Intencionalidad sin naturalismo biológico

(Intentionality without biological naturalism)

Ivar Hannikainen

University of Sheffield
i.hannikainen@sheffield.ac.uk

Recibido: 21 de abril de 2011
Aceptado: 7 de septiembre de 2011

Resumen

La habitación china es una versión del test de Turing que permite a Searle defender su naturalismo biológico, según el cual la computación no es ni suficiente ni constitutiva de la mente. En este ensayo, investigo las dos versiones de su postura anticomputacionalista, sostengo que la computación forma parte de la comprensión del lenguaje natural y sugiero una vía para la reducción fisicalista de la intencionalidad en los actos de habla proposicionales.

Palabras clave: funcionalismo, John Searle, intencionalidad, teoría computacional de la mente.

Abstract

The Chinese Room Argument is a variant of Turing's test which enables Searle to defend his biological naturalism, according to which computation is neither sufficient nor constitutive of the mind. In this paper, I examine both strands of his anti-computationalist stance, argue that computation is constitutive of natural language understanding and suggest a path toward the physicalist reduction of intentionality for propositional speech acts.

Keywords: functionalism, John Searle, intentionality, computational theory of mind.

La *teoría funcionalista de la mente* (TFM) proporciona una solución al problema de las otras mentes. Si consideramos que una determinada tarea, *t*, es una tarea *inteligente*$_t$, atribuimos cierta inteligencia$_t$ a una persona capaz de realizar dicha tarea. ¿Podríamos decir entonces que un ordenador con la misma capacidad sea inteligente$_t$? Ciertos programas de ordenador son capaces de jugar al ajedrez o realizar operaciones aritméticas mucho mejor que la persona de inteligencia media. Cabe decir entonces que los ordenadores , armados con el programa adecuado, exhiben inteligencia$_{ajedrez}$ o inteligencia$_{aritmética}$. Según la TFM, si consideramos que una tarea inteligente o una serie de ellas, T_n, son propias de una mente, entonces podemos atribuir una mente a una persona capaz de realizar T_n. La *tesis fuerte de la Inteligencia Artificial* (TFIA) supone que podemos, por el mismo criterio funcionalista, atribuir una mente a una computadora digital.

Searle ha sido durante décadas recientes uno de los más vociferantes exponentes en contra de ambas (i) la TFIA, y (ii) la TFM. Su famoso *argumento de la habitación china* (AHC) es la "simple y decisiva refutación" de ambas tesis. El AHC se fundamenta en dos axiomas del pensamiento de Searle:

La intencionalidad es la propiedad característica de la mente. **[I]**
Ciertos procesos cerebrales son suficientes para la intencionalidad. **[NB]**

La consecuencia directa de estas dos premisas de Searle es que:

Ciertos procesos cerebrales son suficientes para una mente.

[I] supone que la intencionalidad es el *sine qua non* de la mente: lo característico de las mentes es su capacidad intencional y no su capacidad funcional. Por lo tanto, ningún test funcional, "desde un punto de vista externo," como el test de Turing, es decisivo para la identificación y atribución de una mente.[1] Según [NB], la intencionalidad es efecto de las propiedades causales del cerebro. Los humanos y ciertas especies evolutivamente desarrolladas tienen intencionalidad por ser "un cierto tipo de organismo con cierta estructura biológica (i.e. química y física), y esta estructura, bajo ciertas condiciones, es causalmente capaz de producir percepción, acción, comprensión, aprendizaje y otros fenómenos intencionales."[2] Esta es entonces la alternativa de Searle, el *naturalismo biológico*, al problema de las otras mentes. Searle entiende que la conciencia es una experiencia subjetiva consecuencia de una bioquímica cerebral específica. Presumiblemente, es sólo desde un punto de vista interno – siendo nosotros mismos el sujeto en cuestión – que podríamos real-

[1] Turing (1950).
[2] Searle (1980); cf. Searle (1984, 1990).

mente cerciorarnos de que existe intencionalidad. Mas, sosteniendo la teoría [NB] de la mente, podemos suponer que otras personas como nosotros son capaces de intencionalidad, y por lo tanto, tienen una mente. Puesto que lo que define una mente es su capacidad intencional, y esto es consecuencia de ciertos procesos cerebrales, nadie ni nada que carezca de estos procesos cerebrales podrá tener una mente. De esta manera Searle refuta tanto la TFM como la TFIA. He aquí que Searle sostenga la postura que podríamos llamar anticomputacionalista, según la cual:

La computación no es ni suficiente ni constitutiva de una mente. **[AC]**

Para el tratamiento a lo largo de este ensayo de la postura anticomputacionalista, la dividiremos en sus dos partes, anticomputacionalismo fuerte y débil:

La computación no es constitutiva de una mente. **[AC+]**
La computación no es suficiente para una mente. **[AC-]**

Manteniendo el criterio de la intencionalidad como la propiedad característica de lo mental, [I], y suspendiendo la creencia en el naturalismo biológico de Searle, [NB], en este ensayo preguntaré por las dos versiones de su anticomputacionalismo y plantearé la posibilidad de una reducción fisicalista de la intencionalidad. En la Sección II, profundizaré en el concepto de la intencionalidad con el propósito de hacer una reducción fisicalista (y no biológica como la de Searle) de ésta para los actos de habla proposicionales. En la Sección III, cuestionaré la veracidad de la versión fuerte del anticomputacionalismo de Searle, [AC+]. En la Sección IV, haré lo mismo con la versión débil, [AC-], tanto para una computadora digital (como las que planteaba Turing) como para un sistema híbrido con capacidades sensomotores.

1. Intencionalidad y actos de habla

Recordemos que la intencionalidad, desde Brentano – autor que reintroduce en la psicología fenomenológica este concepto –, es aquella capacidad de los estados mentales de *ser conciencia de* un objeto o estado del mundo:

> Todo fenómeno mental se caracteriza por lo que los Escolásticos de la Edad Media llamaban la inexistencia intencional o mental de un objeto, y lo que podríamos llamar, aunque no del todo inequívocamente, referencia a un contenido, y dirección hacia un objeto (lo que no debe entenderse aquí como significando algo), o objetividad inmanente. Todos incluyen algo como objeto dentro de sí, aunque no todos lo hagan de la misma manera. En la presentación, algo es presentado, en el juicio algo afirmado o negado, en el amor amado, en el odio odiado en el deseo deseado, etc.

> *Esta in-existencia intencional es característica exclusiva de los fenómenos mentales. Ningún fenómeno físico exhibe algo similar.* Podemos, por lo tanto, definir los fenómenos mentales diciendo que son aquellos que contienen objetos intencionales dentro de sí.[3]

Searle ha defendido que la noción de intencionalidad es irreducible; es decir, que "no hay una explicación no-intencional de la intencionalidad."[4] Un repaso por ciertos textos clave en filosofía de la mente sobre la intencionalidad y en la filosofía del lenguaje sobre el significado, servirá para abrir una vía a la reducción de la intencionalidad en los actos de habla proposicionales, α*(Px)*. Empecemos con Husserl y Frege viendo un tratamiento de los juicios ['*Urteile*']. En Husserl:

> ... dos representaciones son en esencia la misma, si sobre la base de cada una de ellas, y tomadas puramente en sí (es decir, analíticamente), enuncian exactamente lo mismo, y nada más, sobre la cosa representada. Y análogamente con respecto a los demás actos. Dos juicios ['*Urteile'*] son esencialmente el mismo juicio, cuando todo lo que podría decirse sobre los hechos a partir de un juicio se aplica también al otro, y no de lo contrario. Su valor de verdad es el mismo, y lo es obviamente cuando el juicio, la esencia intencional como unidad de la cualidad y de la materia del juicio, es el mismo.[5]

Y en Frege:

> Las oraciones 'M dio a N el diploma A', 'el diploma A fue dado a N por M', expresan exactamente el mismo pensamiento: no hay la menor diferencia entre lo que se sabe por medio de cada una de estas oraciones. Por ello tampoco es posible que una de ellas sea verdadera y la otra falsa.[6]

O:

> …los contenidos de dos juicios ['*Urtheile*'] de doble manera pueden ser distintos: primero, que las consecuencias que se puedan derivar de uno, en combinación con otros juicios determinados, se sigan también del otro, en combinación con los mismos otros juicios; en segundo lugar, que este no sea el caso. Las dos proposiciones: 'en Platea derrotaron los griegos a los persas' y 'en Platea fueron derrotados los persas por los griegos,' se distinguen de la primera manera. Aun cuando se puede reconocer una pequeña diferencia en el sentido, la concordancia, no obstante, prevalece. Asi, a aquella parte del contenido que es la *misma* en ambas, la llamo el *contenido conceptual*.[7]

---

[3] Brentano (1874), pp. 115-6, énfasis añadido. Habría que recordar que, para Searle, no es verdad que todos los estados mentales sean intencionales, sino que solamente los estados mentales pueden ser intencionales.

[4] Searle (1979), p. 195.

[5] Husserl (1975), p. 74.

[6] Frege (1969), p. 153.

[7] Frege (1879), pp. 2-3.

Los pares de proposiciones que Frege maneja, al tener el mismo contenido conceptual y el mismo valor de verdad, han de tener también – como señala Husserl – la misma esencia intencional, φ*(Px)*. Cualquier α*(Px)* tiene una φ*(Px)* que es la misma que el contenido conceptual de la proposición, *Px*, que contiene. Para *Px*: 'hoy hace sol', podemos incluir como sus α*(Px)*, la aserción '*Hoy hace sol*', la negación '*Hoy no hace sol*', la creencia '*Creo que hoy hace sol*', el deseo '*Me encantaría que hoy hiciese sol*', y la pregunta '*¿Hace sol hoy?'*. En todos estos α*(Px)*, φ*(Px)* es la misma. Y es la misma también que el contenido conceptual de 'hoy hace sol'.

> *Para cualquier acto de habla proposicional,* α(Px*), la esencia intencional,* φ(Px*), es el contenido conceptual de la proposición que contiene,* Px. **[I_1]**

φ*(Px)* y el contenido conceptual de una proposición tienen otras similitudes. En su tratamiento de la intencionalidad del acto perceptivo, Gurwitsch dice: "El noema, distinto del objeto real al igual que del acto [mental], termina siendo una entidad irreal o ideal que pertenece a la misma esfera que los significados o las significaciones."[8]

Volvamos a Frege, quien nos recuerda que las representaciones o imágenes mentales no son el significado: "No se debe, empero, confundir esta imagen [mental] con el sentido de la palabra 'caballo', pues no hay indicación alguna en la palabra 'caballo' acerca del color del caballo, si está en reposo o en movimiento, el lado desde el que se lo ve, etc."[9] Característico de φ*(Px)* o del noema de la percepción, al igual que del sentido de un término, es esta *unidad ideal*. La unidad ideal de φ*(Px)* es tal que el sujeto "puede revertir a ella un número indefinido de veces."[10]

> *La esencia intencional,* φ(Px*), y el sentido de una proposición son ideales, esto es, no son ni los objetos específicos ni las representaciones mentales de estos.* **[I_2]**

Por último, cabe decir algo sobre la relación entre el estado mental intencional, ψ*(Px)*, y el contenido semántico:

> La conciencia ha de ser definida por su referencia a la esfera del sentido, tal que *experimentar un acto es lo mismo que actualizar un sentido*… ningún estado mental puede ser explicado, salvo en relación al sentido objetivo ['gegenständlicher Sinn'], del que el sujeto se apercibe mediante este acto.[11]

---

8 Gurwitsch (1968), p. 76.
9 Frege (1969), p. 151.
10 Gurwitsch (1968), p. 82.
11 Ibid.

Searle es consciente de este paralelismo entre los α*(Px)* y los ψ*(Px)*. El mismo Searle sostiene que "la ejecución del acto [de habla] es *eo ipso* la expresión del estado intencional correspondiente y el contenido proposicional del acto y del estado son idénticos."[12] Por último, en el *Tractatus Logico-Philosophicus* de Wittgenstein:

> 3.11 El método de proyección es el pensar el sentido de la proposición.[13]

En distintas palabras, Gurwitsch, Searle y Wittgenstein expresan la misma idea: que el estado mental intencional, ψ*(Px)*, es la 'actualización', 'ejecución' o 'proyección' del sentido o contenido semántico de *Px*.

> *Pensar el sentido de una proposición,* Px*, es estar en el estado mental intencional correspondiente,* ψ(Px*).* **[I_3]**

Resumiendo, todo α*(Px)* es intencional. φ*(Px)* es el contenido conceptual de *Px*, que es una unidad ideal. Pensar φ*(Px)* es estar en ψ*(Px)*, que es un solo estado correspondiente a todo α*(Px)*. Para cualquier ψ*(Px)*, φ*(Px)* es el contenido semántico de *Px*. Esto no es lo mismo que decir que ψ*(Px)* es el significado de *Px*, sino que es éste *pensado* o *proyectado*. Entonces, una persona, máquina o ente cualquiera que piense el sentido de la proposición *Px*, capta la esencia intencional de cualquier α*(Px)*, y consecuentemente está en el estado mental intencional correspondiente, ψ*(Px)*.

En esta sección, he intentado un breve repaso de la filosofía de la mente y del lenguaje para sugerir, en contra del alegato de Searle, ciertos paralelos reduccionistas del concepto de la intencionalidad en los actos de habla. De esta manera, convertimos la intencionalidad de un acto mental en algo empíricamente tratable (y no en un argumento total, extensión del [NB], contra cualquier desarrollo de la IA) y físicamente implementable, para abrir la posibilidad de hablar de estados computacionales intencionales.

2. Computación y sintaxis

Recordemos que el test de Turing sugiere una *máquina de Turing* que, para todo input *i*, produce un output *o* indiscernible de la respuesta de un ser humano. Se sigue de la postura de Searle que da igual que clase de instrucciones se le da a un ordenador, que no va a comprender [AC-], ni va a acercarse a la comprensión del lenguaje [AC+]. Esto, recordemos, es lo mismo que decir que la computación no es si suficiente ni constitutiva de una mente.

---

[12] Searle (1979), p. 192.
[13] Wittgenstein (1921).

Supongamos un *i* y un *o* correspondientes a un test de Turing para el francés, tales que *i*: *'Est-ce que un lévrier peut courir?,'* y *o*: '*Oui, un lévrier peut courir.'* Veamos entonces dos habitaciones francesas (HFs) y sus algoritmos correspondientes:

HF1

Cambia el ultimo carácter de *i* de '*?*' a '.' ➔ Empieza por el principio de *i* a borrar caracteres hasta que llegues a un espacio, ' ' ➔ Repite el paso anterior ➔ Inserta '*Oui,*' al principio de *i* ➔ Imprime la cadena como *o*.	**[λ1]**

HF2

Para todo *i*: '*Est-ce que* Px?', donde *Px* es una proposición, si *Px* e Vr, el *o* es "*Oui,* Px."	**[$\lambda 2_i$]**
Px es verdad si el predicado *P* se refiere al objeto *x*.	**[$\lambda 2_j$]**

De acuerdo con el AHC, ninguna de las dos habitaciones – ni HF1 ni HF2 – demuestran comprensión alguna del francés. Da igual si el algoritmo lo ejecuta una persona o un ordenador, la HF no comprende francés. (Si lo ejecuta una persona y ella entiende francés, el sistema es intencional; y si no, no.) En cualquier caso, el algoritmo de la HF no aporta nada en cuanto a comprensión del lenguaje.

Entonces supongamos que un señor llamado Gervasio, aprendiz del francés que sólo conoce algunas palabras, se adentra en las HFs. Su conocimiento del francés no es suficiente para entender *i*, pero por casualidad sabe que, traducido al castellano, *lévrier* es galgo y *courir* es correr. Si metemos a Gervasio en la HF1, tomará *i* y reconocerá dos de las pocas palabras que conoce, *lévrier* y *courir*. Al no conocer el resto de las palabras ni saber gramática del francés, solamente puede intentar adivinar el significado de *i*. Sigue las instrucciones de [λ1], el algoritmo de HF1, y genera *o*. En ningún momento, Gervasio alcanza comprensión alguna del contenido proposicional de *i*. [λ1] es tal que, como en el caso de Searle con el chino, no aporta nada de comprensión.Veamos qué pasa si metemos a Gervasio en la HF2. Tomará *i* y seguirá las instrucciones correspondientes, [$\lambda 2_i$] y [$\lambda 2_j$]. Tras [$\lambda 2_j$], terminará con la siguiente instrucción:

> *un lévrier peut courir* (*Px*) es verdad si el predicado *courir* (*P*) se dice del objeto *lévrier* (*x*).

Ahora, al saber traducir al castellano los signos *courir* y *lévrier*, lo hace; y así entiende conoce las condiciones de verdad de *Px*:

'un lévrier peut courir' es verdad si el predicado *correr* se dice del objeto *galgo*.

Aplicando un poco de gramática del castellano, la cual Gervasio conoce:

'un lévrier peut courir' es verdad si los galgos pueden correr.

Gervasio deduce que *Px* es verdad y – volviendo a [$\lambda 2_i$] – que el *o* del sistema es '*Oui, un lévrier peut courir.*' En el caso de HF2, Gervasio ha comprendido *Px*. El algoritmo de HF2, [$\lambda 2$], ha ayudado a Gervasio a tener una cierta comprensión. Indudablemente su limitado conocimiento de ciertas palabras en francés también ha contribuido; pero, como es evidente en HF1, por si sólo esto no es suficiente para que Gervasio entienda el contenido proposicional de *Px*.

En ejemplos de proposiciones más complejas, el papel de la computación en la comprensión del lenguaje se ve más claramente. Esta vez veamos un acto proposicional condicional:

Est-ce que les enfants puissent jouer sous la pluie sans un superviseur? **[α(*Jn*)]**

El contenido proposicional, *Px*, es: *les enfants peuvent jouer sous la pluie sans un superviseur*. Un parser de gramática del francés como HF2, suficientemente potente, sería capaz de analizarlo como un condicional que consta de dos condiciones: si no (*S*) hay supervisor en patio y (*L*) está lloviendo, entonces (*Jn*) los niños pueden jugar en el patio.

Entonces *Px*, '*les enfants peuvent jouer sous la pluie sans un superviseur*', es verdad siempre y cuando:

$(\neg S \wedge L) \rightarrow Jn$ **[CV(*Jn*)]**

Entender [α(*Jn*)] (o la expresión equivalente en español o inglés) requiere que hagamos este análisis lógico-sintáctico del contenido proposicional. Esto es lo que quiere decir Wittgenstein cuando dice:

> 4.024. Comprender una proposición es saber lo que es el caso si es verdadera.
> (Cabe, pues, comprenderla sin saber si es verdadera.)[14]

[14] Ibid.

El procedimiento de transformación de [α(*Jn*)] en sus condiciones de verdad, [CV(*Jn*)], es indistinguible del procedimiento que debe seguir una persona para entender la pregunta y su contenido proposicional. De acuerdo con Wittgenstein, se puede entender la pregunta sin saber responderla. Y el mero hecho de no tener la más remota idea de cuál es la respuesta adecuada – si afirmativa o negativa – no parece restar de nuestra certidumbre de que entendemos la pregunta.

En definitiva, ¿es cierta [AC+]? ¿La computación es o no es constitutiva de una mente? Parece aquí que Searle está equivocado. La computación forma parte integral de nuestra manera de procesar los pensamientos. Nuestra comprensión de los actos proposicionales parece requerir una manipulación sintáctica (por compleja que a veces sea) que nos permite derivar de cualquier acto proposicional sus condiciones de verdad. Como vimos mediante HF1 y HF2, ciertos algoritmos contribuyen a la actividad inteligente y otros no. No es para nada trivial de cara a la comprensión (como parece sugerir Searle en el AHC) cuáles son las instrucciones que componen el programa de ordenador. El algoritmo correcto de manipulación sintáctica, lo ejecute una persona o una máquina, forma parte de la comprensión de un α(*Px*).

4. Verdad y realidad

4.1 La computadora digital de Turing

Si comprender una proposición es saber lo que es el caso si es verdadera – sus condiciones de verdad – como parece sugerir Wittgenstein en § 4.024, entonces un algoritmo lo suficientemente potente como para derivar de todo α(*Px*) sus condiciones de verdad o satisfacción, comprendería cualquiera de ellos. Pero la consecuencia de esto sería una HF con estados de máquina como:

i: *'Est-ce que le ciel est bleu?'* **[α(*Bc*)]**
Le ciel est bleu es verdadera si *être bleu* es verdad de *le ciel*.

o: '*Oui, si c'est vrai que le ciel est bleu.*'
i: *'Est-ce que les dauphins sont poilus?'* **[α(*Pd*)]**

Les dauphins sont poilus es verdadera si *être poilu* es verdad de *les dauphins*
o: '*Peut être, si c'est le cas que les dauphins sont poilus.*'

Para un funcionalista, amigo del test de Turing, este comportamiento es insuficiente porque no es indistinguible del de una persona inteligente. En una gran variedad de casos parecidos a [α(*Bc*)] y [α(*Pd*)], una persona inteligente sería capaz de

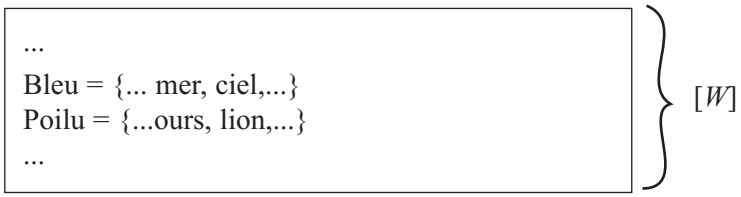
dar una respuesta afirmativa o negativa (un valor de verdad). Para un defensor de [I], la razón sería otra. Si para cualquier *i* del tipo '*Est-ce que* Px*?,*' e independientemente del contenido de *Px*, el tratamiento que hace HF de *i* es siempre el mismo, no es posible que dicho tratamiento sea intencional. Para dos *i*s cualesquiera como [α(*Bc*)] y [α(*Pd*)] tales que φ*(Bc)* y φ*(Pd)* no son equivalentes, y aplicando el mismo algoritmo (digamos [λ2]), los estados computacionales de la HF en ambos casos serán los mismos. Por lo tanto, no es posible que estos estados computacionales capten dos esencias intencionales distintas, ψ*(Bc) y* ψ*(Pd)*, ni que por tanto sean intencionales en absoluto.

Como hemos dicho, para aprobar un test funcionalista, la HF tendría que dar una respuesta, afirmativa o negativa, a [α(*Bc*)] y [α(*Pd*)]. Según Wittgenstein:

> 2.223 Para reconocer si la figura es verdadera o falsa, tenemos que compararla con la realidad.[15]

En los casos de [α(*Bc*)] y [α(*Pd*)], una persona de inteligencia media consultaría su conocimiento adquirido del mundo; mientras que un ordenador consultaría en su memoria su conocimiento de la realidad, *W*, i.e., la unión de conjuntos finitos de proposiciones primitivas verdaderas sobre el mundo:

W = Bleu **U** Rouge **U** ... **U** Poilu **U** Écaillé



Lo que afirma la TFIA es que una máquina de Turing con un *W* en su memoria comparable al conocimiento de la realidad de una persona inteligente, sería capaz de pasar un test de Turing exhaustivo y por lo tanto sería indistinguible funcionalmente de una mente.

Aquí un defensor de [I] debe estar de acuerdo con Searle en que no hay intencionalidad (aunque no necesariamente por [NB], sino por otras razones que se aprecian en el AHC). Un sistema cuyo conocimiento de la realidad se fundamenta en proposiciones primitivas guardadas en su memoria todavía no es intencional. En el *Tractatus*, Wittgenstein explica:

[15] Ibid.

> 2.0211 Si el mundo no tuviera sustancia alguna, el que una proposición tuviera sentido dependería de que otra proposición fuera verdadera.[16]

Un sistema cuyo conocimiento de la realidad se fundamenta en proposiciones primitivas es 'como si el mundo no tuviera sustancia alguna.' Las proposiciones son lo que Wittgenstein llamaba figuras de los *hechos* y no de los *objetos* (que son la *sustancia*), pero:

> 2.026 Sólo si hay objetos puede haber una forma fija del mundo.[17]

Esto quiere decir que el conocimiento del mundo basado en proposiciones primitivas no se hace cargo de la forma fija o sustancia del mundo. *Ser conciencia de* un mundo externo – recordemos – es la definición de lo intencional, por lo tanto la comprensión de un $\alpha(Px)$ mediante un conocimiento como W, no puede constituir $\psi(Px)$: es decir, pensar el mundo en base a proposiciones primitivas no puede ser una actividad intencional.

Esta es otra manera de enfocar la común crítica que dice "la sintaxis no garantiza la semántica." Aún con un conocimiento profundo mediante W de los hechos del mundo, una máquina no tiene conocimiento de los objetos (o la sustancia); y es este conocimiento de los objetos que nos permite interpretar o asignar contenido semántico a las expresiones del lenguaje. Cuando, en última instancia, decimos que una máquina no entiende una proposición como "*Le ciel est bleu*," no es por no saber las condiciones de verdad o el valor de verdad de la proposición, sino por no conocer los objetos del mundo a los que se refiere: no tener la menor idea lo que es el cielo, ni lo que es el azul. Es por esto que decimos que una máquina realmente no sabe ni que el cielo es azul ni que los delfines no son peludos. Puesto que la mente implica un conocimiento de los objetos del mundo, [AC-] es verdad, la computación (en el sentido de la 'computadora digital' de Turing) no es suficiente para una mente.

4.2 El robot perceptivo

El problema de cómo un sistema puede tener conocimiento de los objetos a los que se refieren los símbolos que maneja ha venido a llamarse el *symbol grounding problem* en la obra de Stevan Harnad, siendo la pregunta central:

> ¿Cómo puede hacerse intrínseca al sistema la interpretación semántica del sistema formal de símbolos, en vez de parasítica sobre los significados en nuestras cabezas?[18]

[16] Ibid.
[17] Ibid.
[18] Harnad (1990).

Para Harnad, la solución recae en simular ciertas capacidades no-simbólicas o prelingüísticas de la mente: específicamente, las habilidades de *discriminación e identificación mediante representaciones categóricas*. Para esto, concibió un *sistema híbrido computacional-sensomotor* equipado con transductores y redes neuronales. Entrenadas para categorizar, las redes neuronales construyen representaciones internas que comprimen diferencias dentro de las categorías y las expanden entre las categorías… Los nombres se adjudican mediante conexiones por red a las proyecciones sensoriales de los objetos por los que están.[19] Las redes neuronales, como ya se sabe, son relativamente dadas a aprender a detectar las características invariantes de una serie de inputs visuales y sortearlos de manera específica (i.e. discriminación). Una vez sorteadas las categorías, el programa "sería capaz de asignar una única respuesta – un nombre – a una clase de inputs, tratándolos como equivalentes o invariantes en algún respecto"[20] (i.e. identificación).

Sosteniendo que en ellas se basa la comprensión del lenguaje, Harnad propone un test de Turing (de 'indistinguibilidad funcional total'[21]) para estas habilidades mentales no-simbólicas y prelingüísticas. Así, podemos rescatar la TFM, sosteniendo que el test de Turing tradicional simplemente no es un test funcionalista de la mente *comprehensivo* ya que se olvida de funciones de cruciales para la atribución de una mente: la discriminación e identificación de objetos percibidos. Cangelosi y Harnad diseñaron prototipos de estos sistemas híbridos capaces de identificar, y discriminar entre, distintos tipos de inputs. Para una variedad de inputs – círculos, elipses, cuadrados y rectángulos – los sistemas híbridos de Cangelosi y Harnad fueron capaces de nombrarlos acertadamente. Además fueron capaces de aprender sus propiedades de simetría y asimetría. En otro prototipo, las redes neuronales aprendieron a clasificar caballos y cebras, y aprendieron las propiedades de tener y no tener rayas.[22]

¿Qué implicaciones tiene esto para la interpretación del contenido semántico de los términos? Si después de una etapa de aprendizaje (análoga al aprendizaje que requieren también los niños para procesar términos lingüísticos), para un número indefinido de tests, el sistema es capaz de identificar correctamente el objeto que se presenta ante su aparato perceptivo, podemos decir que la carga semántica ya no recae sobre nuestros hombros. Los términos ya no son, como Searle los llama, 'símbolos formales no-interpretados' ya que el propio sistema híbrido correctamente identifica elementos de la extensión de los términos que emplea: un método perfectamente común para asegurarnos de que alguien conoce la referencia de un término recientemente aprendido. Así, demostrando como un sistema híbrido computacio-

[19] Cangelosi et al. (2000).
[20] Harnad (1990).
[21] Harnad (2002).
[22] Cangelosi et al. (2000).

nal-sensomotor podría fija la referencia de los términos que maneja, Harnad defiende que el *symbol grounding problem* es superado. Por ejemplo:

> *¿Son simétricos los círculos?*
> *Los círculos son simétricos* es verdadera si (P) *ser simétricos* se predica de (x) *los círculos*.
> *¿Tienen rayas los caballos?*
> *Los caballos tienen rayas* es verdadera si (P) *tener rayas* se predica de (x) *los caballos*.

La capacidad de identificación de objetos mediante representaciones categóricas es, a la inversa, la capacidad de conocer la extensión de los términos que emplea (de manera reminiscente de las 'directions of fit' del mismo Searle, i.e., 'word-to-world' y 'world-to-word'). Es decir, cabe suponer que cuando el sistema híbrido procesa un término primitivo, *P* o *x* – tanto cuando lo lee en un *i* como cuando lo escribe en un *o* – ya no entiende por él el mero nombre de un elemento o de un conjunto en *W* (como en el caso de la computadora digital) sino, más bien, una representación categórica formada a partir de visualizaciones del objeto mismo en el mundo, i.e, la abstracción de características invariantes en una recolección de representaciones icónicas de objetos pertenecientes a la extensión del término. La respuesta – es decir, el valor de verdad de *Px* – que da el sistema híbrido computacional-sensomotor ahora depende de un acto de consulta, superposición o comparación de estas representaciones categóricas.

Si dijimos que un estado mental intencional, $\psi(Px)$, era el significado de *Px pensado* o *proyectado*, y que el significado de Px se fundamenta en las representaciones categóricas de los términos simples de *Px*, ¿es posible que la comprensión de un $\alpha(\mathrm{Px})$ mediante representaciones categóricas suponga la proyección de *Px* de manera indistinguible a la de una persona inteligente? Es decir, ¿son intencionales estos actos de consulta de las representaciones categóricas? ¿Son estos estados computacionales que se refieren a objetos del mundo mediante representaciones categóricas como los estados intencionales que *son conciencia de* los objetos o estados del mundo? Al menos, reúnen ciertas características de la intencionalidad sugeridas en la Sección II: hacen referencia a un objeto o estado del mundo; y lo hacen como unidades ideales, que ni son los objetos específicos, ni las representaciones icónicas de estos. Hasta donde una reducción fisicalista permite, son coherentes con lo que constituye el estado mental intencional de las personas.

5. Conclusión

La computación, entendida como el algoritmo correcto de manipulación sintáctica, forma parte integral de nuestra comprensión de los actos de habla proposicio-

nales. Sin embargo, ya que la intencionalidad implica un conocimiento de los objetos del mundo, la computación por sí sola no es suficiente para la intencionalidad. Como Harnad afirma, "aunque el AHC demuestra que la cognición no puede ser *toda* computacional, desde luego no demuestra que no pueda ser *en cierta medida* computacional. Aquí Searle parece haber sacado conclusiones más fuertes de las que el AHC justifica."[23] Es decir, [AC+] es falsa; mientras que, siempre que sostengamos [I], [AC-] será verdadera. Para tener intencionalidad, un sistema computacional tiene que tener conocimiento de los objetos del mundo, lo cual requiere interacción sensomotor con el mundo. Generando representaciones categóricas por las que están los términos que el sistema emplea, éste tiene una relación con el mundo externo adecuada para la superación del *symbol grounding problem*.

De acuerdo con el tratamiento reduccionista de la intencionalidad en los actos de habla, el procesamiento computacional de un acto proposicional compuesto por términos simples fundamentados en representaciones categóricas de objetos del mundo es análogo al estado mental intencional de una persona.

Referencias bibliográficas

Brentano, F. (1874): *Psychologie vom empirische Standpunkt,* Leipzig, Duncker & Humblot.

Cangelosi, A.; Greco, A. & Harnad, S. (2001): "From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories," *Connection Science* 12(2), pp. 143-162. Los penúltimos borradores de todos los artículos escritos o coescritos por Stevan Harnad son accesibles mediante: http://www.ecs.soton.ac.uk/people/harnad/publications.

Frege, G. (1879): *Begriffsschrift: eine der arithmetischen nachgebildete Formelsprache des reinen Denkens.* Halle, Louis Nebert.

Frege, G. (1897): "Logik" en H. Hermes, F. Kambartel y F. Kaulbach (eds.), *Nachgelassene Schriften und Wissensschaftlicher Briefwechsel*, Vol. 1, Hamburgo, Felix Meiner, 1969.

Gurwitsch, A. (1968): "On the intentionality of consciousness," en Farber, M. (ed.), *Philosophical essays in memory of Edmund Husserl,* Nueva York, Greenwood, pp. 65-83.

Harnad, S. (1990): "The symbol grounding problem," *Physica D* 42, pp. 335-346.

Harnad, S. (2002): "Minds, Machines and Searle 2: What's wrong and right about Searle's Chinese Room argument?," en Bishop M. & Preston J. (eds.), *Views into the Chinese Room: new essays on Searle and Artificial Intelligence*, Oxford, Oxford University Press, 2002.

[23] Harnad (2002).

Husserl, E. (1901): *Fünfte logische untersuchung*, Hamburgo, Felix Meiner, 1975.
Searle, J. (1979): "Intentionality and the use of language," en Avishai Margalit (ed.), *Meaning and Use: papers presented at the 2nd Jerusalem Philosophical Encounter, April 1976*, Dordrecht, D. Reidel, pp. 181-197.
Searle, J. (1980): "Minds, brains, and programs," *Behavioral and Brain Sciences* 3(3), penúltimo borrador. http://www.bbsonline.org/ documents/a/00/00/04/84/ bbs00000484-00/bbs.searle2.html.
Searle, J. (1984): *Minds, Brains and Science: The 1984 Reith Lectures*, Cambridge, Harvard University Press.
Searle, J. (1990). "Is the Brain's Mind a Computer Program?," *Scientific American* 262 (1), pp. 26–31.
Turing, A. (1950): "Computing Machinery and Intelligence," *Mind* 59 (236), pp. 433–60.
Wittgenstein, L. (1921): *Logisch-Philosophische Abhandlung,* en Ostwald, W. (ed.), *Annalen der Naturphilosophie* 14, Leipzig, Unesma, pp. 185-262. http://digital.ub.uni-leipzig.de/id15325484L

Ivar Hannikainen
Department of Philosophy
University of Sheffield
i.hannikainen@sheffield.ac.uk