operación** en el funcionalismo robótico (así como el de **entendimiento** y la distinción misma entre **simulación e implantación**) no parecen haber sido presentados con una claridad suficiente como para sostener o introducir una respuesta convincente al argumento del cuarto chino.

6. La prueba total de Turing y el Funcionalismo Robótico

Puesto que la implantación de una simulación exitosa (una que incorpore todas las propiedades causales relevantes) es la medida última que permite afirmar haber entendido las propiedades causales de un determinado mecanismo, Harnad arguye que el funcionalismo robótico es inmune al argumento del cuarto chino. Hemos mostrado varias razones que Harnad ofrece para sustentar esta afirmación. Una razón adicional la fundamenta Harnad en las características de la prueba de Turing; no basta con someter al candidato a una prueba parcial de Turing, según Harnad, sino que ésta debe ser total.

Por ejemplo, no bastaría con que nuestro hipotético cuarto chino "entendiere" la sintaxis del chino: debería ser capaz de interactuar con un hablante nativo de este idioma sobre cualquier tema susceptible de ser tratado por un ser humano. En otras palabras, la prueba de Turing no podría ser reducida a una prueba meramente 'verbal'10: tendría que incluir todas aquellas actividades de las que es capaz un hombre o una mujer normales. Harnad afirma que Searle ha aceptado implícitamente una de las hipótesis centrales de los funcionalistas simbólicos, i.e., que un LN es reducible a sus componentes puramente formales (valga decir,

[8] Nótese que todos los ejemplos de mecanismos 'no naturales' dados por Harnad son aparatos creados por el hombre (un reloj, un molino de viento, un avión, una computadora). Implícitos aquí están los problemas de la **intención** y del **uso**: ¿deben ser considerados éstos a la hora de identificar los rasgos causales de un mecanismo 'no natural'? Harnad soslaya este problema al ofrecer una caracterización muy general de **mecanismo**, pero esto le crea otros problemas igualmente difíciles.

[9] Entendamos por 'edificación' una estructura física que proporcione a sus ocupantes condiciones apropiadas de habitación, lo cual involucra sistemas idóneos de ventilación, de protección contra los elementos, etc.

[10] Es difícil obtener una idea clara de lo que Harnad entiende por 'verbal'. Él mismo parece advertirnos del error de considerar un LN comprensible en términos puramente sintácticos (un error que, por lo demás, casi nadie cometería en la actualidad). Pero si incluimos la semántica (en el sentido lato de la palabra), de alguna manera estamos incluyendo la mayoría de las representaciones o procedimientos que un ser humano emplea al interactuar con su medio. Al menos como nosotros entendemos su argumento, Searle ciertamente incluye la semántica como necesaria para poder decir que realmente se comprende el chino.

sintácticos). Pero en realidad lo que Searle pretende afirmar es, entre otras cosas, precisamente lo contrario: que un organismo que opere sobre la base de reducciones puramente formales es en principio incapaz de entender verdaderamente un LN. El argumento supone que existe un mecanismo (para usar la terminología de Harnad) capaz de imitar el proceso de comprensión de un LN reduciéndolo a manipulaciones puramente formales, para luego mostrar que nunca podría afirmarse que dicho mecanismo realmente **entiende**. El argumento consiste en una especie de reducción al absurdo, donde se acepta una hipótesis que conduce a consecuencias absurdas.

La pregunta que persiste, entonces, es si la propuesta del funcionalismo robótico escapa del cuestionamiento general que Searle hace de la IA, a saber: el cuestionamiento en torno a la posibilidad de construir un organismo artificial que replique la manipulación que un ser humano normal hace de un LN y, en general, que sea capaz de **entendimiento**. Puesto que esta propuesta comparte con Searle la posición de que no todo proceso cognoscitivo es reducible a procedimientos computacionales, queda por ver si plantea una posibilidad verdadera para afirmar que la base biológica no es una condición necesaria para resolver el problema de la **intencionalidad** (y, de paso, del **entendimiento** en general). Esto último está ligado con el problema de determinar cuándo podemos afirmar que los estados interiores de dos cognoscentes (o de un cognoscente y un 'candidato' a serlo) son 'realmente' equivalentes (i.e., y no que meramente aparezcan equivalentes).

Ya hemos visto que, para Harnad, una de las deficiencias del argumento de Searle es que presupone una prueba parcial de Turing, y no una prueba total como debería hacerlo. Searle, nos dice Harnad, no hace otra cosa que conjeturar cómo podríamos simular un mecanismo que comprendiere chino: no nos presenta una simulación exitosa, mucho menos una implantación efectiva. Pero, como el mismo Harnad acepta, podría plantearse un problema de dimensiones considerablemente más reducidas susceptible de ser implantado aun con la tecnología disponible, e.g. jugar al ajedrez. No importa, responde Harnad. Las simulaciones exitosas que podemos traer a colación son todas ellas parciales: ninguna podría calificar como prueba total de Turing, mientras que el argumento de Searle presupone la comprensión de un LN como 'operación' básica, y ésta sí que necesariamente involucraría una prueba total. De nuevo, pareciera que en lugar de impugnar las posiciones de Searle, Harnad está acogiéndolas tácitamente. Veamos los presupuestos claves que subyacen hasta el momento la propuesta del funcionalismo robótico:

i) La caracterización de procedimientos computacionales es insuficiente para poder replicar los procesos cognoscitivos. Hace falta un implantación eficaz, lo cual implica la necesidad de proveer una base material conectada sensorialmente con el medio, en forma directa y autónoma.

ii) La determinación de inteligencia (de capacidad de entendimiento) de un ente cualquiera debe darse en términos de una prueba total de Turing, la cual "es una prueba informal y abierta de si la gente puede o no discriminar el desempeño de la simulación implantada respecto de un ser humano real" [Harnad, 1989: pag. 20]. En otras palabras, ninguna prueba de un dominio específico puede considerarse válida para determinar entendimiento: hace falta una prueba que incorpore la totalidad de las estructuras cognoscitivas en interrelación con su medio.

Como puede verse, de estos puntos se deduce que, lejos de rebatir algunas de las posiciones fundamentales de Searle, Harnad más bien las comparte, alejándose radicalmente en el proceso de las posiciones funcionalistas "clásicas". En particular, su caracterización de la prueba de Turing podría dar pie a la afirmación de Searle de que la base biológica (o una replicación exacta de la misma) es necesaria para la existencia de procesos verdaderamente inteligentes (para la existencia del entendimiento).

7. Intencionalidad

A raíz del argumento de Searle, se espera que cualquier concepción computacional de la mente brinde una explicación del origen de la intencionalidad, y es precisamente éste uno de los aspectos a los cuales Harnad presta atención.

Harnad se plantea el problema de la intencionalidad del siguiente modo:

> "¿Cómo es que los símbolos significan del todo algo? ¿Cómo pueden representar o referirse a objetos y estados de cosas en el mundo? A esto se le llama también el problema de la "intencionalidad" [Harnad, 1989: pag.14].

Es claro que *erróneamente* Harnad ha reducido el problema de la intencionalidad a un aspecto de la semántica (esto es, al problema de la referencia en el lenguaje), pues existen otros estados mentales en los cuales se da la intencionalidad y este aspecto semántico no es parte esencial en ellos. Por ejemplo, las llamadas actitudes proposicionales tales como creencias y temores son estados mentales relacionados con una proposición determinada: esta relación es su intencionalidad. Sin embargo, la intencionalidad en esos casos no consiste en el sentido de una representación, en la cual se encuentren los términos de un lenguaje en relación semántica respecto de la referencia de un signo. Considérese el siguiente caso: **Juan cree que no hay ladrones en el vecindario**. Es claro aquí que el objeto de la creencia de Juan es una proposición, pero aquí no existe una relación semántica de referencia entre esa proposición y el estado mental de creencia de Juan. Considérese también el caso **Juan tiene sed y piensa que desearía un vaso de agua**: aquí el objeto del estado mental es un vaso de agua, y la intencionalidad manifiesta en este estado mental no es, evidentemente, la relación de referencia. Por tanto, Harnad no podría haber resuelto el problema de la intencionalidad, sino a lo sumo un aspecto del problema semántico.

Ahora bien, no queda claro cuál tipo de solución proporciona Harnad al problema mismo de la referencia de los signos. Por una parte, pareciera sugerir que la intencionalidad (como la entiende Harnad) se deriva primariamente de una intencionalidad intrínseca existente en lo que él llama códigos no-simbólicos, los cuales se diferencian de los simbólicos. Un código simbólico es un

> "...conjunto de muestras ("tokens") físicas, las cuales son manipuladas en virtud de su forma (arbitraria) de acuerdo con ciertas reglas formales; la relación entre las muestras y lo que ellas "representan" depende de una convención o sistema notacional interpretativo (...) esto es, la relación muestra/objeto es 'derivada' en el sentido de Searle." [Harnad, 1989: pags. 15-16].

Por otra parte, un código no-simbólico es

> "...aquél en el cual la relación entre las muestras simbólicas y lo que representan no es arbitrario o convencional, sino que está gobernado por la física de alguna manera, e.g. a través de conexiones causales confiables entre propiedades físicas similares, tales como la forma. La relación no-simbólica muestra/objeto es 'intrínseca'" [Harnad, 1989: pag. 16].

La naturaleza de esta relación no-simbólica no es del todo clara; Harnad no desarrolla este punto, lo cual sorprende por la importancia que tiene en toda su argumentación. Pudiera ser que se estuviera refiriendo a relaciones parecidas a las que se darían, por ejemplo, entre un sello y la impresión dejada por éste en un poco de cera. La impresión en la cera y el sello tendrían una forma en común y, de este modo, la relación de representación entre la impresión en la cera y el sello parecería no ser convencional. Según esta interpretación, los códigos no-simbólicos serían especies de "pinturas" de los objetos que representan. Es decir, los códigos existirían respecto de ciertos objetos con los cuales tendrían una relación "pictórica". Sin embargo, no podemos asegurar que esto es, precisamente, lo que Harnad tiene en mente relativo a la relación no-simbólica.

Los códigos no-simbólicos son ejemplificados en el ser humano en el caso de la percepción y constituyen la posible fuente de la relación semántica de referencialidad:

> "El código neural puede ser en parte analógico, esto es, puede preservar la "forma" de la entrada de manera bastante fiel...pero es todavía un código: nunca es la "cosa en sí misma" la que participa en un estado interno, solo su código sensorial. De este modo, si ha de haber una semántica verdadera en la mente (y seguramente la hay), debe derivarse, por lo menos en parte, de la interacción causal con el mundo exterior" [Harnad, 1989: pag. 16].

Tal pareciera que lo que Harnad sostiene es lo siguiente: existe una relación de referencia intrínseca en ciertos elementos del ser humano (tal como la percepción) y de aquí es que surge el concepto semántico de referencialidad, pues esos elementos se refieren a ciertos objetos externos. Suponiendo que existe la forma de intencionalidad (o más bien de referencialidad) intrínseca de la cual habla Harnad, no vemos cómo la mente a partir de esto pueda generar la idea o concepto de referencialidad esencial para la semántica, pues en ningún momento se presentarían ante la mente esos códigos no-simbólicos (fuente de la referencialidad intrínseca) como representando objetos. En otros términos, la mente no tiene acceso directo a la relación de representabilidad de un objeto por un código no-simbólico, porque estos códigos constituyen el único acceso directo de la mente a las cosas en sí mismas. La mente no puede salirse de sí misma, por decirlo así, y contemplar cómo su percepción se encuentra en relaciones de referencia con los objetos de los cuales esas percepciones son sus códigos no-simbólicos.

Por otra parte, Harnad sugiere que la semántica podría provenir de niveles jerarquizados de procesamiento de información:

> "Tal vez la semántica también se derive en parte de la existencia de niveles jerárquicos de procesamiento de información, con los elementos no interpretados de un nivel deviniendo los patrones elementales de otro: por supuesto, resultaría que la mayoría de los procesos cognoscitivos obtendrá por completo debajo del nivel de la consciencia" [Harnad, 1989: pag. 16].

Cómo es que la semántica se deriva de esos supuestos niveles no es, en ningún momento, especificado o explicado por Harnad. Esto significa que es una mera especulación, una hipótesis sin ningún fundamento.

La hipótesis sobre jerarquías de procesamiento de datos, así como la idea de una referencia intrínseca en códigos no-simbólicos, es todo lo que Harnad nos ofrece como explicación del origen de la semántica. Es claro, de nuestras observaciones anteriores, que ninguna de esas dos posibilidades ofrecen la explicación esperada.

8. Un argumento en contra del Funcionalismo Robótico

Como hicimos notar en el apartado **3**, Harnad sostiene que el funcionalismo robótico es inmune al argumento del cuarto chino. En el presente apartado queremos mostrar que esto no es así. Considérese el siguiente escenario:

Supóngase que tenemos un programa que permite reconocer patrones de firmas y una base de datos que incluye las firmas de los ahorrantes del Banco Z. Además de este programa y de esta base de datos, tenemos un "scanner" que nos permite leer firmas directamente de un cheque y pasar la lectura al computador. El sistema completo claramente podría así decidir si la firma es o no de uno de los ahorrantes del banco en cuestión. ¿Podríamos decir en este caso que el sistema como un todo percibe una firma cuando se le presenta un cheque firmado? Creemos que no, y nuestra razón se basa en la posibilidad de formular un razonamiento análogo, para este caso, al del cuarto chino. Tómese un individuo que ignore por completo lo que es una firma. Désele el programa que identifica patrones de firmas y la base de datos de los ahorrantes. Cada vez que se le pasa un papel con algo escrito en él y se le pregunta si es la firma de un ahorrante, entonces el individuo dentro del cuarto identifica lo escrito como un patrón de firma, lo compara con la lista de ahorrantes y responde. Para quien pregunta, el individuo del cuarto definitivamente sabe lo que es una firma, esto es, percibe firmas. Sin embargo, lo único que

percibe el individuo dentro del cuarto es un montón de rayas sin ningún sentido para él. No podemos decir, por tanto, que el sistema *percibe firmas*, sino más bien que recibe datos sensoriales, los cuales compara con patrones y correlaciona con nombres de individuos.

Este caso de las firmas perfectamente escapa lo planteado por Harnad de que al simular las funciones del robot (esto es, al poseer el individuo sus propios transductores) se perciben objetos. El individuo dentro del sistema que identifica firmas asume todas las funciones del robot que identifica firmas, sin que podamos decir que posee la propiedad mental de percibir firmas. El funcionalismo robótico no logra evadir definitivamente el punto esencial del *Argumento del Cuarto Chino*.

9. Conclusiones

Aunque hay ciertos aspectos de la posición de Harnad que no han sido examinados en este artículo (e.g. el argumento de la convergencia, la distinción entre modelación del cerebro y modelación de la mente, la hipótesis de la modularidad, y el tema general de la IA "débil" versus la IA "fuerte"), creemos haber demostrado que el **funcionalismo robótico**, al menos como lo plantea Harnad en su artículo Minds, machines and Searle, no escapa de las dificultades puestas de manifiesto por el argumento del cuarto chino. Además, esperamos haber mostrado algunos de los problemas fundamentales que enfrenta cualquier teoría que asuma una visión computacional de la mente, en especial los problemas centrales de la intencionalidad y de la semántica.

Estos problemas son medulares en ciencias cognoscitivas, no importa cuál sea la posición que se asuma. En artículos posteriores pensamos poder plantear algunas alternativas que, sin partir de presupuestos funcionalistas, ofrecen una esperanza para la comprensión de estos fenómenos, cuya explicación es esencial para cualquier teoría relativa a los procesos cognoscitivos y que se ocupe de la mente en general como objeto de estudio científico.

BIBLIOGRAFIA

Churchland, Paul. M. **Matter and Consciousness** (third printing). MIT Press, Cambridge, MA, 1990.

Harnad, Stevan. *Minds, machines and Searle*. **J. Expt. Theor. Artif. Intell.** 1 (1989) 5-25, 1989.

Hayes, P. & Ford, K. *Escaping the Chinese Room*. Por publicar.

Searle, J. R. *Minds, brains, and programs*. **Behavioral and Brain Sciences**. 3: 417-424, 1980a

Searle, J. R. *Intrisic Intentionality*. **Behavioral and Brain Sciences**. 3: 450-457, 1980b.

Searle, J. R. *The Chinese room revisited*. **Behavioral and Brain Sciences**. 5: 345-348, 1982.

Searle, J. R. *Patterns, symbols, and understanding*. **Behavioral and Brain Sciences**. 8: 742-743, 1985a.

Searle, J. R. **Minds, Brains and Science**. Harvard University Press, Cambridge, MA, 1985b.