John Searle

Mentes, cerebros y ciencia

Traducción de Luis Valdés

CATEDRA
TEOREMA

2

¿PUEDEN LOS COMPUTADORES PENSAR?

En el capítulo anterior he proporcionado, al menos, las líneas generales de una solución al llamado 'problema mente-cuerpo'. Aunque no sabemos con detalle cómo funciona el cerebro, sabemos lo suficiente para tener una idea de las relaciones generales entre los procesos cerebrales y los procesos mentales. Los procesos mentales están causados por la conducta de elementos del cerebro. Al mismo tiempo, se realizan en la estructura que está compuesta por esos elementos. Pienso que esta respuesta es coherente con los enfoques biológicos estándar de los fenómenos biológicos. De hecho, es un género de respuesta de sentido común a la cuestión, dado lo que sabemos acerca de cómo el mundo funciona. Sin embargo, es, con mucho, el punto de vista de una minoría. El punto de vista prevalece en filosofía, psicología e inteligencia artifical es aquél que subraya las analogías entre el funcionamiento del cerebro y el funcionamiento de los computadores digitales. De acuerdo con la versión más extrema de este punto de vista, el cerebro es solamente un computador digital y la mente es solamente un programa de computador. Podría resumirse este punto de vista; yo lo llamo 'inteligencia artificial fuerte', o 'IA fuerte', diciendo que la mente es al cerebro lo que el programa es al *hardware* del computador.

Este punto de vista tiene la consecuencia de que no hay nada esencialmente biológico por lo que respecta a la mente humana. Sucede solamente que el cerebro es uno de un número indefinidamente extenso de diferentes géneros de *hardware* de computador que podrían servir de sostén a los programas que constituyen la inteligencia humana. Según este punto de vista, cualquier sistema físico que tuviese el programa correcto con los *inputs* y los *outputs* correctos tendría una mente en exactamente el mismo sentido que tú y yo tenemos mentes. Así, por ejemplo, si se hiciese un computador con viejas latas de cerveza y se le suministrase energía por medio de molinillos de viento, si tuviera el programa correcto, tendría que tener una mente. Y el punto no es que, dado todo lo que sabemos, podría tener pensamientos y sensaciones, sino más bien que tiene que tener pensamientos y sensaciones, puesto que todo aquello en lo que consiste tener pensamientos y sensaciones es esto: llevar a cabo el programa correcto.

La mayor parte de la gente que mantiene este punto de vista piensa que todavía no hemos diseñado programas que sean mentes. Pero hay bastante acuerdo general entre ellos de que esto es solamente un asunto de tiempo hasta que los científicos computacionales y las personas que trabajan en inteligencia artifical diseñen el *hardware* y los programas apropiados que sean los equivalentes de los cerebros y mentes humanas. Estos serán cerebros y mentes artificiales que son en todos los sentidos los equivalentes de los cerebros y las mentes humanos.

Mucha gente que está fuera del campo de la inteligencia artificial queda completamente pasmada al descubrir que alguien pueda creer un punto de vista como éste. Así pues, antes de criticarlo, permítaseme dar un puñado de ejemplos de las cosas que la gente que trabaja en este campo ha dicho efectivamente. Herbert Simon, de la Universidad de Carnegie-Mellon, dice que ya tenemos máquinas que literalmente pueden pensar. Ya no es cuestión de esperar por ninguna máquina futura, puesto que existen ya computadores digitales que

tienen pensamientos exactamente en el mismo sentido que usted y yo los tenemos. Bien, ¡qué casualidad! Los filósofos han estado preocupados durante siglos por la cuestión de si una máquina podría o no pensar, y ahora descubrimos que en Carnegie-Mellon ya tienen esas máquinas. El colega de Simon, Alan Newell, afirma que hemos descubierto ahora (y obsérvese que Newell dice 'descubierto' y no 'hemos avanzado la hipótesis' o 'hemos descubierto la posibilidad', sino hemos *descubierto*) que la inteligencia es solamente un asunto de manipulación de símbolos físicos; no tiene ninguna conexión esencial con ningún género específico de *wetware* o *hardware* biológico o físico. Más bien, cualquier sistema que sea capaz de manipular símbolos físicos de una manera correcta es capaz de inteligencia en el mismo sentido literal que la inteligencia humana de los seres humanos. Tanto Simon como Newell subrayan que no hay nada metafórico en esas afirmaciones; las proponen de una manera completamente literal. Se cita a Freeman Dyson como el que dijo que los computadores tienen una ventaja sobre el resto de nosotros por lo que respecta a la evolución. Puesto que la conciencia es un asunto de procesos formales solamente, en los computadores esos procesos formales pueden tener lugar en substancias que son mucho más capaces de sobrevivir en un universo que está enfriando, que en seres como nosotros, hechos de nuestros húmedos y sucios materiales. Marvin Minsky, del MIT, dice que la próxima generación de computadores será tan inteligente que deberíamos 'estar contentos si estuvieran dispuestos a mantenernos en torno a la casa como animalitos domésticos'. Mi siempre favorito en la literatura de afirmaciones exageradas a favor de los computadores es John McCarthy, el inventor del término 'inteligencia artificial'. MacCarthy dice que incluso 'puede decirse que máquinas tan simples como los termostatos tienen creencias'. Y de hecho, de acuerdo con él, de toda máquina capaz de resolver problemas puede decirse que tiene creencias. Admiro el coraje de MacCarthy. Una vez le pregunté. '¿Qué creencias tiene su termostato?' Y él me dijo: 'Mi termostato

tiene tres creencias: hace demasiado calor aquí, hace demasiado frío aquí, y aquí hace la temperatura correcta'. Como filósofo me gustan esas tres afirmaciones por una simple razón. A diferencia de muchas tesis filosóficas, son razonablemente claras, y admiten una refutación simple y decisiva. Es esta refutación la que voy a emprender en el presente capítulo.

La naturaleza de la refutación no tiene nada que ver con ninguna etapa particular de la tecnología de los computadores. Es importante subrayar este punto, puesto que la tentación es siempre pensar que la solución a nuestros problemas tiene que esperar a alguna, hasta ahora no creada, maravilla tecnológica. Pero de hecho, la naturaleza de la refutación es completamente independiente de cualquier estado en que se encuentre la tecnología. No tiene nada que ver con la definición misma de computador digital, con lo que un computador digital es.

Es esencial para nuestra concepción de computador digital que sus operaciones puedan especificarse de manera completamente formal; esto es, nosotros especificamos los pasos de la operación del computador en términos de símbolos abstractos —secuencias de ceros y unos impresos en una cinta, por ejemplo. Una 'regla típica de computador determinará que cuando una máquina está en un cierto estado y tiene un cierto símbolo en su cinta, entonces realizará ciertas operaciones tales como borrar el símbolo o escribir otro símbolo y a continuación entrar en otro estado tal como mover la cinta un cuadrado a la izquierda. Pero los símbolos no tienen ningún significado, no tienen ningún contenido semántico, no se refieren a nada. Tienen que especificarse en términos puramente de su estructura formal o semántica. Los ceros y los unos, por ejemplo, son solamente numerales, no están ni siquiera por números. Es más, es esta característica de los computadores digitales la que los hace tan potentes. Uno y el mismo tipo de *hardware,* si se diseña apropiadamente, puede usarse para pasar un rango indefinido de programas diferentes. Y uno, y el

mismo programa puede ser pasado en un rango indefinido de diferentes tipos de *hardware*.

Pero este rasgo de los programas, el que estén definidos de manera puramente formal o sintáctica, es fatal para el punto de vista de que los procesos mentales y los procesos de programas son idénticos. Y la razón puede enunciarse de manera completamente simple. Tener una mente es algo más que tener procesos formales o sintácticos. Nuestros estados mentales internos tienen, por definición, ciertos tipos de contenido. Si estoy pensando en Kansas City, o deseando tener una cerveza fría para beber, o preguntándome si habrá una caída en los tipos de interés, en cada caso mi estado mental tiene un cierto contenido mental además de cualesquiera otros rasgos formales que pueda tener. Esto es, incluso si mis pensamientos se me presentan en cadenas de símbolos tiene que haber más que las cadenas abstractas, puesto que las cadenas por sí mismas no pueden tener significado alguno. Si mis pensamientos han de ser *sobre* algo, entonces las cadenas tienen que tener un *significado* que hace que sean los pensamientos sobre esas cosas. En una palabra, la mente tiene más que una sintaxis, tiene una semántica. La razón por la que un programa de computador no pueda jamás ser una mente es simplemente que un programa de computador es solamente sintáctico, y las mentes son más que sintácticas. Las mentes son semánticas, en el sentido de que tienen algo más que una estructura formal: tienen un contenido.

Para ilustrar este punto he diseñado un cierto experimento de pensamiento. Imaginemos que un grupo de programadores de computador ha escrito un programa que capacita a un computador para simular que entiende chino. Así, por ejemplo, si al computador se le hace una pregunta en chino, confrontará la pregunta con su memoria o su base de datos, y producirá respuestas adecuadas a las preguntas en chino. Supongamos, por mor del argumento, que las respuestas del computador son tan buenas como las de un hablante nativo del chino. Ahora bien, ¿entiende el computador, según esto, chino? ¿Entiende literalmente chino, de la manera en que los hablantes del chino entienden chino? Bien, imaginemos

que se le encierra a usted en una habitación y que en esta habitación hay diversas cestas llenas de símbolos chinos. Imaginemos que usted (como yo) no entiende chino, pero que se le da un libro de reglas en castellano para manipular esos símbolos chinos. Las reglas especifican las manipulaciones de los símbolos de manera puramente formal, en términos de su sintaxis, no de su semántica. Así la regla podría decir: 'toma un signo changyuan-changyuan de la cesta número uno y ponlo al lado de un signo chongyuon-chongyuon de la cesta número dos'. Supongamos ahora que son introducidos en la habitación algunos otros símbolos chinos, y que se le dan reglas adicionalen para devolver símbolos chinos fuera de la habitación. Supóngase que usted no sabe que los símbolos introducidos en la habitación son denominados 'preguntas' de la gente que está fuera de la habitación, y que los símbolos que usted devuelve fuera de la habitación son denominados 'respuestas a las preguntas'. Supóngase, además, que los programadores son tan buenos al diseñar los programas y que usted es tan bueno manipulando los símbolos que enseguida sus respuestas son indistinguibles de las de un hablante nativo del chino. He aquí que usted está encerrado en su habitación barajando sus símbolos chinos y devolviendo símbolos chinos en respuesta los símbolos chinos que entran. Sobre la base de la situación tal como la he descrito, no hay manera de que usted pueda aprender nada de chino manipulando esos símbolos formales.

Ahora bien, lo esencial de la historieta es simplemente esto: en virtud del cumplimiento de un programa de computador formal desde el punto de vista de un observador externo, usted se comporta exactamente como si entendiese chino, pero a pesar de todo usted no entiende ni palabra de chino. Pero si pasar por el programa de computador apropiado para entender chino no es suficiente para proporcionarle a *usted* comprensión del chino, entonces no es suficiente para proporcionar a *cualquier otro computador digital* comprensión del chino. Y nuevamente, la razón de esto puede enunciarse muy simplemente. Si usted no entiende chino, entonces ningún otro computador podría entender chino puesto

que ningún computador digital, en virtud solamente de pasar un programa, tiene nada que usted no tenga. Todo lo que el computador tiene, como usted tiene también, es un programa formal para manipular símbolos chinos no interpretados. Para repetirlo: un computador tiene una sintaxis, pero no una semántica. Todo el objeto de la parábola de la habitación china es recordarnos un hecho que conocíamos desde el principio. Comprender un lenguaje, o ciertamente, tener estados mentales, incluye algo más que tener un puñado de símbolos formales. Incluye tener una interpretación o un significado agregado a esos símbolos. Y un computador digital, tal como se ha definido, no puede tener más que símbolos formales puesto que la operación del computador, como dije anteriormente, se define en términos de su capacidad para llevar a cabo programas. Y esos programas son especificables de manera puramente formal —esto es, no tienen contenido semántico.

Podemos ver la fuerza de este argumento si contrastamos aquello a lo que se parece el ser preguntado y responder a preguntas en algún lenguaje en el que no tenemos conocimiento alguno de ninguno de los significados de las palabras. Imaginemos que en la habitación china se le dan también a usted preguntas en castellano sobre cosas tales como su edad o episodios de su vida, y que usted responde a esas preguntas. ¿Cuál es la diferencia entre el caso del chino y el caso del castellano? Bien, si igual que yo usted no entiende nada de chino y entiende castellano, entonces la diferencia es obvia. Usted entiende las preguntas en castellano porque están expresadas en símbolos cuyos significados le son conocidos. Similarmente, cuando usted da las respuestas en castellano, está produciendo símbolos que son significativos para usted. Pero en el caso del chino no tiene nada de esto. En el caso del chino usted manipula simplemente símbolos formales de acuerdo con un programa de computador y no les añade significado alguno a ninguno de los elementos.

Se han sugerido varias réplicas a este argumento por parte de las personas que trabajan en inteligencia artificial y en psicología, así como en filosofía. Todas ellas

tienen algo en común: todas ellas son inadecuadas. Y hay una razón obvia por la que tienen que ser inadecuadas, ya que el argumento descansa sobre una verdad muy simple, a saber: la sintaxis sola no es suficiente para la semántica y los computadores digitales en tanto que son computadores tienen, por definición, solamente sintaxis.

Quiero clarificar esto considerando un par de argumentos que se presentan a menudo en contra mía.

Algunas personas intentan responder al ejemplo de la habitación china diciendo que la totalidad del sistema entiende chino. La idea es aquí que aunque yo, la persona que está en la habitación manipulando los símbolos, no entiendo chino, yo soy sólo la unidad de procesamiento central del sistema del computador. Ellos argumentan que es todo el sistema, incluyendo la habitación, las cestas llenas de símbolos y los anaqueles que contienen los programas y quizás también otros elementos, tomado como una totalidad, lo que entiende chino. Pero esto está sujeto exactamente a la misma objeción que hice antes. No hay ninguna manera de que el sistema pueda obtener a partir de la sintaxis la semántica. Yo, como unidad de procesamiento central, no tengo ninguna manera de averiguar lo que significa cualquiera de esos símbolos; pero entonces tampoco puede hacerlo todo el sistema.

Otra respuesta común es imaginar que colocamos el programa de comprensión del chino dentro de un robot. Si el robot se moviese e interactuase casualmente con el mundo ¿no sería esto garantía suficiente de que entendía chino? Una vez más la inexorabilidad de la distinción entre semántica y sintaxis derrota esta maniobra. En la medida en que suponemos que el robot tiene solamente un computador por cerebro, aunque pudiese comportarse como si entendiese chino, no habría con todo manera alguna de obtener a partir de la sintaxis la semántica del chino. Esto puede verse si nos imaginamos que yo soy el computador. Dentro de una habitación en el craneo del robot barajo símbolos sin saber que algunos llegan a mí desde cámaras de televisión adosadas a la cabeza del robot y otros salen para mover los brazos y

piernas del robot. En la medida en que todo lo que tengo es un programa de computador formal, no tengo manera de añadirle significado alguno a ninguno de los símbolos. Y el hecho de que el robot esté inmerso en interacciones causales con el mundo exterior no me ayuda a añadir ningún significado a los símbolos a menos que tenga alguna manera de informarse sobre ese hecho. Supongamos que el robot toma una hamburguesa y esto provoca que entre dentro de la habitación el símbolo de una hamburguesa. En la medida en que todo lo que yo tengo es el símbolo sin ningún conocimiento de sus causas o de cómo llegó allí, no tengo ninguna manera de saber lo que significa. Las interacciones causales entre el robot y el resto del mundo son irrelevantes a menos que esas interacciones causales se representen en una mente cualquiera. Pero no hay manera de que puedan serlo si todo en lo que la llamada mente consiste es un conjunto de operaciones puramente formales, sintácticas.

Es importante ver lo que es afirmado y lo que no es afirmado por mi argumento. Supóngase que planteamos la pregunta que mencioné al principio. '¿Puede pensar una máquina?' Bien, en algún sentido, desde luego, todos somos máquinas. Podemos interpretar la materia que tenemos dentro de nuestras cabezas como una máquina de carne. Y desde luego, podemos pensarlo todo. Así, en un sentido de 'máquina', a saber: ese sentido en el que máquina es solamente un sistema físico que es capaz de realizar cierto género de operaciones, en ese sentido, todos somos máquinas, y podemos pensar. Así, trivialmente, hay máquinas que pueden pensar. Pero esta no era la pregunta que nos intrigaba. Así pues, intentemos una formulación diferente de ella. Podría pensar un artefacto? ¿Podría una máquina hecha por el hombre pensar? Bien, una vez más, depende del género de artefacto. Supóngase que hemos diseñado una máquina que fuera molécula-por-molécula indistinguible de un ser humano. Bien, entonces, si se pueden duplicar las causas, entonces presumiblemente pueden duplicarse los efectos. Así, una vez más, la respuesta a esa pregunta es, en principio al menos, trivialmente sí. Si se pudiese construir una máquina que tuviese la misma estructura que un ser

humano, entonces esa máquina sería capaz de pensar. De hecho, sería un substituto de un ser humano. Bien, intentémoslo de nuevo.

La pregunta no es '¿Puede pensar una máquina?' o '¿puede pensar un artefacto?' La pregunta es: '¿Puede pensar un computador digital?' Pero una vez más hemos de ser muy cuidadosos en cómo interpretamos la pregunta. Desde un punto de vista matemático, cualquier cosa puede describirse *como si* fuera un computador digital. Y esto es así porque puede describirse como si instanciase o llevase a cabo un programa de computador. En un sentido completamente trivial, la pluma que está sobre la mesa enfrente de mí puede describirse como un computador digital. Lo único que sucede es que tiene un programa de computador muy aburrido. El programa dice: 'Estáte ahí.' Ahora bien, puesto que en este sentido cualquier cosa es un computador digital, ya que cualquier cosa puede describirse como si estuviera llevando a cabo un programa de computador, entonces, una vez más, nuestra pregunta obtiene una respuesta trivial. Desde luego nuestros cerebros son computadores digitales, puesto que llevan a cabo un número cualquiera de programas de computador. Y, desde luego, nuestros cerebros pueden pensar. Así, una vez más, hay una respuesta trivial a la pregunta. Pero esa no era realmente la pregunta que estábamos intentando plantear. La pregunta que queríamos plantear es ésta: '¿Puede un computador digital, tal como se ha definido, pensar?' Es decir: '¿Es suficiente para, o constitutivo de, pensar el instanciar o llevar a cabo el programa correcto con los *inputs* y *outputs* correctos?' Y a esta pregunta, a diferencia de sus predecesoras, la respuesta es claramente 'no'. Y es 'no' por la razón que hemos puesto de manifiesto reiteradamente, a saber: el programa del computador está definido de manera puramente sintáctica. Pero pensar es algo más que manipular signos carentes de significado, incluye contenidos semánticos significativos. A esos contenidos semánticos es a lo que nos referiremos mediante 'significado'.

Es importante subrayar de nuevo que no estamos hablando sobre un estadio particular del desarrollo de la

tecnología de los computadores. La argumentación no tiene nada que ver con los próximos y pasmosos avances en la ciencia de la computación. No tienen nada que ver con la distinción entre procesos en serie y en paralelo, o con el tamaño de los programas, o la velocidad de las operaciones del computador, o con computadores que pueden interaccionar casualmente con su entorno, o incluso con la invención de robots. El progreso tecnológico se exagera siempre enormemente, pero incluso eliminando la exageración, el desarrollo de los computadores ha sido extraordinariamente notable, y podemos esperar razonablemente que en el futuro se lleven a cabo progresos aún más notables. Sin duda seremos capaces de simular mucho mejor la conducta humana en los computadores de lo que lo podemos hacer en la actualidad, y ciertamente mucho mejor de lo que hemos sido capaces de hacerlo en el pasado. Lo que quiero decir esencialmente es que si estamos hablando sobre tener estados mentales, sobre tener una mente, todas esas simulaciones son simplemente irrelevantes. No importa cuán buena sea la tecnología o cuán rápidos sean los cálculos realizados por el computador. Si se trata realmente de un computador, sus operaciones tienen que definirse sintácticamente, mientras que la conciencia, los sentimientos, los pensamientos, las sensaciones, las emociones, y todo lo demás incluyen algo más que una sintaxis. Por definición el computador es incapaz de *duplicar* esos rasgos por muy poderosa que pueda ser su capacidad para *simular*. La distinción clave aquí es la que se da entre duplicación y simulación. Y ninguna simulación constituye, por sí misma, duplicación.

Lo que he hecho hasta aquí es proporcionar una base al sentido de que esas citas con las que comencé esta charla son realmente tan absurdas como parecen. Hay, sin embargo, una cuestión problemática en esta discusión, y es ésta: '¿Por qué ha pensado alguien alguna vez que los computadores podrían pensar o tener sensaciones y emociones y todo lo demás?' Después de todo, podemos hacer simulaciones computacionales de cualquier proceso del que pueda darse una descripción formal. Así, podemos hacer una simulación computacional

del flujo de dinero en la economía española, o del modelo de distribución de poder en el partido socialista. Podemos hacer una simulación computacional de las tormentas en los términos municipales del país, o de los incendios en los almacenes del este de Madrid. Ahora bien, en cada uno de esos casos, nadie supone que la simulación computacional es efectivamente la cosa real; nadie supone que una simulación computacional de una tormenta nos deje a todos mojados, o que sea probable que una simulación computacional de un incendio vaya a quemar la casa. ¿Por qué diablos va a suponer alguien que esté en sus cabales que una simulación computacional de procesos mentales tiene efectivamente procesos mentales? Realmente desconozco la respuesta a esto, puesto que la idea me parece desde el principio, para decirlo con franqueza, completamente disparatada. Pero puedo hacer un par de especulaciones.

En primer lugar, hay mucha gente que, cuando de la mente se trata, se siente aún tentada a algún tipo de conductismo. Piensan que si algún sistema se comporta como si entendiese chino, entonces realmente tiene que entender chino. Pero ya hemos refutado esta forma de conductismo con el argumento de la habitación china. Otra suposición hecha por mucha gente es que la mente no es parte del mundo biológico, no es parte del mundo de la naturaleza. El punto de vista de la inteligencia artificial fuerte descansa sobre él en su concepción de que la mente es algo puramente formal; que de una manera u otra no puede ser tratada como un producto concreto de procesos biológicos de la misma manera que otro producto biológico. Hay en esas discusiones, para decirlo brevemente, un género de dualismo residual. Los partidarios de IA creen que la mente es algo más que una parte del mundo biológico natural; creen que la mente es especificable de manera puramente formal. La paradoja de esto es que la literatura de IA está llena de recriminaciones contra algún punto de vista llamado 'dualismo', pero de hecho, toda la tesis de la IA fuerte descansa sobre un género de dualismo. Descansa sobre el rechazo de la idea de que la mente

es sólo un fenómeno biológico natural del mundo igual cualquier otro.

Quiero concluir este capítulo uniendo la tesis del capítulo anterior y la tesis de este. Ambas tesis pueden enunciarse de manera muy simple. Y, de hecho, voy a enunciarlas con, quizás, excesiva crudeza. Pero si las unimos pienso que obtenemos una concepción muy poderosa de las relaciones entre mentes, cerebros y computadores. Y el argumento tiene una estructura muy simple, de modo que usted puede ver si es válido o inválido. La primera premisa es:

1. *Los cerebros causan las mentes.*

Ahora bien, esto es realmente demasiado crudo. Lo que queremos decir mediante esto es que los procesos mentales que nosotros consideramos que constituyen una mente son causados, enteramente causados, por procesos que tienen lugar dentro del cerebro. Pero seamos crudos, y abreviemos esto mediante esas cinco palabras —los cerebros causan las mentes. Escribamos ahora la proposición número dos:

2. *La sintaxis no es suficiente para la semántica.*

Esta proposición es una verdad conceptual. Articula justamente nuestra distinción entre la noción de lo que es puramente formal y lo que tiene contenido. Ahora bien, a esas dos proposiciones —que los cerebros causan las mentes y que la sintaxis no es suficiente para la semántica— añadamos una tercera y una cuarta:

3. *Los programas de computador están definidos enteramente por su estructura formal o sintáctica.*

Considero que esta proposición es verdadera por definición, es parte de lo que queremos decir mediante la noción de un programa de computador.

4. *Las mentes tienen contenidos mentales; específicamente, tienen contenidos semánticos.*

Considero que esto es solamente un hecho obvio acerca de cómo funcionan nuestras mentes. Mis pensamientos, y creencias, y deseos son sobre algo, o se refieren a algo, o conciernen a estados de cosas del mundo; y hacen esto porque sus contenidos los dirigen hacia esos estados de cosas del mundo. Ahora bien, a partir de esas cuatro premisas, podemos extraer nuestra primera conclusión; se sigue obviamente de las premisas 2, 3 y 4:

CONCLUSIÓN 1. *Ningún programa de computador es suficiente por sí mismo para dar un sistema, una mente. Los programas, dicho brevemente, no son mentes, y no son suficientes por sí mismos para tener mentes.*

Ahora bien, esto es una conclusión muy poderosa, porque significa que el proyecto de intentar crear mentes diseñando solamente programas está condenado a muerte desde el principio. Y es importante volver a subrayar que esto no tiene nada que ver con ningún estadio particular en el desarrollo de la tecnología, o con ningún estadio particular de la complejidad del programa. Este es un resultado puramente formal, o lógico, obtenido a partir de un conjunto de axiomas en los que están de acuerdo todos (o casi todos) los participantes en la disputa. Es más, incluso la mayor parte de los entusiastas más acérrimos de la inteligencia artifical están de acuerdo en que de hecho, como un asunto de biología, los procesos cerebrales causan estados mentales, y están de acuerdo en que los programas se definen de manera puramente formal. Pero si se unen estas conclusiones con otras cosas que sabemos, entonces se sigue inmediatamente que el proyecto de IA fuerte es incapaz de ser cumplido.

Sin embargo, una vez que hemos obtenido esos axiomas, veamos qué más puede derivarse. He aquí una segunda conclusión:

CONCLUSIÓN 2. *El modo en que las funciones del cerebro causan las mentes no puede ser solamente en virtud de pasar un programa de computador.*

Y en esta segunda conclusión se sigue de poner en conjunción la primera premisa con nuestra primera conclusión. Esto es, del hecho de que los cerebros causan las mentes y del hecho de que los programas no se bastan para llevar a cabo la tarea, se sigue que el modo en que los cerebros causan las mentes no puede ser solamente en virtud de pasar un programa de computador. Ahora bien, pienso que esto es también un resultado importante, puesto que tiene como consecuencia que el cerebro no es, o al menos no es solamente, un computador digital. Vimos anteriormente que cualquier cosa podía describirse trivialmente como si fuera un computador digital, y los cerebros no constituyen ninguna excepción. Pero la importancia de esta conclusión reside en que las propiedades computacionales del cerebro simplemente no bastan para explicar su funcionamiento para producir estados mentales. Y, en efecto, esto debe parecernos una conclusión científica de sentido común en cualquier caso, ya que todo lo que hace es recordarnos el hecho de que los cerebros son motores biológicos; su biología importa. No es solamente un hecho irrelevante sobre la mente, como diversas personas que trabajan en inteligencia artificial han afirmado, el que suceda que está realizada en los cerebros humanos.

Ahora, a partir de nuestra primera premisa, podemos también derivar una tercera conclusión:

CONCLUSIÓN 3. *Cualquier otra cosa que cause las mentes tendría que tener poderes causales equivalentes al menos a los del cerebro.*

Y esta tercera conclusión es una consecuencia trivial de nuestra primera premisa. Es hasta cierto punto similar a decir que si mi motor de gasolina impulsa mi coche a ciento veinte kilómetros/hora, entonces cualquier motor diesel que fuese capaz de hacer eso tendría que

tener una potencia de salida equivalente al menos a la de mi motor de gasolina. Desde luego, algún otro sistema podría causar procesos mentales usando características bioquímicas o químicas enteramente diferentes de las que el cerebro usa de hecho. Podría suceder que hubiese seres en otros planetas, o en otros sistemas solares, que tuviesen estados mentales y usasen una bioquímica enteramente diferente a la nuestra. Supóngase que los marcianos llegasen a la tierra y concluyésemos que tienen estados mentales. Pero supóngase que cuando abriésemos sus cabezas se descubriese que todo lo que había allí dentro era una mucosidad verde. Bien, con todo, la mucosidad verde, si funcionase de manera tal que produjese conciencia y el resto de su vida mental, tendría que tener poderes causales iguales a los del cerebro humano. Ahora bien, de nuestra primera conclusión, que los programas no bastan, y nuestra tercera conclusión, que cualquier otro sistema tendría que tener poderes causales iguales a los del cerebro, se sigue inmediatamente la conclusión cuatro:

CONCLUSIÓN 4. *Para cualquier artefacto que pudiéramos construir que tuviese estados mentales equivalentes a los estados mentales humanos, el desarrollo de un programa de computador no sería suficiente por sí mismo. Más bien, el artefacto tendría que tener poderes equivalentes a los del cerebro humano.*

El resultado de esta discusión es recordarnos algo que ya sabíamos desde el principio, a saber: que los estados mentales son fenómenos biológicos. La conciencia, la intencionalidad y la causación mental son todas ellas parte de la historia de nuestra vida biológica, junto con el crecimiento, la reproducción, la secreción de la bilis y la digestión.