SIMULANDO A SEARLE[*] [†]

Simulating Searle

Antonio Manuel Liz Gutiérrez
Universidad de La Laguna

RESUMEN

Después de varias décadas de discusión en torno a las críticas de John Searle al Computacionalismo, y en especial a la Inteligencia Artificial, es conveniente tomar cierta perspectiva. A pesar de la gran influencia de sus planteamientos, ni su experimento mental de "La habitación china" ni su distinción entre simular y duplicar pueden ser tomados como mostrando la inviabilidad del computacionalismo. El contraste entre "realidades meramente simuladas" y "auténticas realidades" no es ontológico sino epistemológico. Y su distinción entre simular y duplicar descansa en un uso muy ambiguo de las nociones de "causa" y de "poderes causales". En contra de lo pretendido por Searle, no pueden establecerse *a priori* argumentos concluyentes que apoyen el rechazo del computacionalismo.
Palabras clave: Mente, computacionalismo, inteligencia artificial, simulación, poderes causales.

ABSTRACT

After some decades of discussions about John Searle's criticism of computationalism, and especially of Artificial Intelligence, it is necessary to take those controversies in perspective. In spite of the great influence of Searle's approach, nor his mental experiment of "The Chinesse Room", nor his distinction between simulation and duplication can be taken as showing the failure of Computationalism. The contrast between "merely simulated realities" and "genuine realities" is not ontological, but epistemological. And his distinction between simulation and duplication is supported by a very

* **Recibido** Septiembre de 2008; **aprobado** Noviembre de 2008.
† Este trabajo ha sido llevado a cabo en el seno de los Proyectos de Investigación HUM2005-03848 y FF12008-01205 (España). Agradezco a los revisores anónimos de la revista sus comentarios y sugerencias a una versión anterior del mismo.

ambiguous use of the notions of "cause" and "causal powers". Against what is intended by Searle, there are not conclusive *a priory* arguments for a rejection of Computationalism.
Key Words: Mind, computationalism, artificial intelligence, simulation, causal powers.

Desde hace ya varias décadas, John Searle ha criticado con dureza el computacionalismo dominante dentro de la actual Ciencia Cognitiva y, más concretamente, el ambicioso programa de la Inteligencia Artificial. Y sus críticas han llegado a ser enormemente influyentes y populares a pesar de los numerosos rechazos y enfrentamientos a los que se han visto sometidas. Seguramente sea ya el momento de mirar atrás con cierta perspectiva y preguntarnos qué conclusiones debemos obtener de todas estas controversias.

Básicamente, los planteamienos de Searle siguen dos líneas argumentales. Una de ellas se centra en su famoso experimento mental de "La habitación china". La otra se centra en la distinción entre simular y duplicar. En ambas está en juego la noción de simulación. Pero esta noción interviene en dos diferentes contrastes. Mientras que en la primera línea argumental "simular una vida mental" se opone a "tener auténticamente esa vida mental", en la segunda línea argumental "simular una vida mental" se opone a "duplicarla".

En las páginas que siguen examinaremos más de cerca estos dos contrastes. Y a través de nuestros análisis, obtendremos una importante conclusión. El primer contraste, que vertebra crucialmente el experimento mental de Searle de "La habitación China", carece de fuerza suficiente como para apoyar, como pretende Searle, un rechazo completo al computacionalismo. Creo que este resultado es muy claro. Pese a las apariencias (y nunca mejor dicho), la distinción entre "realidades meramente simuladas" y "auténticas realidades" no es ontológica, sino epistemológica. Es una distinción que trazamos desde dentro de nuestras creencias, hipótesis, teorías, interpretaciones y atribuciones. Por sí mismas, las que damos en llamar "realidades meramente simuladas" tienen tanto derecho a ser partes genuinas de la realidad como aquellas que llamamos "auténticas realidades". Por decirlo así, la distinción entre lo "meramente simulado" y lo "auténtico" la ponemos nosotros. Como digo, este resultado es muy claro. Incluso debería ser claro para el propio Searle[1]. Y lo sorprendente es que haya sido obviado. Lo

[1] Sobre todo, teniendo en cuenta sus tesis de que la propia "información" tiene una realidad ontológica subjetiva. Aunque sea epistemológicamente objetiva, la información y las relaciones informacionales serían siempre propiedades dependientes de un observador. No son rasgos intrínsecos del mundo. Siendo esto así, ¿cómo podría tener un carácter ontológico

sorprendente es que haya permanecido oculto en la mayor parte de las polémicas.

Así pues, en contra de cierto estado de opinión firmemente establecido, el experimento mental de "La habitación china" no puede conducir de manera concluyente y, por decirlo así "a priori", a un rechazo del computacionalismo. Y aquí es donde interviene, no menos crucialmente, la segunda línea argumental empleada por Searle. Gran parte de la responsabilidad de que haya sido obviado el hecho de que toda distinción entre lo "meramente simulado" y lo "auténtico" es, en algún sentido relevante, supuesta por nosotros se deriva de la fuerte retórica empleada por Searle a propósito del otro contraste en el que interviene la noción de simulación: el contraste entre "simular" y "duplicar". Searle repite una y otra vez que simular una vida mental no es lo mismo que duplicarla. Y ciertamente no puede decirse nada en contra de tal afirmación. Pero a la hora de explicar con mayor precisión en qué consistiría tal "simulación" y en qué consistiría una "duplicación", utiliza de manera muy ambigua las nociones de "causa" y "de poderes causales". Y como veremos, aquí es donde surgen los problemas. Pues si, por un lado, pensamos en términos de poderes causales, el que duplicar una vida mental requiera la presencia de los poderes causales que la hacen posible se convierte en una verdad trivial. El que cualquier cosa capaz de generar cierto efecto ha de tener el poder causal de producirlo es una tautología que se sigue de la noción misma de poder causal. Y si, por otro lado, pensamos en términos de causas, no puede resultar nada extraño que se intente duplicar una vida mental a través de causas diferentes de las que pudieron producirla. Simplemente estaríamos aplicando el principio general de que la mayoría de las cosas que ocurren en el mundo de hecho pueden ser producidas por causas diversas. Para identificar otras causas diferentes de una vida mental utilizaríamos simulaciones de la producción de algunos de sus efectos causales. Y en esto consistiría justamente el proyecto de la Inteligencia Artificial en su sentido más ambicioso.

Tampoco parece que esta segunda vía argumental de Searle tenga todo el peso contundente que habitualmente se le atribuye. Más bien, su fuerza es sólo retórica. ¿Qué podremos, pues, aprender de toda la larga polémica iniciada por Searle hace ya casi tres décadas? Mi conclusión será que, sobre todo, debemos aprender que no hay argumentos concluyentes que, de forma *a priori*, permitan rechazar completamente el computacionalismo, ni siquiera en sus versiones más extremas. Y como sabemos que ningún ar-

la distinción entre "meras simulaciones" y "auténticas realidades"? En sí mismo, nada es una simulación. Las simulaciones sólo son "simulaciones" desde el punto de vista de un sujeto que las puede llegar a "confundir" con las "auténticas realidades".

gumento empírico puede tener tampoco ese carácter concluyente, no nos quedará más remedio que seguir viviendo en el arriesgado terreno de las hipótesis y conjeturas más o menos fundadas. Otra cosa sería vivir en la ilusión.

1. Computacionalismo e Inteligencia Artificial

El computacionalismo es la posición dominante dentro de las actuales Ciencias Cognitivas. Y para el computacionalismo, lo psicológico ocupa un espacio muy bien delimitado. Su naturaleza consiste en ciertas tramas de relaciones funcionales que definen una muy especial estructura computacional. Como es sabido, Searle pone en cuestión esta concepción resaltando la distinción entre sintaxis y semántica. La estructura computacional en la que suelen fijarse las variedades más clásicas del computacionalismo es una estructura formal de tipo sintáctico. Es la estructura computacional de carácter sintáctico típica de los programas informáticos. Y Searle argumenta que la sintaxis sola nunca es suficiente para conseguir una semántica ni, por consiguiente, para generar una genuina vida mental llena de significado, llena de todo tipo de contenidos mentales. La auténtica vida mental en su integridad, es decir, una intencionalidad llena de contenidos, con una plena conciencia cualitativa de los mismos, etc., es un producto causal de nuestros cerebros. La sintaxis por sí sola únicamente podrá ofrecer simulacros de vida mental. Simulaciones que jamás conseguirán producir o duplicar una auténtica vida mental. Y que tampoco servirán para explicar nada.

La Inteligencia Artificial es una disciplina integrada en el amplio campo de las Ciencias Cognitivas. Estudia los procesos informacionales y las reglas subyacentes a ciertas tareas que requieren inteligencia. Tareas muy generales como las de resolver problemas, tomar decisiones, reconocer patrones, formar hipótesis a la luz de ciertas evidencias, etc., o tareas más concretas como las de jugar al ajedrez, subir y bajar escaleras o mantener una conversación.

Searle distingue tajantemente entre dos clases de proyectos dentro del campo de la Inteligencia Artificial: la Inteligencia Artificial en sentido débil y la Inteligencia Artificial en sentido fuerte[2]. Para la primera, el computador es sólo una herramienta de trabajo para estudiar ciertos procesos abstractos realizables mediante procesos físicos. Especialmente, procesos capaces de representar determinadas funciones mentales. Para la segunda, en

[2] Esta distinción, así como la distinción entre sintaxis y semántica, es constante en casi todos los trabajos de Searle dedicados a la filosofía de la mente. Véase, por ejemplo, Searle (1980, 85, 87, 90, 92, 97 y 98).

cambio, los computadores no son sólo herramientas de trabajo. Un computador apropiadamente programado tendría literalmente esas funciones mentales, tendría mente.

Searle no tiene ninguna objeción que hacer al proyecto de la Inteligencia Artificial en su sentido débil. En el fondo, el computador es aquí simplemente una prolongación del bolígrafo y del papel. Sin embargo, el segundo proyecto, el proyecto de la Inteligencia Artificial en sentido fuerte, le parece completamente desencaminado. Este proyecto no sólo subrayaría las analogías entre el funcionamiento del cerebro y el funcionamiento de ciertos mecanismos computacionales, sino que afirmaría categóricamente que la mente es al cerebro lo que el «software» de un computador, el programa que está ejecutando, es a su «hardware», a su maquinaria específica. Y esto es, para Searle, un tremendo desatino.

La argumentación de Searle es muy conocida. Pocos experimentos mentales de uso filosófico han sido tan discutidos, y sobre todo tan divulgados, como el suyo relativo a "La habitación china". Y la reacción habitual, fuera de los ámbitos más académicos, ha sido muy favorable al carácter conclusivo del mismo. No es ésta mi opinión. Creo que hay algo erróneo en la manera en que Searle usa ese experimento mental para criticar el proyecto de la Inteligencia Artificial en su sentido fuerte. Así pues, voy a ir contra corriente. Pero creo que la posición a la que finalmente llegaré es mucho más adecuada, incluso intuitivamente, que la sugerida por Searle. Debo advertir, sin embargo, que no es mi propósito argumentar que la semántica pueda surgir de la sintaxis. Seguramente no pueda. Ni tampoco que nuestra mente no sea más que un programa informático. Seguramente no lo es. Lo que quiero criticar es la forma peculiar en la que Searle desarrolla su argumentación.

2. ¿Puede pensar una máquina? El juego de imitación de Türing

Türing propuso analizar el problema de la atribución de propiedades mentales a través de un curioso test, el llamado "juego de imitación". Se trata de lo siguiente:

Imaginemos la siguiente situación. Un hombre A y una mujer B contestan a las preguntas que un interrogador C les hace con objeto de descubrir cuál de ellos es el hombre y cuál la mujer. La comunicación se realiza a través de algún medio que impida su identificación directa, por ejemplo a través de unos monitores donde aparecen escritas las preguntas y las respuestas. A intenta que C se equivoque al realizar la identificación. B, en cambio, procura ayudar a C. El interrogador C, a veces acierta en la identificación y otras veces se equivoca. Si, en este juego, hacemos que una máquina (un computador) realice perfectamente el papel asignado a A; es de-

cir, produzca en el monitor respuestas adecuadas a las preguntas de C, y a las intenciones que A tenía de confundir a C en su identificación, y si la frecuencia con la que C se equivoca resulta ser la misma, o aproximadamente la misma, antes y ahora, entonces qué razón, se preguntaba Türing, podrá hacer que C sospeche que A es una máquina.

El peculiar test que acabamos de presentar acerca de la identificación o atribución de propiedades mentales fue originalmente propuesto por Türing a mediados del siglo XX[3]. Los desarrollos matemáticos de Türing, y otros, contribuyeron decisivamente a la constitución de la disciplina que se conoce hoy día como «Inteligencia Artificial». La pregunta ¿puede pensar una máquina? sería sustituible, sugiere Türing, por preguntas como ¿pueden ciertas máquinas jugar siempre bien nuestro juego de imitación? o, más en general, ¿pueden ciertas máquinas ser capaces de jugar perfectamente este tipo de juegos en cualquier situación?

Son muchas las cuestiones que inmediatamente surgen en torno al juego de imitación. ¿Es lo mental solamente una capacidad para mantener coherentemente «diálogos» de cierto tipo? ¿De cualquier tipo? ¿Consiste lo mental en llevar a cabo ciertas conductas lingüísticas, en manipular sintácticamente bien un conjunto de símbolos? Tal vez Türing pensase que sí, que cualquier pregunta interesante acerca de la mente podía ser reformulada como una pregunta acerca de ciertas capacidades de manipulación de símbolos. También Descartes, por ejemplo, consideraba nuestra capacidad lingüística como la pieza clave que, haciendo intervenir los elementos más esenciales de nuestra vida mental, elementos vinculados a la libertad y a la conciencia, nos distinguía de los animales y demás máquinas sin pensamiento. El computacionalismo no está muy alejado de esta posición. Una simulación perfecta de nuestras habilidades simbólicas sería, entonces, suficiente para atribuir correctamente todas las propiedades mentales. Y nada más quedaría por decir. Las entidades a las que atribuyéramos, en esas condiciones, ciertas propiedades mentales tendrían realmente esas propiedades mentales. Desde esta perspectiva, toda propiedad mental se identifica con algún tipo de capacidad de manipulación de símbolos empíricamente discernible. Y la constatación de esta capacidad garantizaría la posesión de esas propiedades mentales.

¿Qué reacciones suscita el anterior planteamiento? Podrían argüirse, al menos, tres objeciones inmediatas. La primera de ellas consistiría en sostener que existen aspectos de lo mental que no se refieren tanto a la manipulación de símbolos cuanto a la capacidad de realizar acciones adecuadas en relación con un entorno. Pero es fácil escapar a esta objeción. Nada nos

[3] Véase Türing (1950).

impide complicar las cosas para la máquina que sustituye a A en nuestro juego. Podemos hacer de ella un sofisticado robot con capacidad de movimiento y control motor. Además de «hablar» con la máquina, C podría verla «actuar», incluso «cooperar» o «competir» con ella. El peculiar programa seguido por la máquina adquiriría aquí una relevancia máxima. El objetivo ahora es conseguir unos determinados efectos conductuales no sólo de tipo lingüístico. Pero el medio de alcanzar ese objetivo sigue basándose en la ejecución de un determinado programa. La identificación de Türing se ampliaría. Las propiedades mentales se identificarían aquí con otras propiedades computacionales más complejas. No sólo con propiedades referidas a manipulaciones simbólicas que desemboquen en conductas lingüísticas, sino también con propiedades simbólicas relativas a otros tipos de conductas. El juego de imitación, simplemente, sería más sofisticado. Y la objeción no sería definitiva. Podría ser respondida, de un modo muy sencillo, ampliando el juego.

La segunda y la tercera objeción, sin embargo, se dirigen al núcleo mismo de la propuesta de Türing. Pueden iniciarse señalando que imitar, o simular, lo mental es algo muy distinto de tener una auténtica mente. Podemos llevar hasta el extremo la apariencia humana de nuestro robot. Por encima y por debajo de su piel. Podemos hacer que su «conversación», sus «acciones» y, también, su «interior» sean ahora, para nosotros, indistinguibles de nuestra conversación, nuestras acciones y nuestro interior. Cuando nos situamos en este límite, cuando la imitación o simulación es perfecta en todos los sentidos, ¿qué razón podríamos tener para pensar siquiera en una simple imitación, en una simulación? Tal vez ninguna, pero las dos objeciones coincidirían en sostener que aunque estuviéramos entonces plenamente autorizados para atribuir propiedades mentales, las entidades a las que atribuyéramos esas propiedades podrían no tenerlas realmente.

Sin embargo, las razones para mantener esto último son distintas en cada tipo de objeción. Para la segunda objeción, aunque nosotros no apreciemos diferencia alguna entre los originales auténticos y las imitaciones, tal vez sí la haya, y sea decisiva, en relación a los diversos materiales de que estén hechos los originales y las imitaciones. Lo mental no sería indiferente a la clase de materia, de «hardware», de que estén hechas las cosas a las que atribuimos propiedades mentales.

La distinción entre propiedades del «software», propiedades computacionales (funcionales, formales, etc.) y propiedades del «hardware» (propiedades relativas a la constitución material), es siempre relativa a ciertas formas de descripción. Puede ser, por tanto, muy difícil precisar los límites de lo simulable. No obstante, debe siempre existir alguna propiedad constitutiva, no formal sino de tipo material, de la clase de entidades que tengan

"auténticas" propiedades mentales. Y esa propiedad constitutiva material debe ser capaz de distinguir a tales entidades, bajo las mismas descripciones computacionales (funcionales, formales, etc.), de todas aquellas entidades que sólo "simulen" tener esas propiedades mentales. Una simulación perfecta para nosotros, perfecta en cualquier límite imaginable, aunque acaso justifique la atribución de propiedades mentales, no garantiza por ello su posesión.

Vayamos ahora a la tercera objeción. Esta objeción sostendría que «tener mente» es una propiedad cuya instanciación (ejemplificación, etc.) sólo puede ser conocida con seguridad por aquellas entidades en las que se encuentre instanciada. Si una piedra tuviera mente, sabría que la tiene. Sin embargo, no podría saber con seguridad si nosotros tenemos o no mente. Las entidades que tienen mente, saben que la tienen. Pero ninguna entidad con mente puede saber nunca con seguridad si algo distinto de ella tiene mente o no. El ejemplo privilegiado de esta concepción de lo mental son las sensaciones. Por mucho que nuestra máquina imite a A, en el juego de imitación de Türing, no poseerá mente a menos que sea capaz de tener cosas como sensaciones (emociones, sentimientos, afectos, etc., estados mentales con algún componente "cualitativo"). Y «ser consciente», «comprender», «tener creencias o deseos», «recordar», etc., involucraría siempre, en esta perspectiva, algo así como tener sensaciones de cierto tipo (o emociones, sentimientos, afectos, etc.).

Tener mente se convierte aquí en una propiedad de uso exclusivamente "privado", imposible de enajenar bajo ninguna presunta identificación computacional o material. Ninguna simulación, por muy perfecta e ideal que sea, en cualquier extensión y profundidad, podrá garantizar jamás el que algo tenga mente, incluso aunque todos nuestros criterios para atribuir propiedades mentales resulten plenamente satisfechos. Y tampoco bastará ninguna supuesta identidad material. Esta era la objeción que el propio Türing consideraba más seriamente. Pero era una objeción que, según él, inevitablemente conducía al solipsismo[4].

Así pues, hemos llegado a tres resultados distintos. En la línea marcada por Türing, y salvando la primera objeción, nos encontramos en primer lugar con una concepción computacional (funcionalista, formal, etc.) de lo mental, concepción que inspira gran parte de los desarrollos de las actuales Ciencias Cognitivas y, por supuesto, de lo que hemos denominado Inteligencia Artificial en sentido fuerte. Según esta concepción, la mente sería

[4] Véanse al respecto los interesantes comentarios de Türing (1950) a las posibles objeciones a su propuesta. Como veremos, habrá también una peligrosa deriva solipsista en la manera como Searle quiere interpretar su experimento mental de "La habitación china".

una "forma" de organización de la "materia". Un segundo resultado, completamente diferente, se desprendería de la segunda objeción expuesta. Ello supondría una revisión profunda del computacionalismo. La última palabra sobre lo mental seguramente la tengan, entonces, disciplinas como la neurofisiología. Y un tercer resultado que puede ser descrito como una inamovible posición privilegiada (¡cartesiana!) para lo mental. Los fenómenos mentales involucrarían propiedades últimas e irreductibles a cualquier otro tipo de propiedades, ya sean éstas computacionales (funcionales, formales, etc.) o materiales.

Como veremos a continuación, Searle arremete contra cualquier concepción de lo mental cercana a los planteamientos de Türing. Pero reconoce, no obstante, el carácter poco decisivo del primer tipo de objeción. Ciertamente, sería muy fácil salvar esta objeción, al menos en principio, simplemente con más de lo mismo[5]. Searle se moverá ambiguamente entre las objeciones segunda y tercera.

3. Türing en la habitación china

Searle realiza una vuelta de tuerca sobre el juego de la simulación de Türing. Una decisiva manera de contrastar cualquier teoría de la mente, sostiene Searle, consiste en preguntarnos qué sería para nosotros tener mente, tener estados y procesos mentales, si nuestra mente funcionara según los principios de tal teoría. Y a fin de hacer esto con la concepción de la mente subyacente al proyecto de la Inteligencia Artificial en sentido fuerte, un computacionalismo de tipo sintáctico basado en programas informáticos, Searle propone el siguiente experimento mental, conocido como "La habitación china"[6]:

Supongamos que estamos encerrados en una habitación donde nos van introduciendo una serie de símbolos (que, de hecho, resultan ser frases en chino), y que se nos pide que enviemos fuera de la habitación otros conjuntos de símbolos, respetando ciertas reglas. El intercambio se lleva a cabo a través de una hoja de papel o a través de un monitor de ordenador, esto no importa mucho. Lo que sí es importante es que, para realizar la tarea, debamos consultar un manual de instrucciones que nos dice cómo realizar los intercambios. El manual está escrito en castellano y nos capacita para correlacionar conjuntos de símbolos exclusivamente en virtud de su «forma», en virtud de sus características formales o sintácticas. Supongamos que no

[5] Sin embargo, ha sido computacionalmente mucho más fácil diseñar máquinas capaces de jugar al ajedrez, o simplemente calculadoras de bolsillo, que máquinas capaces de subir y bajar escaleras con soltura, ¡por no hablar de máquinas capaces, por ejemplo, de jugar al futbol!

[6] Ver principalmente Searle (1980 y 85).

sabemos una palabra de chino. Al cabo de cierto tiempo, sin embargo, podríamos "aparentar", vistos desde fuera de la habitación, que somos tan capaces de mantener conversaciones en chino como podría hacerlo cualquier hablante competente chino. Al hacerlo, nos estaríamos comportando exactamente igual a un computador programado para mantener conversaciones en chino. Pero esto no sería más que una pura apariencia, una mera simulación. En el fondo, seguiríamos sin saber una palabra de chino. Hasta podríamos ignorar que esos conjuntos de símbolos son realmente frases en chino y que el manual de instrucciones es realmente un manual para mantener conversaciones en chino.

La ejecución de las reglas de mi manual no me hace entender chino[7]. Paralelamente, la explicitación de las reglas seguidas por mí al mantener en la anterior situación aparentes conversaciones en chino, desde una cierta perspectiva "externa", no constituiría ninguna explicación suficiente de la habilidad que un hablante chino tiene para hablar y entender chino.

Cualesquiera que sean los principios y reglas formales incluidos en el manual de instrucciones, Searle argumenta que no serán suficientes para conseguir entender chino. Un ser humano podría seguir esos principios y reglas formales sin llegar jamás a entender chino, y sin llegar jamás tampoco ni a darse cuenta siquiera de que aparentemente está hablando chino en lugar de, por ejemplo, llevando a cabo un extraño juego.

Searle examina y rechaza una serie de posibles réplicas a su planteamiento. Muy brevemente, son siguientes[8]:

1. La réplica en términos de sistemas: Consiste en decir que el sujeto situado en la habitación china (llamaremos así al escenario descrito en el experimento mental) es sólo parte de un sistema más amplio formado por él, el manual de instrucciones, los papeles o monitores en los que se envían o devuelven los mensajes, las relaciones entre todas estas cosas, etc., y que sería ese sistema global el que sí entiende chino.

La respuesta de Searle a esta réplica es rotunda. Supongamos que todos esos componentes del sistema entero están interiorizados en el sujeto. El sujeto conoce de memoria el manual y no necesita papeles ni monitores. De alguna forma recibe ciertos «inputs» y devuelve unos «outputs», pero todo lo que media entre ellos se encuentra ahora dentro de él. Es fácil concluir, dice Searle, que no por ello el sujeto conseguirá, mejor que antes, entender chino. El pensamiento, en contra de lo que mantendría la Inteligencia Arti-

[7] Como a continuación argumentaremos, siendo completamente precisos deberíamos decir que "no me permiten afirmar que entienda chino". Y éste será el punto crucial.

[8] Estas objeciones son consideradas y rechazadas por Searle en muchos de sus trabajos, desde Searle (1980).

ficial en sentido fuerte, no es únicamente un asunto de manipulación formal de símbolos.

2. La réplica de la robótica: Supongamos, diría este tipo de réplica, que instalamos dentro de un robot un computador que sigue el manual para mantener conversaciones en chino. El robot "percibe", "se mueve", etc., de manera enteramente similar a como nosotros percibimos y nos movemos en nuestro entorno. Este robot sí que tendría ahora un genuino entendimiento de las conversaciones que se mantienen en chino. Y sí que tendría genuinos estados y procesos mentales.

Searle responde a esta réplica sugiriendo que nos imaginemos a la habitación china donde está el sujeto, donde también podemos estar nosotros mismos, instalada dentro del robot. A esa habitación podrían llegar los mismos «inputs» que antes desde el punto de vista formal. Y los «outputs» también serían los mismos. Los «inputs» provendrían ahora de los mecanismos perceptivos del robot y algunos «outputs» irían acompañados de ciertos movimientos en el robot. Por ejemplo, el brazo del robot podría señalar ciertos lugares cuando yo dijera, en chino, «allí». Yo recibiría información del aparato perceptual del robot y, a veces, enviaría ciertas órdenes a su aparato motor. Pero, insiste Searle, nada de esto cambiaría las cosas. Yo seguiría sin entender una palabra de chino si estuviera dentro de esa habitación china instalada, a su vez, dentro del robot.

3. La réplica de la simulación del cerebro: Supongamos ahora que nuestro manual de instrucciones, nuestro programa, es de un tipo muy diferente. Construimos un computador capaz de procesar información de manera enteramente similar a como la procesan nuestros cerebros (masivamente en paralelo, de forma no simbólica, etc.). Y hacemos que el computador así construido simule completamente los procesos informacionales que tienen lugar en el cerebro de un hablante chino cuando desarrolla conversaciones en chino. Negar aquí que nuestro computador entiende chino implicaría, según esta réplica, negar que los propios hablantes chinos entienden chino.

Este tipo de réplica seguiría aún dependiendo, para Searle, del supuesto general de toda la Inteligencia Artificial en sentido fuerte. Se sigue suponiendo que no necesitamos conocer cómo es, en detalle, el cerebro para conocer cómo funciona la mente. Se sigue suponiendo que existe un nivel formal y computacional que constituye la esencia de la mente con independencia de la materia concreta de nuestros cerebros[9]. Y el problema con la

[9] Véase al respecto la interesante polémica entre Searle y los Churchland publicada en la revista *Investigación y Ciencia* de Marzo de 1990 (versión en castellano). El artículo de Paul y Patricia Churchland se titulaba "¿Podría pensar una máquina?", el artículo de John Searle se titulaba "¿Es la mente un programa informático?".

simulación del cerebro es que simula tan sólo la estructura formal de las secuencias de activaciones neuronales. A esa simulación no le importan las propiedades causales del cerebro, propiedades causales entre las que puede estar justamente su capacidad para producir estados intencionales y conciencia.

4. La réplica de la combinación de factores: Si tomamos juntas las tres réplicas anteriores, podemos tener un caso más plausible de comprensión del chino. Imaginemos, sugiere esta réplica, un completo sistema integrado formado por un robot y un sistema computacional instalado en su interior que simule perfectamente todos los procesos que tienen lugar en el cerebro de un hablante chino. ¿Qué más se puede pedir?, preguntarían los proponentes de esta réplica.

Pero, según Searle, sí se puede pedir más. Si por separado las anteriores réplicas son criticables, para Searle también lo serían tomadas conjuntamente. Simular que se entiende chino nunca es suficiente para entenderlo. La materia de nuestros cerebros es sumamente importante. De ella debe depender, en último término, que se produzca o no realmente una vida mental apropiada, que se produzca o no comprensión del chino.

5. La réplica del conocimiento de otras mentes: ¿Cómo sabemos que otros sujetos realmente comprenden chino? Según este tipo de réplica, sólo a partir de su conducta. Si fueran superados todos los criterios conductuales relevantes, no podríamos decir que no se entiende chino. Y esto se aplicaría tanto al computador y al robot como a nosotros mismos.

Esta nueva réplica se desmantelaría, dice Searle, distinguiendo entre, por un lado, conocer que otros sujetos tienen ciertos estados y procesos mentales y, por otro lado, conocer qué son esos estados y procesos (Searle, 1992, desarrolla en profundidad este argumento). Además, si fuéramos nosotros mismos los que estuviéramos en la habitación china, no podría ya plantearse esta réplica. Si estuviéramos en la habitación china, nos daríamos cuenta, insiste Searle, de que seguiríamos sin saber ni una palabra de chino.

6. La réplica del desarrollo futuro de la Inteligencia Artificial: La Inteligencia Artificial en sentido fuerte tal vez pueda llegar a desarrollar máquinas capaces de repetir, en otros medios materiales no orgánicos, los poderes causales que en nuestros cerebros sean constitutivos de su vida mental. De esta forma, la Inteligencia Artificial en sentido fuerte tal vez pueda llegar a crear, literalmente, máquinas inteligentes. Y no tenemos ninguna base, dado el estado actual de nuestra ciencia y tecnología, para negar esta posibilidad.

Searle no tiene ninguna objeción específica para esta réplica. Pero señala que se situaría ya fuera del programa de la Inteligencia Artificial en su

sentido fuerte, programa que partía de la tesis de que los procesos mentales eran esencialmente procesos computacionales realizados sobre elementos formalmente definidos.

A través de todas las anteriores objeciones y réplicas, Searle insiste una y otra vez en los mismos puntos. Sus réplicas a las objeciones presentadas remiten siempre a la situación originalmente descrita en su experimento mental de la habitación china. Tener una vida mental no puede derivarse del hecho de que se siga ningún programa. Ningún conjunto de reglas sintácticas de manipulación de símbolos o, en general, ningún conjunto de reglas formales, seguidas por un sistema, puede hacer que tal sistema tenga auténticos estados y procesos mentales.

4. Otra vuelta más de tuerca: simulando a Searle

Incluso aunque se lleve la simulación a su límite, Searle insiste en que debe existir una última diferencia ontológica entre tener una auténtica vida mental y sólo simular o aparentar que se tiene. Pero, ¿es así? ¿Debe existir siempre esa "última diferencia ontológica"?

Searle no considera una posible situación como la que presentaremos a continuación. Se trataría de una situación capaz de generar la ilusión de esa diferencia ontológica y que, por lo tanto, puede hacernos dudar de que realmente la distinción entre una auténtica vida mental y una vida mental tan sólo simulada pueda establecerse en unos términos tan exclusivamente ontológicos como pretende Searle.

Mi objetivo es mostrar que dicha distinción sólo es (¡sólo puede ser!) una distinción epistemológica. En otras palabras, que estrictamente la distinción entre "auténticas vidas mentales" y "vidas mentales tan sólo simuladas" no existe en la realidad misma. Existe tan sólo dentro de los límites de lo que creemos conocer de ella. Simplemente, fuera de la perspectiva subjetiva de nuestro conocimiento ¡no hay simulaciones! Sin una perspectiva subjetiva, sin un determinado punto de vista epistemológico sobre la realidad, no hay más simulación de vida mental en un robot, o en cualquier otro de los ingenios ofrecidos por la Inteligencia artificial, que la simulación que pueda haber, por ejemplo, en una alcachofa, o en cualquier simple piedra escogida al azar.

5. La ilusión de una diferencia ontológica

Describamos ya esa situación en la que podría generarse la ilusión de que existe una diferencia ontológica última entre nuestra propia vida mental y una vida mental tan sólo simulada.

Nos basaremos en dos supuestos. En primer lugar, supongamos que conseguimos construir máquinas que simulan perfectamente todos los rasgos

formales de nuestra vida mental. Y que consiguen simular así, por ejemplo, la vida mental del propio Searle. Cada una de estas máquinas es un robot prácticamente indistinguible de nosotros mismos (similar a la del propio Searle). Su actividad en cualquier entorno físico o social es enteramente similar a la nuestra. Su organización funcional interna, su estructura formal, simula perfectamente nuestros sistemas neuronales, etc., etc., etc. Podemos suponer en este punto una combinación de todos los factores que intentaba integrar la cuarta réplica, la réplica basada en la combinación de factores.

Nuestro segundo supuesto es el siguiente. Instalamos cierto número de esas máquinas en la habitación china. Y las organizamos de manera que sean los componentes de "un nuevo sistema integrado". Esto es, las máquinas instaladas en la habitación china se organizan simulando de nuevo todos nuestros procesos neuronales. Y se instalan en un nuevo robot, también prácticamente indistinguible de nosotros mismos, aunque tal vez de un tamaño un poco mayor. Como decimos, conseguimos así un segundo sistema mucho más complejo y sofisticado.

Imaginemos ahora que las primeras máquinas «se nieguen a reconocer» que este segundo sistema pueda llegar a entender chino en esta nueva situación. Se «niegan a reconocerlo» de una forma enteramente similar a como nosotros nos negaríamos a reconocerlo. ¡Se "niegan a reconocerlo" de forma enteramente similar a como Searle se niega a reconocer que las primeras máquinas puedan hacerlo!

Nos falta una pequeña convención: expresaremos mediante las comillas los reparos que mostraría Searle a que estemos realmente ante un auténtico estado mental. «Negarse a reconocerlo» sería, en principio, tan sólo aparentar que se niegan a reconocerlo. Con todo esto, hemos diseñado una curiosa y compleja situación. Y podemos interpretarla de dos maneras radicalmente distintas. De las siguientes dos maneras:

Primera interpretación: Podemos interpretarla, en primer lugar, rechazando tanto que el segundo sistema integrado realmente entienda chino como que las primeras máquinas que simulan nuestra vida mental (¡la del propio Searle!) realmente tengan una auténtica vida mental. Como se sospechará, ésta sería la interpretación preferida por Searle. ¡Pero no es la única!

Segunda interpretación: También podemos admitir que nuestras primeras máquinas sí tienen realmente una vida mental como la nuestra, que esa vida mental es del mismo tipo que la que también tiene el segundo sistema integrado y, por último, que hay algo en esa vida mental, compartida por nosotros, por cada una de las primeras máquinas y por el segundo sistema

integrado, que produce la ilusión, algo así como una especie de inevitable «ilusión de usuario»[10], de que tal vida mental no puede ser el resultado de una simple reproducción de los aspectos computacionales (funcionales, formales, etc.) de nuestra vida mental. Esas máquinas, como también muchos de nosotros, "suponen" (ahora deberíamos decir, suponen) que siempre debe haber algo más en su vida mental de lo que pueda haber en el segundo sistema integrado.

En esta última interpretación, el segundo sistema integrado podría tener una vida mental similar a la vida mental de las primeras máquinas que lo componen. Y también similar a la nuestra. Sin embargo, al igual que haríamos muchos de nosotros (¡al igual que haría Searle!), cada una de esas primeras máquinas "niega" (mejor dicho, niega) rotundamente que para entender chino baste tener dentro de sí la estructura computacional (funcional, formal, etc.) que tiene el segundo sistema integrado. Podemos incluso imaginar a algunas de esas primeras máquinas "comentando con íntima satisfacción" (mejor dicho, comentando con íntima satisfacción) el experimento mental de Searle de "La habitación china".

El caso es que tengan o no tengan las primeras máquinas de nuestro ejemplo una genuina vida mental, rechazarán (o acaso tan sólo «rechazarán») que el segundo sistema integrado la tiene. Y por lo tanto, y éste es el punto crucial, ese simple rechazo (o tal vez «rechazo») ¡no podrá ser utilizado como criterio para descubrir si el segundo sistema integrado tiene o no una auténtica vida mental! Repitámoslo: ¡no puede ser un criterio!

La existencia de tal rechazo (o «rechazo») no puede servirnos para elegir una interpretación del primer tipo frente a una interpretación del segundo tipo. Aunque parezca incontestable que no conseguiríamos comprender chino en la habitación china, no es incontestable. Ni es obvio que esto baste para poner a prueba nuestra teoría de la mente. En contra de lo que sostiene Searle, tal rechazo (o, insistimos una vez más, "rechazo") ¡no puede servir como piedra última de toque de una teoría de la mente!

Searle posiblemente señalaría con vehemencia todos nuestros anteriores paréntesis e insistiría en que debe haber una última diferencia ontológica entre el auténtico rechazo que podríamos hacer nosotros y el sospechoso «rechazo» de nuestras máquinas. Pero, ¿cómo dar sentido a esta presunta

[10] En el sentido empleado por Dennett (1991). Junto con Patricia y Paul Churchland, desde sus primeras formulaciones Dennett ha sido una de las figuras más combativas en contra de los planteamientos de Searle. Como ya hemos dicho, fuera del estrecho mundo académico filosófico, todas estas críticas han quedado generalmente silenciadas por la gran popularidad que ha llegado a alcanzar la posición de Searle.

diferencia ontológica sin cometer una petición de principio? Dicho de otro modo, ¿cómo distinguir el rechazo del «rechazo» sin utilizar, en algún momento, esa misma distinción? Éste es el problema fundamental.

En definitiva, la existencia de ese rechazo (o «rechazo») no puede ser decisoria. El criterio que Searle pone siempre en juego apela a un "nosotros" que acaba siendo sólo un «sí mismo» (y aquí las comillas significan ya otra cosa). Pero, entonces, el diagnóstico debe estar ya claro. Pues un criterio de uso exclusivamente privado, ¡no sirve como criterio!

Y aún hay más. No sólo no tiene valor decisorio la existencia de ese rechazo (o «rechazo») sino que, hasta cierto punto, también puede ser irrelevante. Tal vez esa «ilusión de la diferencia» únicamente se base en el hecho de que ni esas máquinas, ni nosotros mismos, tendríamos ningún conocimiento directo e inmediato de la vida mental del segundo sistema integrado. Veamos. Si las primeras máquinas tienen vida mental, el segundo sistema también debe poder tenerla. La razón sería que el segundo sistema puede tener todo lo que tiene cada una de las primeras máquinas. Nada en nuestra descripción lo impide. Es cierto que algunas partes del segundo sistema también tendrían una vida mental propia, pues algunos de sus componentes son justamente máquinas del primer tipo. Pero, ¿puede significar esto algún obstáculo? Si quitáramos esas vidas mentales y las sustituyéramos por otros dispositivos apropiados, obtendríamos una nueva máquina muy semejante a las primeras. ¿Por qué, entonces, proporcionar vida mental a algunas partes de un sistema, a algunos de sus componentes, va a restar vida mental a tal sistema? La vida mental de algunas de las partes del sistema integrado no jugaría en todo esto ningún papel relevante. Y de ser así las cosas, la «ilusión de una diferencia ontológica última» surgiría, en el fondo, del simple hecho de ser sujetos diferentes.

Podemos dejar simplemente las cosas en este punto, y ya habríamos puesto en cuestión el carácter conclusivo del experimento mental de Searle. Pero también podemos continuar sugiriendo algunos pequeños cambios. Algunos pequeños cambios en nuestras creencias (¿o acaso "creencias"?), y en las creencias" (¿o "creencias"?) de las primeras máquinas, podrían dar lugar al reconocimiento (¿o "reconocimiento"?) de que el segundo sistema integrado sí entiende efectivamente chino. Y también al reconocimiento (¿o "reconocimiento"?) de que así es como se entiende chino. Y esos pequeños cambios en unas y otras creencias (¿o "creencias"?) no serían ni más ni menos que ¡la aceptación de los supuestos de la Inteligencia Artificial en su sentido más fuerte!

De todas formas, seguramente la inteligencia artificial en su sentido fuerte esté desencaminada. Sus supuestos seguramente sean erróneos. Los materiales concretos que ejemplifiquen las estructuras computacionales perti-

nentes pueden ser efectivamente muy importantes. Tal vez tanto los materiales concretos como las estructuras computacionales sólo sean, por separado, condiciones necesarias pero no suficientes para la existencia de fenómenos mentales. O tal vez ni tan siquiera sean condiciones necesarias. Tal vez sólo puedan ofrecer, de manera conjunta o combinada, condiciones suficientes para lo mental. Todo esto es cierto. Pero una cosa es esto y otra el argumento de Searle que estamos analizando. Y el argumento de Searle no es capaz de mostrar el despropósito de la inteligencia artificial en sentido fuerte con la supuesta rotundidad con la que muchas veces ha sido presentado.

Hemos intentado mostrar que la segunda interpretación de nuestro peculiar juego de imitación, en el que se generaba la ilusión de una diferencia ontológica, no es de entrada inaceptable. Sin embargo, seguimos teniendo dos interpretaciones posibles. ¿Cuál de ellas sería la más adecuada? Searle, por supuesto, se inclinaría por la primera. Ha de existir siempre una diferencia ontológica fundamental y última entre simular algo y duplicarlo. Tengamos o no acceso a esta diferencia, Searle sostiene que se trata de una diferencia ontológica básica.

6. Duplicar y simular

No es lo mismo hablar de causas que hablar de poderes causales. Los planteamientos de Searle a la hora de afianzar sus posiciones en relación a la interpretación correcta del experimento mental de "La habitación china" (es decir, la primera interpretación) son muy ambiguos en este punto. Y cuando se analiza esta ambigüedad, podemos descubrir que algunos de los argumentos más repetidos de Searle son o bien rechazables, o bien puramente tautológicos. A continuación, nos ocuparemos de esta cuestión. Y también del peligro de cierto inadvertido chauvinismo en nuestro empeño por conocer la mente, y de la necesidad de tener en cuenta otros posibles tipos de mente.

La distinción clave en nuestro conocimiento de la mente, para Searle, se establecería entre simular y duplicar. Simular algo no es lo mismo que duplicarlo. Simular un proceso no implica repetir tal proceso. Y, paralelamente, tampoco ofrecer una simulación conduce inmediatamente a poder ofrecer una explicación. Al simular algo, no estamos de ninguna manera obligados a repetir las que pudieron haber sido sus causas reales.

Nuestras mentes son muy parecidas, en muchos importantes aspectos, a algunos mecanismos computacionales. Pero no bastaría llevar a cabo correctamente un programa de computador para tener mente. A lo más, simularíamos algo mental. No lo crearíamos. Ninguna simulación es, por sí misma, una duplicación de aquello que se simula. Una simulación de un proce-

so inteligente no crea o duplica esa inteligencia. Duplicar la inteligencia, o la mente, implica duplicar los poderes causales de aquello que típicamente tiene mente. Y en nuestro caso, esto parece entrañar una necesaria duplicación de muchos de los poderes causales de nuestros complejos sistemas nerviosos[11].

Según Searle, la intencionalidad en los seres humanos y en los animales, una intencionalidad siempre conectada a la conciencia, es un producto causal de sus sistemas nerviosos, principalmente de sus cerebros. Searle considera que esto debe ser aceptado como un hecho empírico. Como un hecho empírico acerca de las relaciones causales que se establecen entre los procesos mentales y ciertos procesos neuronales. En otras palabras, ciertos procesos neuronales deben ser aceptados como condiciones causales suficientes de la intencionalidad[12].

Para Searle, la ejecución de un programa de computador, lo complejo que se quiera, no puede ser suficiente para conseguir la intencionalidad deseada. Un ser humano instalado en el escenario ofrecido por "La habitación china" no tiene, simplemente por ello, los estados mentales requeridos para hablar chino. Y una consecuencia inmediata de esto, según Searle, es que cualquier mecanismo capaz de producir intencionalidad, una intencionalidad como la nuestra, debe tener unos poderes causales iguales a esos poderes causales del cerebro que en nosotros son responsables de la producción de la intencionalidad. Somos organismos cuya composición y estructura biológica es decisiva y causalmente responsable de su vida mental. Tal vez otros procesos físico-químicos tengan similares capacidades causales. Esto vuelve a ser una cuestión empírica. Tal vez los "marcianos" estén hechos de un material diferente y ello no les impida pensar y sentir como nosotros. Lo importante es que sólo tendrán una vida mental como la nuestra, según Searle, aquellos seres cuyos cuerpos tengan los mismos poderes causales que en nosotros son responsables de nuestra vida mental.

Lo psicológicamente relevante, pues, de las operaciones del cerebro no son las características formales de los procesos neuronales, sino sus propiedades causales. Las manipulaciones formales de símbolos, por sí mismas, no generan ninguna intencionalidad. Son exactamente iguales a las manipulaciones de símbolos sin ningún significado. Son procesos que únicamente tienen una sintaxis, no una semántica. Y esto que decimos de los símbolos se aplicaría también a cualquier otra estructura computacional o, en general, formal. La intencionalidad que parecen tener algunos compu-

[11] Véase, sobre todo, Searle (1980, 84, 85, 87 y 90).
[12] Searle (1984 y 85). Véase también D. Pérez Chico (1999).

tadores y algunos robots es una intencionalidad prestada, una intencionalidad que sólo existe realmente "en la mente del espectador".

Repitámoslo una vez más. Según Searle, la intencionalidad existe gracias a los poderes causales de nuestros cerebros. Las propiedades computacionales (funcionales, formales, sintácticas, etc.) no son suficientes para la intencionalidad. Esas propiedades no tienen, por sí mismas, más poderes causales que los poderes causales que se derivarían de su capacidad para producir nuevos formalismos una vez que son ejemplificadas por ciertos estados de cosas. A la pregunta ¿podría pensar una máquina?, respondería Searle que obviamente sí, nosotros mismos somos máquinas de este tipo, máquinas capaces de pensar. Ahora bien, ¿podría pensar una máquina construida por nosotros? La respuesta de Searle a esta nueva pregunta sería que a menos que esa máquina consiga duplicar exactamente los poderes causales que nos hacen a nosotros pensar, no se duplicarán los efectos, no se producirán auténticos pensamientos como los nuestros.

7. Malas respuestas y respuestas tautológicas

Pero una vez llegados a este punto, es conveniente aclarar algo. Algo que puede llegar a ser tremendamente importante. Searle suele hablar de "poderes causales", y de "causas" también en el sentido de poderes causales. Nosotros, asimismo lo hemos hecho. Pero una cosa son los poderes causales y otra, muy distinta, las causas. Causas diferentes que produzcan un mismo efecto comparten un mismo poder causal. El poder causal de producir ese efecto. Los poderes causales clasifican las causas, que generalmente pueden ser de un tipo muy variado, por los efectos que pueden llegar a producir. A iguales posibles efectos, iguales poderes causales.

Asumiendo esta matización, Searle no podría haber ofrecido una buena respuesta a la segunda de las anteriores preguntas (¿podría pensar una máquina construida por nosotros?) en los siguientes términos: a menos que esa máquina consiga duplicar exactamente "las causas" que nos hacen a nosotros pensar, no se duplicarán los efectos, no se producirán pensamientos como los nuestros. Ésta no sería una buena respuesta porque nada impide (sobre todo, nada impide *a priori*) que causas diferentes tengan los mismos efectos relevantes. Esto es perfectamente posible. Y por lo demás muy común. Causas muy distintas pueden producir los mismos efectos.

Pero si interpretamos la respuesta de Searle como una respuesta no basada ya en las causas, sino basada en los poderes causales (o, en todo caso, basada en las causas en el sentido de poderes causales), nos encontramos con una respuesta tautológica. Nos encontramos con una respuesta demasiado vacía como para ser explicativa. Por supuesto que sólo podrán conseguirse los efectos relevantes si se comparte un mismo poder causal. Esto es así por definición. Por definición de lo que es un poder causal.

Lo que debería importarnos no es ni una clase particular de causas ni, tampoco, los poderes causales. Lo que debería importarnos son clases diferentes de causas. Lo que quisiéramos, y lo que quiere la propia Inteligencia Artificial, es descubrir si, a través de causas en principio bastante diferentes, ciertas máquinas podrían tener o no los mismos poderes causales que tienen nuestros cerebros a la hora de producir pensamientos. Esto nos permitiría identificar, por decirlo así, la auténtica naturaleza de lo mental. Ahora bien, sólo podremos descubrir esa auténtica naturaleza de lo mental comparando causas y efectos. Mejor dicho, manteniendo un mismo tipo relevante de efectos y variando las clases de causas posibles de tales efectos.

Tal vez ciertas máquinas tengan los mismos poderes causales que nuestros cerebros, por lo que se refiere a la producción de nuestros pensamientos, a través de causas muy diferentes de las que entran en juego en nuestros cerebros. Y tal vez lo común a todas esas causas sea justamente su capacidad para instanciar (ejemplificar, realizar, implementar, etc.) una cierta estructura computacional (funcional, formal, sintáctica, etc.). Seguiríamos teniendo entonces los mismos poderes causales, y los mismos pensamientos, pero distintas clases de causas. Distintas composiciones o estructuras materiales. Y esas distintas composiciones o estructuras materiales 1) quedarían agrupadas desde un punto de vista computacional (funcional, formal, sintáctico, etc.) como pertenecientes a una misma "clase natural", y 2) compartirían el poder causal de producir determinados estados mentales.

Ciertamente, el planteamiento de Searle es muy ambiguo. Si, por un lado, lo interpretamos en términos de "poderes causales", resulta indistinguible de una simple tautología. Y si, por otro lado, se reformula en términos más empíricos, en términos de "causas", se vuelve tremendamente implausible. Literalmente, se vuelve erróneo. Pero justamente de esa ambigüedad consigue toda su gran fuerza retórica. Al hablar de causas reales, parece pretender llegar al fondo de la cuestión. La forma como habla de esas causas hace que parezca que se ha conseguido llegar a ese fondo. Pero se trata de un fondo muy superficial. Porque Searle habla de causas en el sentido de poderes causales. Y así, ¡apenas se ha dicho algo más que una tautología!

Uno de los más inquietantes problemas respecto a lo mental es su integración en el orden natural del mundo. ¿Qué implicaciones tendrían las tesis de Searle en relación a este problema? ¿Qué consecuencias se seguirían respecto a la cuestión de la reducción de las propiedades mentales a algún tipo de propiedades no mentales? De una parte, hemos visto que Searle coloca lo mental en los niveles más básicos de nuestra estructura material. Y esto nos invita a considerar seriamente la posibilidad de reducir lo men-

tal a propiedades físico-químicas de nuestros sistemas nerviosos. Pero, por otro lado, encontramos dos importantes problemas implícitos en sus planteamientos, problemas que podrían frenar y hasta paralizar nuestras ansias reduccionistas.

El primero de tales problemas se refiere al tipo de propiedades causales que, en último término, puedan ser responsables de lo mental. Tal vez no sean propiedades físicas o químicas básicas. Y tal vez también, aunque lo sean, no puedan ser reducibles a otras propiedades físicas o químicas. En otras palabras, a fin de explicar la producción causal de lo mental, tal vez haya que admitir nuevas propiedades en nuestras ciencias básicas, nuevas propiedades irreducibles a cualesquiera otras[13].

El segundo problema tiene que ver con la exigencia de que para producir una vida mental como la nuestra sea necesario duplicar exactamente los poderes causales que en nuestros cerebros hacen posible nuestra vida mental.

8. Reduccionismo chauvinista

Aunque la anterior exigencia no exprese más que un simple enunciado tautológico (por definición de «poder causal», duplicar esos poderes causales ha de producir una vida mental como la nuestra, y viceversa), parece que nos acercamos ya a algo fijo que podemos tomar como punto de referencia. Podemos intentar reducir nuestra vida mental a todo aquello capaz de tener los mismos poderes causales que nuestros cerebros. Sin embargo, si hacemos esto, si optamos por esta estrategia reductiva, ¿qué ocurrirá con las posibles vidas mentales que no sean "como la nuestra"? ¿No estaríamos, en este caso, pecando de "chauvinismo"?

Si tomamos como referencia vidas mentales como la nuestra, o vidas mentales muy similares a la nuestra, parece que, de alguna forma, podemos contar con los elementos suficientes para ensayar ciertas reducciones. A través de variaciones en las posibles causas, y manteniéndose los mismos efectos relevantes, intentaríamos descubrir un cierto poder causal. Pero, en tal caso, nuestras reducciones podrán estar dejando fuera otros muchos posibles fenómenos mentales acaso muy importantes. Fenómenos mentales que, vinculados a otros posibles efectos relevantes, nuestro particular equipamiento biológico acaso no permita. Y si, por el contrario, queremos explicar y dar cuenta de todas las formas posibles de mentalidad, tendremos que depender siempre de vías de acceso a lo mental que no sean sólo

[13] Algo de esto parece sugerir, por ejemplo, Penrose (1989 y 94) . Y como es sabido, también lo sugería hace años Sellars a propósito del problema de integración de la "imagen manifiesta" en la "imagen científica".

neurológicas. Tendremos que tener en cuenta cosas como la conducta, tanto lingüística como no lingüística, la organización funcional, etc. Y, entonces, la posibilidad de una reducción de lo mental, al menos de una reducción neurofisiológica, vuelve a ponerse en cuestión.

Este dilema es inevitable y dramático en el tratamiento dado por Searle a la intencionalidad. Y se hace más dramático aún cuanto más nos empeñemos en que sólo una duplicación de los poderes causales de "un cerebro como el nuestro" conseguirá producir una intencionalidad genuina, auténtica. Aunque tengamos descripciones reductivas de todos los sistemas capaces de duplicar esos poderes causales, seguiremos anclados en un tipo muy específico y particular de intencionalidad.

Como vemos, a menos que nos empeñemos en dar a la noción de "poderes causales capaces de generar una mente" un sentido completamente vacío de contenido y puramente tautológico, la distinción entre duplicar realmente una vida mental y meramente simularla solo podremos trazarla desde dentro de nuestras concepciones, teorías, interpretaciones, atribuciones, etc. (... ¡y simulaciones!) acerca de lo que puede ser una auténtica vida mental. Y en este terreno, con todos sus riesgos y oportunidades, el computacionalismo sigue siendo un jugador más. ¡Sigue siendo un jugador más a pesar de los argumentos de Searle basados en su experimento mental de "La habitación china"!

Podemos acabar preguntándonos, ¿en qué pueden consistir, con más detalles, esas relaciones causales que se establecerían entre nuestros cerebros y nuestros pensamientos? ¿Cuánto se necesitaría duplicar o repetir de nuestros cerebros a fin de conseguir reproducir nuestra vida mental? Como ya hemos dicho, descubrimos los poderes causales de las cosas discriminando y examinando sus causas y efectos. Y duplicando exactamente las causas que nos hacen a nosotros pensar, sin duda se duplicarían sus efectos, se conseguirían pensamientos como los nuestros. Todo esto parece indudable. Pero, ¿no podrían conseguirse los mismos efectos relevantes a través de causas diferentes (lo cual, por otra parte, es de lo más habitual en cualquier otro contexto)? Más aún, ¿no son posibles otros "efectos relevantes"? En otras palabras, ¿no son posibles otras formas de intencionalidad, de vida mental, diferentes de la nuestra? ¿Cómo descubrir o rechazar tales posibilidades? Estas son algunas de las preguntas importantes que Searle no responde. Y son justamente el tipo de preguntas que nosotros debemos seguir intentando responder.

9. En algún lugar entre Searle, Fodor y Dennett

Las críticas de Searle al computacionalismo afectan muy directamente al tipo de computacionalismo defendido por autores como Jerry Fodor[14]. En nuestro trabajo hemos querido mostrar que los argumentos de Searle no son concluyentes. O al menos, que no son tan concluyentes como a primera vista pueda parecer. Pero esto no significa una reivindicación directa del computacionalismo abanderado por Fodor. Es difícil negar que Searle puede tener razón en muchos de sus reparos a la metáfora de la mente como un computador. Como poco, sería una metáfora bastante desorientadora. Y esto no es todo. De hecho, las ciencias cognitivas en los últimos años se han ido poco a poco orientando hacia otras formas de computacionalismo muy distintas del computacionalismo clásico defendido por Fodor (por ejemplo, hacia el "conexionismo", o hacia planteamientos "híbridos") y hacia perspectivas mucho más centradas en la neurología y neurofisiología. Todo ello también parece estar dando la razón a Searle. La mente puede estar más cerca del "hardware" que del "software".

Sin embargo, hay otro autor que nos debe hacer tomar con mucha precaución afirmaciones como la que acabamos de hacer. Ese autor es Daniel Dennett. Frente a la tenaz defensa que hace Searle de la existencia de intencionalidades originarias, y nuestros cerebros serían los lugares privilegiados en los que se producen tales intencionalidades, Dennett rechaza la misma idea de una intencionalidad originaria. Para Dennett, toda intencionalidad es siempre derivada. No existirían intencionalidades originarias que sean la fuente primigenia de los fenómenos mentales. Por las mismas razones que consideramos como derivada la intencionalidad de un texto o de un artefacto, derivada respecto a la intencionalidad que suponemos presente en nuestras mentes, deberíamos considerar como derivada la intencionalidad de nuestras mentes respecto a las interpretaciones intencionales, o teleológicas, que podemos hacer de la evolución biológica. Somos un producto de la evolución biológica en el mismo sentido en el que un texto o un artefacto son productos nuestros. Simplemente, podemos interpretarnos a nosotros mismos también como "productos". Y por ello toda intencionalidad no puede ser sino derivada[15].

[14] Una discusión sumamente interesante entre Searle y Fodor se recogía en *The Behavioral and Brain Sciences*, 3 (septiembre, 1980). Véase Searle (1980a y 1980b) y Fodor (1980). Véase también cómo continúa la polémica en Searle (1991) y Fodor (1991).

[15] La opinión de Dennett se ha mantenido constante en este punto desde Dennett (1971). Más recientemente, podemos encontrar nuevos argumentos en Dennett (1990), comparando las atribuciones de mente con la interpretación de textos y artefactos, en Dennett (1991), a propósito de diversos problemas involucrando la conciencia, y en Dennett (1995), en relación a la evolución biológica. En Dennett (1996:50-5), dirigiéndose explícitamente contra

Fodor es tremendamente exclusivista en sus posiciones a favor del computacionalismo clásico. Para él, la mente sólo puede ser un programa informático. Dennett es instrumentalista y ficcionalista respecto a la existencia de la mente. Lo que más bien hay, para Dennett, son atribuciones de mente desde una cierta perspectiva intencional. Y según veíamos, Searle rechaza todo computacionalismo a la vez que se aferra a un realismo extremo en relación a un cierto tipo de mente. Seguramente la verdad esté en algún lugar entre Searle, Fodor y Dennett. En cualquier caso, creo que deberíamos estar mucho más abiertos que Searle respecto a las diferentes posibilidades que pueden existir para lo mental. Por otro lado, deberíamos ser mucho menos exclusivistas que Fodor respecto a las capacidades explicativas del computacionalismo. Y también deberíamos ser mucho más realistas que Dennett.

Si hay algo que no podemos perder de vista a la hora de entender la mente es que lo que estamos intentando entender es la mente, no nuestras atribuciones de propiedades mentales. Pues, después de todo, "atribuir" propiedades mentales es algo que sólo es posible si existen mentes. Atribuir, interpretar, usar criterios, adoptar perspectivas, etc., son cosas que sólo pueden hacer las mentes. Algunas clases de mentes.

Referencias bibliográficas

Churchland, P. y Churchland, P. (1990): "¿Podría pensar una máquina?", *Investigación y Ciencia*, marzo.

Dennett, D. (1971): "Intentional System", *Journal of Philosophy*, 68, 4º.

———— (1990): "The Interpretation of Texts, People and Other Artifacts", *Philosophy and Phenomenological Research*, vol. L (Supplement, Fall).

———— (1991): *Consciousness Explained*, Boston, Little Brown.

———— (1995): *Darwin's Dangerous Idea*, Londres, Penguin Books.

———— (1996): *Kinds of Minds*, Nueva York, Basic Books.

Penrose, R. (1989): *The Emperor's New Mind*, Oxford, Oxford Univ. Press.

———— (1994): *Shadows of The Mind. A Search for the Missing Science of Consciousness*, Oxford, Oxford Univ. Press.

Fodor, J. (1980): "Searle on What Only Brains Can Do", *The Behavioral and Brain Sciences*, 3.

———— (1991): "Yin and Yang in the Chinesse Room", en D. Rosenthal (ed.) [1991].

Pérez Chico, D. (1999): "Naturalismo biológico y el problema del mundo externo", *Teorema* xviii/1.

Searle, se condensa el tipo de rechazo a la idea de una intencionalidad originaria que hemos señalado.

Rosenthal, D. (ed.) (1991): *The Nature of Mind*, Oxford, Oxford Univ. Press.
Searle, J. (1980a): "Minds, Brains, and Programs", *The Behavioral and Brain Sciences*, 3.
———— (1980b): "Author's Response", *The Behavioral and Brain Sciences*, 3.
———— (1984): "Intentionality and its place in nature", *Synthese*, 61.
———— (1985): *Minds, Brains and Science. The 1984 Reith Lectures*, Cambridge, Harvard Univ. Press [*Mentes, Cerebros y Ciencia*, Madrid, Cátedra, 1985].
———— (1987): "Minds and Brains without Programs", en Blakemore, C. & S. Greenfield (eds.) [1987] *Mindwaves: Thoughts on Intelligence, Identity and Consciousness*, Nueva York, Basil Blackwell.
———— (1990): "¿Es la mente un programa informático?", *Investigación y Ciencia*, Marzo.
———— (1991): "Yin and Yang Strike Out", en D. Rosenthal (ed.) [1991].
———— (1992): *The Rediscovery of The Mind*, Cambridge, MIT Press.
———— (1997): *The Mystery of Consciousness*, Nueva York, New York Review of Books [*El misterio de la conciencia*, Barcelona, Paidós, 2000].
———— (1998): *Mind, Language and Society. Philosophy in the Real World*, Nueva York, Basic Books [*Mente, Lenguaje y Sociedad. La filosofía en el mundo real*, Madrid, Alianza, 2001].
Turing, A. (1950): "Computing Machinery and Intelligence", *Mind*, vol. 59, nº 236.