Mentes, cerebros y programas[12]

John R. Searle 1980

¿Qué significado deberíamos atribuir a los recientes esfuerzos por simular computacionalmente las capacidades cognitivas humanas? Al responder esta pregunta consideraré útil distinguir entre lo que llamo IA "fuerte" e IA "débil" o "cauta". Según la IA débil, el valor fundamental del computador en el estudio de la mente radica en que nos brinda una herramienta muy poderosa. Por ejemplo, nos permite formular y poner a prueba hipótesis de manera más rigurosa y precisa que antes. Pero de acuerdo a la IA fuerte, el computador no es una mera herramienta en el estudio de la mente; más bien, un computador programado apropiadamente es realmente una mente, en el sentido que se puede decir que los computadores con los programas apropiados pueden literalmente *comprender* y tener otros estados cognitivos. Y de acuerdo a la IA fuerte, debido a que el computador programado tiene estados cognitivos, los programas no son meras herramientas que nos permiten poner a prueba explicaciones psicológicas; más bien, los programas son en sí mismos esas explicaciones. No tengo objeciones acerca de los postulados de la IA débil, al menos en lo que concierne a este artículo. Mi discusión estará dirigida a los postulados que he definido como IA fuerte, especialmente el que señala que un computador programado de manera apropiada literalmente tiene estados cognitivos, y que los programas, por consiguiente, explican la cognición humana. Cuando me refiera a IA, es la versión fuerte expresada en estos dos postulados la que tengo en mente.

Consideraré el trabajo de Roger Schank y sus colegas en Yale (ver, por ejemplo, Schank y Abelson 1977), debido a que estoy más familiarizado con éste que con otros postulados similares, y debido a que provee un ejemplo claro del tipo de trabajo que deseo examinar. Pero nada de lo que sigue depende de los detalles de los programas de Schank. Los mismos argumentos se aplicarían al SHRDLU de Winograd (1972), al ELIZA de Weizenbaum (1965), y de hecho a cualquier simulación de los fenómenos mentales humanos en una máquina de Turing.

Brevemente, y dejando de lado los detalles, podemos describir el programa de Schank de acuerdo a lo siguiente: el objetivo del programa es simular la habilidad humana para comprender historias. Es característico de las habilidades humanas para comprender historias que permiten responder preguntas acerca de éstas, aún cuando la información que entreguen no haya sido mencionada explícitamente en el texto. Entonces, por ejemplo, supongamos que a Ud. le entregan la siguiente historia: "Un hombre fue a un restaurante y pidió una hamburguesa. Cuando la hamburguesa llegó, estaba completamente quemada, y el hombre salió furiosamente del restaurante sin pagar la

[1] En Haugeland, J. (1997) *Mind Design II*. Cambridge, MA: MIT Press. Segunda edición 1997.

[2] Traducción de Lucía Castillo I., estudiante de Magíster en Estudios Cognitivos, Universidad de Chile.

hamburguesa o dejar propina". Ahora, si se le entrega la siguiente pregunta: "¿Se comió el hombre la hamburguesa?", usted probablemente responderá "No, no se la comió". De manera similar, si le entregan la siguiente historia: "Un hombre fue a un restaurante y pidió una hamburguesa; cuando la hamburguesa llegó, el hombre se sintió muy satisfecho con ella; y cuando se fue del restaurante le dejó al mozo una gran propina, antes de pagar la cuenta", y le preguntan "¿Se comió el hombre la hamburguesa?", probablemente responderá "Sí, se comió la hamburguesa".

Ahora, las máquinas de Schank pueden, similarmente, responder preguntas sobre restaurantes de esta manera. Para hacer eso, poseen una "representación" del tipo de información que los seres humanos tienen acerca de los restaurantes, que les permite responder preguntas como las hechas anteriormente, dada esa clase de historias. Cuando a la máquina se le entrega la historia y se le hacen las preguntas, imprimirá respuestas similares a las que esperaríamos que los seres humanos entregaran ante historias similares. Los partisanos de la IA fuerte proclaman que en esta secuencia de preguntas y respuestas la máquina no sólo está simulando una habilidad humana, sino también:

(a) Puede decirse que la máquina literalmente *comprende* la historia y responde las preguntas; y

(b) Lo que la máquina y su programa hacen *explica* la habilidad humana para comprender la historia y responder preguntas acerca de ésta.

Me parece que los postulados (a) y (b) no se sostienen en ninguna medida a partir del trabajo de Schank, e intentaré demostrarlo en los párrafos siguientes[i].

Una manera de poner a prueba cualquier teoría de la mente es preguntarse cómo sería si nuestra propia mente realmente funcionara bajo los principios que la teoría sostiene para todas las mentes. Apliquemos esta prueba al programa de Schank mediante el siguiente *Gedankenexperiment*. Supongamos que estoy encerrado en una pieza y se me entrega un gran lote de escritos en chino. Supongamos, además, como es realmente el caso, que no conozco nada de chino, ni escrito ni hablado, y que ni siquiera confío en que pudiera distinguir escritura china de escritura japonesa, o de dibujos sin sentido. Supongamos, en seguida, que luego de este primer lote de escritura china se me entrega un segundo lote de escritos en chino junto a un conjunto de reglas para correlacionar el segundo lote con el primer lote. Las reglas están en inglés, y comprendo estas reglas tan bien como cualquier hablante nativo de inglés. Me permiten correlacionar un conjunto de símbolos formales con otro conjunto de símbolos formales, y lo único que significa "formal" aquí es que puedo identificar estos símbolos enteramente por sus formas. Ahora supongamos que se me entrega también un tercer lote de símbolos chinos junto a algunas instrucciones, nuevamente en inglés, que me permiten correlacionar elementos de este tercer lote con los otros dos lotes, y estas reglas me indican cómo debo entregar de vuelta ciertos símbolos chinos de ciertas formas en respuesta a ciertas formas que se me entregaron en el tercer lote.

Sin ser de mi conocimiento, las personas que me entregan estos símbolos llaman al primer lote un “guión”, al segundo una “historia”, y al tercero “preguntas”. Aún más, llaman a los símbolos que yo entrego de vuelta en respuesta al tercer lote “respuestas a las preguntas”, y al conjunto de reglas en inglés que me entregaron, “programa”. Para complicar un poquito este asunto, imagine que estas personas también me entregan historias en inglés, las que entiendo, y luego me hacen preguntas en inglés acerca de estas historias, y yo les respondo en inglés. Supongamos también que luego de un rato me vuelvo tan bueno siguiendo las instrucciones para manipular estos símbolos chinos, y los programadores se vuelven tan buenos escribiendo los programas, que, desde un punto de vista externo –esto es, desde el punto de vista de alguien fuera de la pieza en la que estoy encerrado-, mis respuestas a las preguntas son indistinguibles de las que darían hablantes nativos de chino. Nadie que viese mis respuestas podría decir que no hablo una palabra de chino. Supongamos, además, que mis respuestas a las preguntas en inglés son, como lo serían sin duda, indistinguibles de las de otros hablantes nativos de inglés, por la simple razón de que soy un hablante nativo de inglés. Desde un punto de vista externo, desde el punto de vista de alguien leyendo mis “respuestas”, las respuestas a las preguntas en chino y a las preguntas en inglés son igualmente buenas. Pero en el caso chino, a diferencia del caso inglés, produzco las respuestas manipulando símbolos formales sin interpretar. En lo que concierne al chino, simplemente me comporto como un computador; realizo operaciones computacionales sobre elementos formalmente especificados. Para los propósitos del chino, soy simplemente la realización de un programa computacional.

Ahora, los postulados de la IA fuerte plantean que los computadores programados entienden las historias, y que el programa en algún sentido explica el entendimiento humano. Pero estamos ahora en la posición de examinar estos postulados a la luz de nuestro experimento mental.

(a) En cuanto al primer postulado, me parece obvio, en el ejemplo, que no entiendo una sola palabra de las historias chinas. Tengo inputs y outputs que son indistinguibles respecto de los de un hablante nativo de chino, y puedo tener el programa formal que se le antoje, pero aún así no entiendo nada. El computador de Schank, por las mismas razones, no entiende nada de ninguna historia, ya sea en chino, inglés o el idioma que sea, ya que en el caso chino el computador soy yo, y en los casos en que el computador no soy yo, el computador no tiene nada más que lo que yo tengo en el caso en que no entiendo nada.

(b) En cuanto al segundo postulado –que el programa explica el entendimiento humano- podemos ver que el computador y su programa no entregan condiciones suficientes para la comprensión, dado que el computador y el programa están funcionando y no hay comprensión alguna. Pero, ¿entrega siquiera alguna condición necesaria, o alguna contribución significativa al entendimiento? Uno de los postulados de los defensores de la IA fuerte es el siguiente: cuando yo comprendo una historia en inglés, lo que estoy haciendo es exactamente lo mismo –o quizás más de lo mismo- que lo que estaba haciendo

> en el caso de manipular símbolos chinos. Es simplemente más manipulación formal de símbolos lo que distingue el caso en inglés, donde entiendo, del caso en chino, donde no entiendo. No he demostrado que este postulado es falso, pero ciertamente parecería increíble en el ejemplo.

La plausibilidad del supuesto se ha derivado de la suposición de que podemos construir un programa que tenga los mismo inputs y outputs que los hablantes nativos, y, además, de que asumimos que los hablantes tienen algún nivel de descripción en el que también son implementaciones de un programa. Sobre la base de estas suposiciones, asumimos que aún si el programa de Schank no explica completamente la comprensión, al menos explica una parte del asunto. Esto es, supongo, una posibilidad empírica, pero no se ha dado la más mínima razón para suponer que sea cierta, dado que lo que el ejemplo sugiere –aunque de hecho no lo ha demostrado- es que el programa computacional es irrelevante en mi comprensión de la historia. En el caso chino poseo todo lo que la inteligencia artificial podría asignarme en términos de un programa, y no entiendo nada; en el caso inglés lo comprendo todo, y no hay hasta el momento ninguna razón para suponer que mi comprensión tiene algo que ver con programas computacionales –esto es, con operaciones computacionales en elementos especificados sólo formalmente.

En tanto el programa se define en términos de operaciones computacionales sobre elementos definidos exclusivamente de manera formal, lo que el ejemplo sugiere es que éstos, en sí mimos, no poseen ninguna conexión interesante con la comprensión. No constituyen, ciertamente, condiciones suficientes, y no se ha entregado la más mínima razón para suponer que constituyan condiciones necesarias, o siquiera que su contribución a la comprensión sea significativa. Es necesario considerar que la fuerza del argumento no radica simplemente en que máquinas diferentes puedan tener los mismos inputs y outputs operando bajo principios formales distintos –ése no es el punto en lo absoluto- sino en que cualquier principio formal que pongas en el computador será insuficiente para dar cuenta del entendimiento, dado que un ser humano será capaz de seguir esos principios formales sin entender nada, y no se ha entregado ninguna razón para suponer que sean necesarios o siquiera que contribuyan, ya que no hay razón para suponer que cuando comprendo inglés estoy operando bajo algún programa formal.

¿Qué es, entonces, lo que tengo en el caso de las oraciones en inglés, que no tengo en el caso de las oraciones en chino? La respuesta obvia es que sé lo que las primeras significan, pero no tengo idea alguna del significado de las segundas. ¿En qué consiste esto, y por qué no podríamos dárselo a una máquina, sea lo que sea? ¿Por qué no podría dársele a una máquina lo que sea que tengo que me hace comprender las oraciones en inglés? Volveré a estas preguntas luego de desarrollar mi ejemplo un poco más.

HE TENIDO LA OPORTUNIDAD de presentar este ejemplo a muchos investigadores en inteligencia artificial y, curiosamente, no parecen estar de acuerdo respecto de cuál sería la respuesta correcta. Obtengo una variedad sorprendente de respuestas, y en lo que sigue consideraré las más comunes (especificadas de acuerdo a sus orígenes geográficos). En primer lugar, quiero aclarar algunos malentendidos comunes acerca de

la “comprensión”. En muchas de estas discusiones uno se encuentra con maniobras muy elaboradas acerca de la palabra “comprensión”. Mis críticos señalan que hay distintos grados de comprensión, que “comprende” no es simplemente un predicado, que hay incluso distintos tipos y niveles de comprensión, y que frecuentemente la ley del medio excluido no se aplica directamente a afirmaciones de la forma “x comprende y”, que en muchos casos si *x comprende y* es un tema sobre el que hay que decidir y no un simple hecho. Y así sucesivamente.

Ante todos estos puntos quiero decir: “Por supuesto, por supuesto”. Pero no tienen nada que ver con lo que está en discusión. Hay casos claros en los que “comprende” se aplica y casos claros en los que no se aplica; y estos casos son todo lo que necesito para este argumento[ii]. Yo comprendo historias en inglés; en un grado menor puedo comprender historias en francés; en un grado aún menor, historias en alemán; y en chino, no comprendo nada. Mi auto y mi calculadora, por el contrario, no comprenden nada; no están en esta línea de trabajo.

Generalmente atribuimos “comprensión” y otros predicados cognitivos metafóricamente y por analogía a automóviles, sumadoras y otros artefactos; pero estas atribuciones no prueban nada. Decimos “La puerta *sabe* cuando abrir debido a sus celdas fotoeléctricas”, “La sumadora *sabe como (comprende como, es capaz de)* hacer sumas y restas pero no divisiones”, y “El termostato *percibe* cambios en la temperatura”. La razón por la que hacemos estas atribuciones es interesante y tiene que ver con el hecho que extendemos nuestra propia intencionalidad en los artefactos[iii]; nuestras herramientas son extensiones de nuestros propósitos, y por lo tanto encontramos natural atribuirles intencionalidad de manera metafórica. El sentido en que una puerta automática “comprende las instrucciones” de sus celdas fotoeléctricas no es en lo absoluto el sentido en que yo comprendo inglés.

Si el sentido en que los computadores programados de Schank comprenden historias se supone que sea el sentido metafórico en que la puerta comprende, y no el sentido en que yo entiendo inglés, no valdría la pena discutir el asunto. Newell y Simon escriben que el sentido de “comprensión” que atribuyen a sus computadores es exactamente el mismo que el de los seres humanos. Me gusta lo explícito de este postulado, y es este tipo de afirmaciones el que consideraré. Defenderé que, en sentido literal, el computador programado comprende lo mismo que el automóvil y la sumadora: absolutamente nada. La comprensión del computador no es sólo (como en el caso de mi comprensión del alemán) parcial o incompleta; es nula.

Ahora a las respuestas.

I LA RESPUESTA DEL SISTEMA (Berkeley): Mientras es cierto que una persona individual encerrada en la pieza no comprende la historia, el hecho es que ésta es sólo una parte de un sistema completo, y el sistema sí comprende la historia. La persona tiene frente a sí un gran libro donde están escritas las reglas, tiene un montón de papel para escribir y lápices para hacer los cálculos, tiene “bancos de datos” de conjuntos de

símbolos chinos. Ahora, la comprensión no se adscribe al individuo por sí sólo; sino al sistema completo del cual él es parte.

Mi respuesta a la teoría del sistema es simple. Dejemos que el individuo internalice todos los elementos del sistema. Él memoriza las reglas del libro y los bancos de datos de símbolos chinos, y hace todos los cálculos en su cabeza. El individuo entonces incorpora el sistema entero. No hay nada en el sistema que no pueda abarcar. Podemos incluso deshacernos de la pieza y suponer que trabaja en el exterior. Del mismo modo, no entiende nada de chino, y *a fortiori* tampoco el sistema, dado que no hay nada en el sistema que no esté en él. Si él no entiende nada, entonces no hay modo en que el sistema pudiese entender, porque el sistema no es más que una parte de él.

De hecho, me siento un poco avergonzado del sólo hecho de dar esta respuesta a la teoría del sistema, debido a que la teoría me parece tan inverosímil desde un principio. La idea es que mientras una persona no entiende chino, de algún modo la *conjunción* de esa persona y algunos pedazos de papel podrían entender chino. No me resulta fácil imaginar cómo alguien fuera de la ideología podría encontrar la idea verosímil en lo absoluto. Sin embargo, creo que mucha gente comprometida con la ideología de la IA fuerte podría, al fin y al cabo, sentirse inclinada a decir algo muy parecido a esto; por lo que analicémoslo un poco más. De acuerdo a una de las versiones de esta visión, mientras el hombre en el ejemplo de los sistemas internalizados no entiende chino en el sentido que lo comprende un hablante nativo de chino (porque no sabe, por ejemplo, que la historia se refiere a restaurantes y hamburguesas y ese tipo de cosas), “el hombre como sistema de manipulación formal de símbolos” *sí comprende chino*. El subsistema del hombre constituido por el sistema formal de manipulación de símbolos chinos no debiese ser confundido con el subsistema para inglés.

Así que realmente hay dos subsistemas en el hombre; uno que comprende inglés, y el otro chino, y “es sólo que los dos sistemas tienen muy poco que ver el uno con el otro”. Pero, quisiera responder, no sólo tienen poco que ver el uno con el otro, sino que no son ni remotamente similares. El subsistema que comprende inglés (asumiendo que nos permitimos hablar en esta jerga de “subsistemas” por un momento) sabe que las historias son acerca de restaurantes y comer hamburguesas y cosas así. Pero el sistema chino no sabe nada de esto; mientras el sistema inglés sabe que “hamburguesas” se refiere a hamburguesas, el sistema chino sólo sabe que luego de “garabato#$” viene “garabato@%”. Lo único que sabe es que varios símbolos formales se introducen en un lado y se manipulan de acuerdo a reglas escritas en inglés, y que luego otros símbolos se expresan en el otro lado.

El punto importante del ejemplo original era discutir que esa manipulación de símbolos en sí misma no puede ser suficiente para comprender chino en ningún sentido literal, porque el hombre podría escribir “garabato@%” luego de “garabato#$” sin entender nada de chino. Y no debilita el argumento el postular subsistemas dentro del hombre, ya que los subsistemas no mejoran en nada la condición original del hombre; todavía no tienen nada ni remotamente parecido a lo que tiene el hablante (o subsistema) de inglés.

De hecho, de acuerdo a la descripción del caso, el subsistema chino es simplemente una parte del subsistema inglés, una parte que se hace cargo de la manipulación de símbolos sin sentido de acuerdo a las reglas del inglés.

Preguntémonos qué se supone que motivó la respuesta del sistema en primer lugar –esto es, ¿qué bases *independientes* podrían asumirse para decir que el agente debe tener un subsistema dentro de sí que literalmente comprende historias en chino? Hasta donde puedo ver, la única base está en que en el ejemplo tengo el mismo input y output que un hablante nativo de chino, y un programa que va de uno a otro. Pero el punto del ejemplo está en demostrar que eso no podría ser suficiente para la comprensión, en el sentido en que yo comprendo historias en inglés, porque una persona, o el conjunto de sistemas que constituyen una persona, podría tener la combinación adecuada de input, output y programa, y aún así no entender nada en el sentido literal relevante en el que yo entiendo inglés.

La única motivación para decir que *debe* haber un subsistema dentro de mí que comprende chino es que tengo un programa y puedo pasar el test de Turing: puedo engañar a hablantes nativos de chino (ver Turing 1950). Pero precisamente uno de los puntos en discusión es la adecuación del test de Turing. El ejemplo muestra que podrían haber dos "sistemas", y que ambos podrían pasar el test de Turing, pero que de los dos hay uno sólo que comprende; y no constituye un argumento contra este punto decir que, ya que ambos pueden pasar el test de Turing, ambos deben comprender, dado que este supuesto no apunta al argumento de que el sistema en mí que entiende inglés tiene mucho más que el sistema que sólo procesa chino. En buenas cuentas, la respuesta del sistema da por sentado que el sistema debe comprender chino, al insistir en esto sin argumento alguno.

Además, la respuesta del sistema aparentemente conduce a consecuencias independientes absurdas. Si vamos a concluir que debo poseer cognición sobre la base de que tengo ciertos inputs y outputs y un programa entre unos y otros, parece ser que todo tipo de subsistemas no-cognitivos se volverían cognitivos. Por ejemplo, mi estómago tiene cierto nivel de descripción en el que procesa información, e implementa un número indeterminado de programas computacionales, pero doy por sentado que no pretendemos afirmar que tiene entendimiento. Sin embargo, si aceptamos la respuesta del sistema, parece difícil evitar decir que estómago, corazón, hígado, etc. son todos subsistemas que comprenden, dado que no hay un aspecto fundamental que nos permita distinguir la motivación para decir que el subsistema chino comprende, de la motivación para decir que el estómago comprende. (A propósito, decir que el sistema chino tiene información como input y output mientras el estómago tiene comida y productos de esa comida no es una respuesta apropiada, dado que, desde el punto de vista del agente, desde mi punto de vista, no hay información ni en el chino ni en la comida; el chino no es más que un montón de garabatos. La información en el caso chino radica únicamente en los ojos de los programadores e intérpretes, y no hay nada que evite que traten los inputs y outputs de mis órganos digestivos como información si así lo desean.)

Este último punto plantea ciertos problemas independientes para la IA fuerte, y es válido apartarse un momento del tópico central para explicarlos. Si se pretende que la IA fuerte sea una rama de la psicología, debe ser capaz de distinguir entre sistemas genuinamente mentales y aquellos que no lo son. Debe ser capaz de distinguir los principios que regulan el trabajo de la mente de aquellos que regulan los sistemas no-mentales; de otra manera no nos ofrecerá ninguna explicación acerca de qué es específicamente mental en lo mental. Y la distinción mental/no-mental no puede radicar simplemente en el ojo del observador –debe ser intrínseca a los sistemas. Porque de otra manera sería atributo de cualquier observador tratar a la gente como no-mental, y, por ejemplo, a los huracanes como mentales, si así lo quisiera.

Pero frecuentemente en la literatura sobre IA la distinción se pierde de maneras que en el largo plazo serán desastrosas para sostener la validez de la IA como investigación cognitiva. McCarthy, por ejemplo, sostiene: “Puede decirse que máquinas tan simples como un termostato tienen creencias, y tener creencias parece ser una característica de muchas máquinas capaces de resolver problemas” (1979). Cualquiera que crea que la IA tiene alguna oportunidad como teoría de la mente tiene que considerar las implicancias de esta afirmación. Se nos pide que aceptemos como un descubrimiento de la IA fuerte que el pedazo de metal en la pared que usamos para regular la temperatura tiene creencias en exactamente el mismo sentido que nosotros, nuestras esposas y nuestros niños, y más aún, que “la mayoría” de las demás máquinas en la pieza –teléfono, grabadora, calculadora, interruptor de luz, etc.- también tienen creencias en este sentido literal. No es el objetivo de este artículo argumentar contra el punto de McCarthy, por lo que simplemente afirmaré lo siguiente sin mayor argumento. El estudio de la mente parte del hecho que los humanos tienen creencias, y que los termostatos, teléfonos y calculadoras no las tienen. Si logras generar una teoría que refute este punto, habrás producido un contra-ejemplo de la teoría, y la teoría será falsa.

Uno se queda con la impresión de que la gente en IA que escribe este tipo de cosas piensa que puede salirse con la suya simplemente porque no las toman en serio, y creen que nadie más lo hará. Propongo, por un momento al menos, tomarlas en serio. Concentrémonos por un minuto en qué sería necesario para establecer que ese pedazo de metal ahí en la pared tiene creencias reales, creencias con dirección de ajuste, contenido proposicional, y condiciones de satisfacción; creencias que tienen la posibilidad de ser fuertes o débiles; creencias nerviosas, ansiosas o seguras; creencias dogmáticas, racionales o supersticiosas; fe ciegas o elucubraciones vacilantes; todo tipo de creencias. El termostato no es un candidato. Tampoco el estómago, hígado, calculadora o teléfono. Sin embargo, dado que estamos tomando la idea en serio, consideremos que su veracidad sería fatal para la propuesta de la IA fuerte como ciencia de la mente, ya que ahora la mente estaría en todas partes. Lo que queríamos saber era qué distingue a la mente de termostatos, hígados y demases. Y si McCarthy estuviese en lo cierto, la IA fuerte no tendría ninguna esperanza de decírnoslo.

II LA RESPUESTA DEL ROBOT (Yale): Supongamos que escribimos un programa distinto del de Schank. Supongamos que ponemos un computador dentro de un robot, y

que este computador no sólo acepta y entrega símbolos formales como input y output, sino que además opera el robot de manera tal que el robot hace algo muy parecido a percibir, caminar, moverse, clavar clavos, comer, beber –todo lo que Ud. quiera. El robot podría, por ejemplo, tener asociada una cámara de televisión que le permitiera ver, brazos y piernas que le permitieran actuar, y todo esto controlado por su cerebro de computador. Un robot así podría, a diferencia del computador de Schank, tener entendimiento genuino y otros estados mentales.

Lo primero que hay que considerar acerca de la respuesta del robot es que concede tácitamente que la cognición no es solamente manipulación formal de símbolos, dado que esta respuesta agrega un conjunto de relaciones causales con el mundo exterior.

Pero la réplica a la respuesta del robot es que agregar capacidades "perceptuales" y "motoras" no agrega nada al entendimiento, en particular, ni a la intencionalidad, en general, del programa original de Schank. Para evaluar esto, considere que el mismo experimento mental se aplica al caso del robot. Suponga que, en vez de un computador dentro del robot, me pone a mí dentro de la pieza, y me entrega, nuevamente, como en caso chino original, más símbolos chinos con más instrucciones en inglés para ordenar símbolos chinos con símbolos chinos y entregar símbolos chinos al exterior.

Ahora suponga que, sin que yo lo sepa, algunos de los símbolos chinos que se me entregan vienen de una cámara de televisión asociada al robot, y otros símbolos chinos que entrego sirven para que los motores dentro del robot muevan los brazos o las piernas. Es importante enfatizar que todo lo que estoy haciendo es manipular símbolos formales; no sé nada de estos otros hechos. Estoy recibiendo "información" de los aparatos "perceptuales" del robot, y estoy entregando "instrucciones" a sus aparatos motores, sin tener conocimiento de nada de esto. Soy el homúnculo del robot, pero a diferencia del homúnculo tradicional, no tengo idea de lo que está pasando. No entiendo nada salvo las reglas para la manipulación de símbolos. Ahora, en este caso, quiero afirmar que el robot no tiene estados intencionales en lo absoluto; simplemente se mueve como resultado de su cableado eléctrico y su programa. Y además, al implementar el programa, yo tampoco tengo estados intencionales del tipo relevante. Todo lo que hago es seguir instrucciones formales para manipular símbolos formales.

III LA RESPUESTA DEL SIMULADOR DE CEREBROS (Berkeley y MIT): Supongamos que diseñamos un programa que no representa la información que tenemos acerca del mundo, como la información de los guiones de Schank, sino que simula la secuencia real de activación de neuronas que se produce en las sinapsis del cerebro de un hablante nativo de chino cuando éste está comprendiendo historias en chino y dando respuestas acerca de las mismas. La máquina recibe historias en chino y preguntas acerca de esas historias como input, simula la estructura formal de un cerebro chino real al procesar esas historias, y entrega respuestas en chino como output. Podemos incluso imaginar que la máquina no opera con un solo programa serial, sino con todo un conjunto de programas corriendo en paralelo, de acuerdo al modo en que suponemos que operan los cerebros humanos reales al procesar un lenguaje natural. Ahora,

probablemente en este caso tendríamos que decir que la máquina comprende las historias; y si nos negamos a decir esto, ¿no tendríamos que negar también que los hablantes nativos de chino entiendan las historias? Al nivel de sinapsis, ¿cuál sería la diferencia entre el programa del computador y el programa del cerebro chino?

Antes de enfrentar esta respuesta, quisiera apartarme del tema un momento para considerar que esta es una respuesta muy extraña para un partisano de la inteligencia artificial (funcionalismo, entre otros). Yo creía que la idea central de la IA fuerte era que no necesitamos saber cómo funciona el cerebro para saber cómo funciona la mente. La hipótesis básica, o al menos eso creía yo, era que hay un nivel de operaciones mentales que consiste en procesos computacionales sobre elementos formales que constituye la esencia de lo mental, y que puede ser ejecutado en diferentes soportes computacionales. Sobre los supuestos de la IA fuerte, la mente es al cerebro como el programa es al *hardware*, y por tanto podemos entender la mente sin necesidad de hacer neurofisiología. Si tuviéramos que saber cómo funciona el cerebro para hacer IA, no nos preocuparíamos de la IA.

Sin embargo, aún acercarnos tanto a las operaciones del cerebro no es suficiente para producir comprensión. Para sostener esto, imagine que en vez de tener a un hombre monolingüe en una pieza manipulando símbolos tenemos a un hombre operando un complejo sistema de cañerías de agua con válvulas que las conectan entre sí. Cuando el hombre recibe los símbolos chinos mira en su programa, escrito en inglés, cuáles válvulas debe abrir y cuáles debe cerrar. Cada conexión de las cañerías de agua corresponde a una sinapsis en el cerebro chino, y el sistema en su conjunto está organizado para que, luego de las necesarias activaciones –esto es, luego de abrir y cerrar las válvulas necesarias- la respuesta en chino emerja como output al final de la serie de cañerías.

Ahora, ¿dónde está la comprensión en este sistema? Recibe chino como input, simula la estructura formal de las sinapsis del cerebro chino, y entrega chino como output. Pero ciertamente el hombre no entiende chino, y tampoco las cañerías. Y si nos sentimos tentados a adoptar la absurda teoría de que, de algún modo, la *conjunción* del hombre *y* las cañerías sí comprende, recordemos que, en principio, el hombre puede internalizar la estructura formal de las cañerías de agua y hacer todas las "activaciones" necesarias en su imaginación. El problema del simulador de cerebros es que lo que simula sobre el cerebro no es lo que habría que simular. Mientras simule sólo la estructura formal de la secuencia de activaciones neuronales en las sinapsis, no habrá simulado lo relevante acerca del cerebro: su habilidad para producir estados intencionales. Y el hecho que las propiedades formales no son suficientes para las propiedades causales queda demostrado en el ejemplo de las cañerías de agua. Podemos remover todas las propiedades formales de las propiedades causales neurobiológicas relevantes.

IV LA RESPUESTA COMBINATORIA (Berkeley y Stanford): Mientras cada una de las tres respuestas anteriores podría no ser suficientemente convincente en sí misma como refutación del contra-ejemplo de la pieza china, si las tomamos las tres

colectivamente resultan mucho más convincentes e incluso decisivas. Imagine un robot con un computador organizado como un cerebro insertado en su cavidad craneal; imagine este computador programado con todas las sinapsis de un cerebro humano; imagine que la totalidad del comportamiento del robot resulta imposible de distinguir del comportamiento humano real; y ahora imagine este conjunto como un sistema unificado, y no sólo como un computador con inputs y outputs. Seguramente en este caso sí adscribiríamos intencionalidad al sistema.

Concuerdo completamente con la afirmación de que, en este caso, sería racional e incluso irresistible aceptar la hipótesis de que el robot tiene intencionalidad, siempre y cuando no supiésemos nada más acerca de él. De hecho, fuera de la apariencia y el comportamiento, el resto de los elementos de la combinación son irrelevantes. Si pudiésemos construir un robot cuyo comportamiento resultase indistinguible de un conjunto importante de comportamientos humanos, le atribuiríamos intencionalidad, hasta que no apareciese alguna razón para no hacerlo. No necesitaríamos saber por adelantado que su cerebro era un análogo formal del cerebro humano.

Pero, realmente, no veo cómo esta propuesta pueda ser de alguna ayuda a los supuestos de la IA fuerte, y aquí está el porqué. De acuerdo a la IA fuerte, implementar un programa formal con los inputs y outputs correctos es condición suficiente, y de hecho constitutiva, de intencionalidad. Según Newell (1980), la esencia de lo mental es la operación de un sistema de símbolos físicos. Pero nuestra atribución de intencionalidad al robot en este ejemplo no tiene nada que ver con programas formales. Simplemente se basa en la suposición de que si el robot se ve y se comporta suficientemente como nosotros, supondremos, hasta que se nos pruebe lo contrario, que debe tener estados mentales como los nuestros, que causen y sean expresados por su comportamiento, y que debe tener un mecanismo interno capaz de producir dichos estados mentales. Si supiéramos, independientemente, cómo dar cuenta de su comportamiento sin necesidad de esas suposiciones, no le atribuiríamos intencionalidad, especialmente si supiéramos que tiene un programa formal. Y este es precisamente el punto de mi argumento anterior para la respuesta del robot.

Supongamos que sabíamos que el comportamiento del robot provenía completamente del hecho que tenía un hombre en su interior que recibía símbolos formales sin interpretar desde los receptores sensoriales del robot y que remitía otros símbolos formales sin interpretar al exterior a través de sus mecanismos motores, y que el hombre estaba manipulando estos símbolos de acuerdo a un montón de reglas. Además, supongamos que el hombre no sabe nada acerca de la realidad del robot; todo lo que sabe es qué operaciones realizar sobre cuáles símbolos sin significado. En un caso así consideraríamos que el robot no es más que un ingenioso muñeco mecánico. La hipótesis de que el muñeco tiene una mente parecería ahora injustificada e innecesaria, dado que ya no hay razón para adscribir intencionalidad al robot o al sistema del que el robot es parte (a excepción, por supuesto, de la intencionalidad del hombre al manipular los símbolos). Las manipulaciones formales de símbolos continúan, los inputs y outputs se articulan correctamente, sin embargo el único *locus* real de intencionalidad es el

hombre, y éste no sabe nada acerca de los estados intencionales relevantes; él no *ve*, por ejemplo, lo que aparece ante los ojos del robot, no *intenta* mover el brazo del robot, y no *entiende* nada de lo que el robot dice o escucha. Ni tampoco, por las razones expuestas anteriormente, lo hace el sistema del que el hombre y el robot forman parte.

Para comprender este punto, contrastemos este caso con otros casos en los que encontramos completamente natural adscribir intencionalidad a miembros de ciertas especies de primates, tales como simios y monos, y a ciertos animales domésticos, como los perros. Las razones por las que lo encontramos natural son, a grandes rasgos, dos: no podemos encontrarle sentido al comportamiento de los animales sin adscribirles intencionalidad, y podemos ver que las bestias están hechas de materia similar a la nuestra –un ojo, una nariz, piel, etc. Dada la coherencia del comportamiento del animal y la suposición de la misma materia causal detrás de éste, asumimos tanto que el animal debe tener estados mentales tras su comportamiento, y que estos estados mentales deben ser producidos por mecanismos hechos de la materia que se parece a nuestra materia. Seguramente haríamos suposiciones similares acerca del robot a menos que tuviéramos alguna razón para no hacerlo; pero apenas supiésemos que su comportamiento era el resultado de un programa formal, y que las propiedades causales reales de la sustancia física eran irrelevantes, abandonaríamos la suposición de intencionalidad.

Hay otras dos respuestas a mi ejemplo que aparecen frecuentemente (y que por lo tanto son dignas de discutirse), pero que se alejan totalmente de la idea en cuestión.

V LA RESPUESTA DE LAS OTRAS MENTES (Yale): ¿Cómo sabemos que otras personas entienden chino o cualquier otra cosa? Sólo por su comportamiento. Ahora, el computador puede pasar los test de comportamiento tan bien como ellos (en principio), así que si va a atribuirles cognición a otras personas, debe, en principio, atribuírsela también a los computadores.

La objeción merece sólo una breve respuesta. El problema en esta discusión no es cómo sé que otras personas tienen estados cognitivos, sino qué es lo que les estoy atribuyendo cuando les atribuyo estados cognitivos. El *quid* del argumento es que éstos no pueden ser simplemente procesos computacionales y sus outputs, porque pueden haber procesos computacionales y outputs sin estados cognitivos. No es una respuesta válida ante este argumento el fingir anestesia. En “ciencias cognitivas” uno presupone la realidad y cognoscibilidad de lo mental de la misma manera que en las ciencias físicas uno debe presuponer la realidad y cognoscibilidad de los objetos físicos.

VI LA RESPUESTA DE LAS MÚLTIPLES MANSIONES (Berkeley): Su argumento en su totalidad presupone que la IA se trata sólo de computadores análogos y digitales. Pero resulta que ése es sólo el estado actual de la tecnología. Independiente de cuáles sean estos procesos causales que Ud. dice que son esenciales para la intencionalidad (asumiendo que Ud. está en lo cierto), eventualmente seremos capaces de construir dispositivos que tengan estos procesos causales, y eso será inteligencia artificial. Por lo tanto, sus argumentos no están de ninguna manera dirigidos a la habilidad de la inteligencia artificial para producir y explicar la cognición.

No tengo ninguna objeción contra esta respuesta a excepción de decir que, en efecto, trivializa el proyecto de la IA fuerte, al redefinirla como cualquier cosa que produce y explica artificialmente la cognición. El interés de la afirmación original respecto a la inteligencia artificial era que ésta era una tesis precisa y bien definida: los procesos mentales son procesos computacionales sobre elementos formalmente definidos. He estado ocupado en cuestionar esa tesis. Si la afirmación es redefinida y ésa ya no es la tesis, mis objeciones ya no se aplican, porque no habría una hipótesis susceptible de ser puesta a prueba sobre la que se pudieran aplicar.

VOLVAMOS a las preguntas que prometí que trataría de responder. Habiendo sido establecido en mi experimento original que entiendo el inglés y no entiendo el chino, y habiéndose establecido de acuerdo a esto que la máquina no entiende ni inglés ni chino, aún así debe haber algo en mí que causa que entienda inglés, y algo correspondiente que falta, causando que no entienda chino. Ahora bien, ¿por qué no podríamos dotar a la máquina de ese algo, sea lo que sea? No veo ninguna razón, en principio, de porqué no podríamos dotar a una máquina de la capacidad de comprender inglés o chino, dado que, en un sentido importante, nuestros cuerpos con nuestros cerebros son precisamente esas máquinas. Pero si veo argumentos sólidos para decir que no podemos darle esa capacidad a una máquina donde la operación de la máquina esté definida solamente en términos de procesos computacionales sobre elementos definidos formalmente –esto es, donde la operación de la máquina esté definida como la realización de un programa computacional. No es porque soy la realización de un programa computacional que comprendo inglés y tengo otras formas de intencionalidad. (Soy, supongo, la realización de un número indeterminado de programas computacionales). Más bien, hasta donde sabemos, es porque soy cierto tipo de organismo con cierta estructura biológica (esto es, física y química), y esta estructura bajo ciertas condiciones es capaz causalmente de producir percepción, acción, comprensión, aprendizaje, y otros fenómenos intencionales. Y parte del presente argumento es que sólo algo que tenga esos poderes causales podrá tener tal intencionalidad. Quizás otros procesos físicos y químicos podrían producir exactamente los mismos efectos; quizás, por ejemplo, los marcianos también tengan intencionalidad, pero sus cerebros estén hechos de algo completamente distinto. Esa es una pregunta empírica, más parecida a la pregunta de si la fotosíntesis puede ser producida por algo con una constitución química diferente a la de la clorofila.

Pero el punto central del presente argumento es que ningún modelo puramente formal va a ser nunca, por sí mismo, suficiente para dar cuenta de la intencionalidad, porque las propiedades formales no son en sí mismas constitutivas de intencionalidad, y no tienen en sí mismas ningún poder, excepto el poder, cuando son realizadas, de producir el siguiente estado del formalismo cuando la máquina está andando. Y cualquier otra propiedad causal que posea una realización particular del modelo formal es irrelevante para el modelo formal, debido a que podemos ubicar el mismo modelo formal en una realización diferente, donde esas propiedades causales estén obviamente ausentes. Aún si por algún milagro los hablantes de chino realizaran exactamente el programa de

Schank, podemos poner el mismo programa en hablantes de inglés, cañerías de agua, o computadores, ninguno de los cuales entiende chino, a pesar del programa.

Lo que interesa de las operaciones del cerebro no es la sombra formal que brinda la secuencia de sinapsis, sino las propiedades reales de esas secuencias. Todos los argumentos de la versión fuerte de la IA que he visto insisten en trazar una línea alrededor de las sombras que brinda la cognición, y luego afirmar que esas sombras son la realidad.

A MODO DE CONCLUSIÓN quiero definir algunos de los elementos filosóficos generales implícitos en el argumento. Para asegurar la claridad trataré de hacerlo en un estilo de preguntas y respuestas, y comenzaré con esa vieja y repetida pregunta:

- ¿Puede una máquina pensar?

La respuesta es, obviamente: Sí. Somos precisamente tales máquinas.

- Sí, pero, ¿puede un artefacto, una máquina construida por el hombre, pensar?

Asumiendo que es posible producir artificialmente una máquina con sistema nervioso, neuronas con axones y dendritas, y todo lo demás, suficientemente parecida a nosotros, nuevamente la respuesta parece ser obviamente: Sí. Si se pueden duplicar exactamente las causas, se pueden duplicar los efectos. Y de hecho podría ser posible producir consciencia, intencionalidad, y todo lo demás, usando principios químicos diferentes de los usados por los seres humanos. Es, como dije, una pregunta empírica.

- De acuerdo, pero ¿puede un computador digital pensar?

Si por “computador digital” nos referimos a cualquier cosa que tiene un nivel de descripción en el que puede ser correctamente descrito como la implementación de un programa computacional, entonces, ya que nosotros somos implementaciones de un número desconocido de programas computacionales y podemos pensar, nuevamente la respuesta es, por supuesto: Sí.

- Pero, ¿puede algo pensar, entender, etc. solamente en virtud de ser un computador con el tipo de programa adecuado? ¿Puede ser la implementación de un programa, el programa adecuado por supuesto, ser en sí misma condición suficiente para el entendimiento?

Creo que ésta es la pregunta correcta, aunque se la confunde comúnmente con alguna de las anteriores, y la respuesta a esta pregunta es: No.

- ¿Por qué no?

Porque las manipulaciones formales de símbolos en sí mismas no tienen intencionalidad. No tienen significado –no son siquiera manipulación de *símbolos*, dado que los “símbolos” no simbolizan nada. En jerga lingüística, tienen sólo sintaxis pero no semántica. La intencionalidad que parecen tener los computadores está únicamente en la

mente de los que los programan y utilizan, los que envían el input y los que interpretan el output.

El objetivo del ejemplo de la pieza china era tratar de demostrar esto, al mostrar que, en cuanto ponemos algo que realmente tiene intencionalidad, un hombre, dentro del sistema, y programamos al hombre con el programa formal, se hace evidente que el programa formal no brinda intencionalidad adicional de ninguna clase. No agrega nada, por ejemplo, a la habilidad de un hombre de comprender chino.

Precisamente el rasgo de IA que parecía más atractivo –la distinción entre el programa y su realización- resulta ser fatal para la afirmación de que la simulación podría constituir duplicación. La distinción entre el programa y su realización en el hardware parece ser paralela a la distinción entre el nivel de operaciones mentales y el nivel de operaciones cerebrales. Y si pudiésemos describir el nivel de las operaciones mentales como un programa formal, parecería que podríamos describir lo esencial de la mente sin necesidad de recurrir a la psicología introspectiva o a la neurofisiología del cerebro. Pero la ecuación "Mente es a cerebro como programa es a hardware" se quiebra en numerosos puntos, entre ellos los tres siguientes.

En primer lugar, la distinción entre programa y realización tiene la consecuencia de que el mismo programa podría tener toda clase de locas realizaciones sin ninguna forma de intencionalidad. Weizenbaum (1976), por ejemplo, muestra en detalle cómo construir un computador usando un rollo de papel higiénico y un montón de piedras pequeñas. De manera similar, el programa de comprensión de historias en chino puede ser programado en una secuencia de cañerías de agua, un conjunto de máquinas de viento, o un hablante monolingüe de inglés –ninguno de los cuales adquiere en virtud de éste algún conocimiento de chino. Piedras, papel higiénico, viento y cañerías de agua no son el tipo de cosas que podrían tener intencionalidad, en primer lugar (sólo algo que tenga los mismos poderes causales que el cerebro puede tener intencionalidad), y, aunque el hablante de inglés tiene el tipo de constitución apropiada para la intencionalidad, podemos ver fácilmente que no obtiene intencionalidad extra al memorizar el programa, debido a que memorizarlo no le va a enseñar chino.

En segundo lugar, el programa es meramente formal, pero los estados intencionales no son formales en ese sentido. Son definidos según su contenido, no según su forma. La creencia de que está lloviendo, por ejemplo, se define no según cierto aspecto formal, sino según su contenido mental, con condiciones de satisfacción, dirección de ajuste, entre otros (ver Searle 1979). De hecho, una creencia como ésa no tiene siquiera un aspecto formal en este sentido sintáctico, dado que la misma creencia puede ser articulada en un número indefinido de expresiones sintácticas diferentes en distintos sistemas lingüísticos.

En tercer lugar, como mencioné anteriormente, los eventos y estados mentales son productos de la operación del cerebro, pero el programa no es, en este sentido, producto de la operación del computador.

- Entonces, si los programas no son, de forma alguna, constitutivos de los procesos mentales, ¿por qué tanta gente cree lo contrario? Esto requiere al menos alguna explicación.

No sé cómo responder a esto. La idea de que las simulaciones computacionales pudiesen ser la realidad de los procesos mentales debió haber parecido sospechosa desde un principio, porque el computador no está, de ninguna manera, limitado a la simulación de procesos mentales. Nadie supone que la simulación computacional de un gran incendio quemará el vecindario hasta las cenizas, o que la simulación de una tormenta nos va a dejar a todos empapados. ¿Entonces por qué podría alguien suponer que una simulación computacional del entendimiento realmente entiende algo? Se dice en ocasiones que sería extremadamente difícil lograr que los computadores sintiesen dolor o se enamoraran, pero el amor y el dolor no son ni más fáciles ni más difíciles que la cognición o cualquier otra cosa. Para simular, todo lo que se necesita son los inputs y outputs correctos y un programa en el medio que transforme los primeros en los segundos. Eso es todo lo que el computador posee para hacer lo que hace. Confundir simulación con duplicación, ya se trate de dolor, amor, cognición, incendios o tormentas, constituye siempre el mismo error.

Aún así, hay numerosas razones por las que la IA debe haber parecido, y quizás aún parece a los ojos de numerosas personas, capaz de reproducir y por tanto explicar los fenómenos mentales. Y creo que no tendremos éxito en remover estas ilusiones mientras no hayamos expuesto completamente las razones que las originan.

En primer lugar, y quizás en primer orden de importancia, está la confusión acerca de la noción de “procesamiento de información”. Mucha gente en la ciencia cognitiva cree que el cerebro humano y su mente hacen algo llamado “procesamiento de información”, y, de manera análoga, el computador con su programa hacen procesamiento de información; pero los incendios y las tormentas, por el contrario, no hacen procesamiento de información en lo absoluto. Entonces, aunque el computador puede simular las características formales de cualquier proceso, mantiene una relación especial con la mente y el cerebro porque, cuando el computador está programado apropiadamente, idealmente con el mismo programa que el cerebro, el procesamiento de información es idéntico en ambos casos, y este procesamiento de información constituye realmente la esencia de lo mental.

Pero el problema con este argumento radica en que descansa sobre una ambigüedad en la noción de “información”. En el sentido en que la gente “procesa información” cuando reflexiona acerca de un problema aritmético o cuando lee y responde preguntas acerca de historias, el computador programado no hace “procesamiento de información”. Lo que hace es manipular símbolos formales. El hecho de que el programador y quien interpreta el output del computador usan esos símbolos para referirse a objetos en el mundo está totalmente fuera del alcance del computador. El computador, repito, tiene sintaxis pero no semántica. Entonces si usted tipea en el computador “2 más 2 igual ?”, el computador responderá “4”. Pero no tiene idea de que “4” significa 4, o de que

significa algo en lo absoluto. Y el punto no es que carezca de algún tipo de información de segundo orden acerca de la interpretación de los símbolos de primer orden, sino que sus símbolos de primer orden no tienen ningún tipo de interpretación en lo que concierne al computador. Todo lo que el computador posee son más símbolos.

La introducción de la noción de "procesamiento de información" produce, en consecuencia, un dilema. O construimos la noción de "procesamiento de información" de manera tal que implique intencionalidad como parte del proceso, o no lo hacemos. Si optamos por lo primero, el computador programado no hace procesamiento de información, sólo manipula símbolos formales. Si optamos por lo segundo, entonces, aunque el computador procesa información, sólo lo hace en el sentido en que las calculadoras, máquinas de escribir, estómagos, termostatos, tormentas y huracanes procesan información –esto es, en el sentido de que existe un nivel de descripción en el que podemos describirlos recibiendo información por un lado, transformándola, y produciendo información como respuesta. Pero en este caso depende de los observadores externos interpretar el input y el output como información en el sentido tradicional. Y no se establece ninguna similitud entre el computador y el cerebro en términos de alguna similitud en el procesamiento de la información en ninguno de los dos casos.

En segundo lugar, en buena parte de la IA permanece cierto conductismo u operacionalismo residual. Dado que los computadores programados apropiadamente pueden tener patrones de input/output similares a los de los seres humanos, nos sentimos tentados a postular estados mentales similares a los de los seres humanos en los computadores. Pero una vez que vemos que es posible conceptual y empíricamente que un sistema posea capacidades humanas en algún aspecto sin poseer intencionalidad alguna, debiésemos ser capaces de resistir este impulso. Mi calculadora de escritorio tiene la capacidad de calcular, pero no tiene intencionalidad; y en este artículo he tratado de demostrar que un sistema podría tener capacidades de input y output que duplicaran las de un hablante nativo de chino y aun así no comprender chino, independientemente de cómo haya sido programado. El Test de Turing es característico de la tradición de ser desvergonzadamente conductistas y operacionalistas, y creo que si los investigadores de la IA realmente repudiaran el conductismo y el operacionalismo, gran parte de la confusión entre simulación y duplicación sería eliminada.

En tercer lugar, este operacionalismo residual se une a una forma residual de dualismo; de hecho, la IA fuerte sólo tiene sentido dada la presunción dualista de que, en lo que respecta a la mente, el cerebro no tiene importancia. En la IA fuerte (y en el funcionalismo también) lo que importa son los programas, y los programas son independientes de su realización en máquinas; de hecho, hasta donde concierne a la IA, el mismo programa podría ser realizado en una máquina electrónica, una sustancia mental Cartesiana, o un espíritu del mundo Hegeliano. El descubrimiento individual más sorprendente que he hecho en la discusión de estos problemas es que muchos de los investigadores de la IA se sienten consternados ante mi idea de que los fenómenos mentales humanos reales podrían depender de las propiedades físico-químicas reales de

los cerebros humanos reales. Sin embargo, no debería haberme sorprendido; ya que, a menos que se acepte alguna forma de dualismo, el proyecto de la IA fuerte no tiene ninguna oportunidad.

El proyecto es reproducir y explicar lo mental diseñando programas; pero a menos que la mente sea conceptual y empíricamente independiente del cerebro, no se puede llevar a cabo el proyecto, ya que el programa es completamente independiente de cualquier realización. A menos que se crea que la mente es separable del cerebro tanto conceptual como empíricamente –una forma fuerte de dualismo- no se puede esperar reproducir lo mental escribiendo y ejecutando programas, ya que los programas deben ser independientes de los cerebros o cualquier otra forma de implementación. Si las operaciones mentales consisten en operaciones computacionales sobre símbolos formales, no han de tener ninguna conexión interesante con el cerebro, y la única conexión habría de ser que sucede que el cerebro es uno de los infinitos tipos de máquinas capaces de implementar ese programa. Esta forma de dualismo no es la variedad Cartesiana tradicional que afirma que existen dos tipos de substancias, pero sí es Cartesiano en el sentido que insiste en que lo específicamente mental acerca de la mente no tiene conexión intrínseca alguna con las propiedades reales del cerebro. Este dualismo subliminal se nos presenta camuflado por el hecho que la literatura acerca de la IA contiene frecuentes ataques contra el "dualismo". Lo que los autores parecen desconocer es que su posición presupone una versión fuerte del dualismo.

- ¿Puede pensar una máquina?

Mi punto de vista es que sólo *una* máquina puede pensar, y de hecho una clase muy especial de máquinas, llamadas cerebros, y las máquinas que tengan los *mismos poderes causales* que los cerebros. Y ésa es la razón principal de por qué la IA fuerte tiene tan poco que decirnos acerca del pensamiento: no tiene nada que decirnos acerca de las máquinas. Por su propia definición se refiere a programas, y los programas no son máquinas. Lo que sea que constituya además la intencionalidad, ha de ser un fenómeno biológico, y es muy probable que sea tan causalmente dependiente de la bioquímica específica de sus orígenes como la lactancia, la fotosíntesis, o cualquier fenómeno biológico. Nadie supondría que podemos producir leche y azúcar al ejecutar una simulación computacional de las secuencias formales de la lactancia y la fotosíntesis[iv]; pero cuando se trata de la mente, mucha gente está dispuesta a creer en tal milagro, debido a un profundo y determinante dualismo: la mente, suponen, depende de procesos formales, y es independiente de las causas materiales específicas, a diferencia de la leche y el azúcar.

En defensa de este dualismo, se expresa frecuentemente la esperanza de que el cerebro sea un computador digital. (A propósito, los primeros computadores eran llamados frecuentemente "cerebros electrónicos".) Pero eso no ayuda en nada. Por supuesto que el cerebro es un computador digital. Dado que todo es un computador digital, los cerebros también lo son. El punto es que la capacidad causal del cerebro de producir intencionalidad no puede consistir en su implementación de un programa

computacional, ya que para cualquier programa es posible que algo lo instancie y aún así no tenga estados mentales. Lo que sea que el cerebro hace para producir intencionalidad, no puede consistir en implementar un programa, ya que ningún programa es en sí mismo suficiente para dar cuenta de la intencionalidad.

Notas

[i] No estoy diciendo, por supuesto, que el propio Schank esté comprometido con estos supuestos.

[ii] Además, la "comprensión" implica tanto la posesión de estados mentales (intencionales) como la veracidad (validez, éxito) de estos estados. Para los propósitos de esta discusión, nos interesa sólo la posesión de los estados.

[iii] La intencionalidad es, por definición, la característica de ciertos estados mentales por la que están dirigidos a o son acerca de objetos y estados de las cosas en el mundo. Por lo tanto, creencias, deseos e intenciones son estados intencionales; formas indirectas de ansiedad y depresión no lo son. (Para mayor discusión, ver Searle 1979)