DEBATE SOBRE LAS ANTROPOLOGÍAS. THÉMATA. Núm. 35, 2005, págs. 267-275.

Cómo salir de la Habitación China: Conciencia e intencionalidad en las otras mentes

Jesús Navarro. Universidad de Sevilla

Esta comunicación tiene dos intenciones aparentemente opuestas: en la primera parte sostengo que el argumento de la habitación china, si se aceptan sus premisas, es capaz de rebatir cualquier modelo de inteligencia artificial fuerte que se proponga, incluidas las redes neurales que desarrollan su propio lenguaje en procesos prolongados de aprendizaje. En la segunda parte, cuestiono la conveniencia de aceptar esas mismas premisas: en caso de asumir que la perspectiva de primera persona es infalible, y esta es una condición indispensable del argumento de Searle, el problema de las *otras mentes* se hace insoluble, abocándonos al solipsismo.

1. Sintaxis, semántica y procesos de aprendizaje

Las pretensiones de los ordenadores de llegar a pensar como nosotros se vieron truncadas hace unos 25 años, cuando a John Searle se le ocurrió diseñar mentalmente una habitación un tanto peculiar, que llamó "La Habitación China"[1]. El objetivo del experimento mental era demostrar que la mera manipulación de símbolos conforme a reglas formales preestablecidas no es suficiente para desarrollar una inteligencia artificial en sentido fuerte, es decir, para crear una *mente* similar la mente humana, con el mismo tipo de conciencia e intencionalidad. Los requisitos del experimento eran los siguientes:

a) Tomamos a un sujeto que no conozca en absoluto el idioma chino, por ejemplo, el propio Searle. b) Introduzcámoslo en una habitación que disponga de dos orificios: uno de entrada y otro de salida, por los que atravesarán textos escritos en chino. c) Dotemos la habitación de un manual de instrucciones, redactado en inglés (idioma conocido por el sujeto), que permita, dada una cadena cualquiera de caracteres chinos, mediante la aplicación de reglas puramente sintácticas (es decir, basadas exclusivamente en la forma de los caracteres y sus posiciones relativas), ofrecer una respuesta perfectamente coherente, similar a la que cabría esperar de cualquier hablante chino.

Es difícil imaginar que las reglas del manual de instrucciones puedan ser puramente formales y sintácticas, pero ahí reside la fuerza del argumento: Searle, dentro de la habitación, sólo manipula signos, sin conocer en absoluto lo que esos signos significan, es decir, sin saber cuál es el referente en el mundo exterior de las palabras chinas que utiliza. Si el libro de instrucciones fuera suficientemente bueno, y Searle es lo suficientemente hábil, cualquier interlocutor chino fuera de la habitación llegaría a estar convencido de que el sujeto del interior ha comprendido el texto que se le ha dado, y que

[1] Searle ha publicado el argumento en diversos lugares. La primera aparición tuvo lugar en "Minds, Brains and Programs". En *Behavioral and Brain Sciences* 3 (3), pp. 417-57. Otras versiones: *Minds, Brains and Science*. Londres: British Broadcasting Corp., 1984; *The Rediscovery of Mind*. Cambridge, Mass: MIT Press, 1992; "Twenty-One Years in the Chinese Room". En J. Preston y M. Bishop (eds.). *Views into the Chinese Room. New Essays on Searle and Artificial Intelligence*. Oxford: Clarendon Press, 2002. pp. 51-69.

está manteniendo una verdadera conversación en chino, entendiendo lo que se le dice. Pero, por muy convencido que esté el interlocutor chino, Searle puede asegurar que no tiene ni idea de a qué se refieren los caracteres que utiliza. No entiende absolutamente nada: sólo maneja signos.

El experimento mental tiene interés para la discusión acerca de la inteligencia artificial porque los ordenadores, por principio, sólo son capaces de hacer eso: manejar signos con criterios sintácticos. Pero la semántica, dice Searle, no puede ser reproducida en un sistema computacional, basado en principios exclusivamente sintácticos. Y, sin una semántica, los signos carecen de referente y, en sentido estricto, ni siquiera puede decirse que sean verdaderos signos[2]. Los ordenadores, por muy potentes y veloces que puedan llegar a ser, nunca serán capaces de salir de sus particulares habitaciones chinas: nunca podrán comprender que los signos que utilizan son *intencionales*, es decir, que apuntan hacia algo distinto de ellos mismos.

Para bien o para mal, los ordenadores cuentan en el mundo académico con buenos abogados, los llamados *funcionalistas*, que aceptaron defenderlos contra la denuncia por necedad presentada por Searle. Hasta entonces había bastado una argucia elaborada por Alan Turing[3], según el cuál, básicamente, si un ordenador puede engañarte por completo y hacerte creer que entiende lo que le dices, entonces no tendrás más remedio que aceptar que en sus chips, gracias al programa informático adecuado, se ha creado una mente. Si la imitación es suficientemente perfecta, y llega a hacerse indiscernible del original, habríamos de aceptar que se trata de *otro original*; lo contrario sería puro chovinismo antropoide. Pero, por culpa de Searle, el test de Turing había dejado de ser suficiente: puede que el ordenador logre engañarnos por completo, pero no por eso habrá entendido nada. Era preciso diseñar nuevas estrategias para la defensa.

El buffet funcionalista se puso manos a la obra y contraatacó a principios de los 80 con una serie de argumentos que pretendían abrir puertas y ventanas en la habitación, mientras que Searle modificaba su posición hábilmente para devolver los golpes. Así fueron apareciendo distintas versiones del argumento, cada vez más enrevesadas: Searle llegaba a memorizar el libro para evitar su presencia en la habitación; en unas ocasiones, las instrucciones dejaban de referirse a los ideogramas chinos y reproducían el funcionamiento de un cerebro chino, neurona a neurona; en otras, llegaba a tomar forma física dentro de un enorme robot que realizaba en el mundo acciones correlacionadas con la conversación mantenida. En ningún caso, sostiene Searle, él, que está dentro de la habitación, ni el sistema completo al que pertenece, llega a comprender nada de lo que hace o dice. No hay forma de escapar de la habitación desarrollando programas de software, basados en mera computación; por ese camino no hay modo de saltar de la sintaxis a la semántica, de la ciega manipulación de signos a la genuina comprensión humana. Para crear una inteligencia artificial fuerte no basta con aumentar la complejidad de los programas y de los procesadores: añadir más de lo mismo, dice Searle, no va a solucionar el problema.

Para que haya auténtica intencionalidad, es decir, verdadera referencia del lenguaje al mundo, es preciso un cerebro humano, o algún artefacto que sea capaz de reprodu-

---

[2] Por eso, en sus últimas versiones del argumento, Searle llega a negar que el ordenador tenga siquiera una sintaxis: eso implicaría que está utilizando símbolos que, para ser considerados como tales, han de tener referente. En el ordenador, si lo consideramos independientemente de cualquier observador externo, no se estarían procesando símbolos, sino que sólo se estarían llevando a cabo ciegos procesos eléctricos.

[3] "Computing Machinery and Intelligence". En *Mind* 59 (1950). pp. 433-60.

cir complejas habilidades como son ser *consciente* o tener estados *intencionales*. Ninguna réplica digitalizada, basada en la computación sintáctica, puede llegar a reproducir esas habilidades. Cómo lo hace un cerebro humano, es algo que todavía no sabemos. Pero Searle se muestra confiado en que el misterio se resolverá más pronto que tarde. Sólo una vez que hayamos averiguado cómo lo hace el cerebro humano, seremos capaces de diseñar máquinas capaces de adquirir una semántica: pero para eso sería precisa una tecnología completamente distinta de la actual, algo que todavía no podemos ni imaginar.

Un solo argumento no basta para dar al traste con un gran proyecto de investigación interdisciplinar. Por eso, viendo la ineficacia del buffet funcionalista, los ordenadores decidieron contratar los servicios de otros abogados: los llamados *conectivistas*, que hicieron resurgir una cierta esperanza en el lado de la defensa[4]. Tal vez sea posible hacer evolucionar la inteligencia artificial de un modo cualitativamente distinto, que no consista en ofrecer más de lo mismo, aumentando la complejidad de los procesos computacionales. La nueva esperanza tiene un nombre: *redes neurales* (*Neural Networks*). A diferencia de los programas de software convencionales, las redes neurales carecen de un código rígido predefinido. Por el contrario, van generando su propio software a medida que interactúan con su entorno, mediante un proceso de aprendizaje prolongado en el tiempo. Esta posibilidad parece marcar una diferencia cualitativa con los modelos de inteligencia artificial atacados por Searle a principios de los 80. A lo largo de ese proceso de aprendizaje con el medio, el ordenador sería capaz de desarrollar su propia semántica, y no limitarse a la aplicación de instrucciones formales preestablecidas. La intencionalidad estaría garantizada al no tratarse de un lenguaje impuesto sin referencia al exterior, sino de un lenguaje generado a través de la relación entre la máquina y el entorno circundante. Este tipo de máquinas son mucho más flexibles que las anteriores, y se adaptan a las circunstancias cambiantes del medio de un modo que para el ordenador convencional era imposible. Parece ciertamente que aprendan por ellas mismas, que comprendan por ellas mismas las relaciones que se establecen entre las cosas.

Una de las ideas que pretendo sostener aquí es que, a pesar del entusiasmo que pueda despertar este proyecto, el argumento de la habitación china puede ser modificado de forma que atrape en sus redes no sólo a los ordenadores convencionales, anclados en sus rígidos softwares, sino también a las flamantes redes neurales recién estrenadas. Para mostrar esto quisiera que imagináramos una nueva versión de la habitación china, ligeramente más sofisticada. Supongamos que el libro de instrucciones entregado a Searle era perfectamente correcto cuando él entró en la habitación: en aquel momento, era posible simular a la perfección la conversación de cualquier hablante chino nativo. Pero seamos lo suficientemente desconsiderados como para mantener a Searle encerrado en la habitación durante un largo periodo de tiempo, años incluso. ¡Al fin y al cabo entró allí por su propia voluntad! Ahora supongamos que el chino, como es el caso de cualquier idioma vivo, va evolucionando con el paso del tiempo, de modo que, poco a poco, el manual de instrucciones con el que Searle entró en la habitación ha ido quedándose obsoleto. Hay ocasiones en las que sus respuestas

[4] Para una exposición actualizada, cfr. H. A. Simon y S. A. Eisenstadt. "A Chinese Room That Understands" y I. Alexander. "Neural Depictions of 'World' and 'Self': Bringing Computational Understandign to the Chinese Room", ambos en J. Preston y M. Bishop (eds). *Views into the Chinese Room*.

ya no son válidas, y sus interlocutores chinos comienzan a desconfiar de la capacidad de comprensión del inquilino de la habitación.

En ese momento entramos en juego nosotros, los encargados de mantener el buen funcionamiento del invento. Viendo la perplejidad de los interlocutores, decidimos introducir una pequeña modificación en el diseño: cada vez que Searle responda utilizando una regla que ha quedado obsoleta, procederemos a aplicarle una pequeña descarga eléctrica. En ese momento pensará sin duda que ha debido realizar alguna operación equivocada y, tras repasar su manual, comprobará que no es así: que ha aplicado las reglas correctamente. Es lógico pensar que, ante el sufrimiento repetido por las descargas eléctricas, Searle supondrá que debe haber algo en el mundo exterior que ha cambiado, haciendo que determinada regla sea ya inadecuada. Entonces es probable que identifique, por el momento en que recibe la descarga, de qué regla se trata, y que vaya probando por ensayo y error otras alternativas. Imaginemos que acierta, y encuentra una nueva regla cuya aplicación no implica descarga eléctrica. Con ayuda de un bote de typex y un bolígrafo (que no es mucho pedir como material extra para la habitación), Searle corrige el libro para adaptarlo a la nueva situación. Habría transformado el "software" original en función del medio cambiante con el que se relaciona.

Con el tiempo podemos imaginar que, poco a poco, signo a signo, el manual de instrucciones acaba siendo totalmente distinto de aquel original que le fue entregado años atrás. El código de software habría dejado de ser rígido: Searle habría imitado el proceso de aprendizaje de una red neural, transformando su programa mediante una prolongada interacción con el medio. Pero, desgraciadamente, lo más probable es que siga sin comprender lo más mínimo acerca de a qué se refieren los términos. Puede que haya adquirido con el paso de los años esa inmensa paciencia que caracteriza a los artesanos chinos, pero no habrá adquirido en absoluto una semántica, y seguirá indefinidamente sin saber a qué se refieren los ideogramas que maneja.

La introducción en el experimento de un sencillo proceso de *feed-back* aversivo, puede transformar la habitación china en un argumento mucho más potente, pues por este camino ¡podemos incluso llegar a concebirlo sin necesidad de ningún libro de instrucciones original! Searle podría entrar en la habitación sin saber en absoluto cuál es la tarea que tiene que realizar. El lugar del libro de instrucciones podría ser ocupado por un largo y paulatino proceso de aprendizaje. En primer lugar, se introducirían signos muy sencillos, cuyo correlato sería enormemente fácil de encontrar. Un saludo, por ejemplo. Por ensayo y error, Searle sería capaz de encontrar el signo que ha de devolver, a riesgo de acabar chamuscado por las descargas. Poco a poco, podríamos ir aumentando la complejidad hasta que, finalmente, Searle llegara a construir por sí mismo un completo y perfecto libro de instrucciones. Sólo entonces estaría listo para enfrentarse a conversaciones cotidianas con sujetos chinos, superando con éxito el test de Turing. Esta nueva habitación no es mucho más difícil de imaginar que la primera, y sí es en cambio más potente, pues no depende tanto de la intencionalidad del programador, que siempre había sido la deficiencia del modelo original.

Parece por lo tanto que el argumento de la habitación china puede ser adaptado para hacer frente a las redes neurales que son, como dijimos antes, la nueva esperanza de los ordenadores para desarrollar una inteligencia artificial en sentido fuerte. Dado que los procesos de aprendizaje siguen desarrollándose según criterios sintácticos, no ofrecen un contacto con el mundo más auténtico que los programas convencionales.

Nada les permite dar el salto de la sintaxis a la semántica: seguirán indefinidamente sin saber a qué se refieren sus signos.

Viendo que la balanza no se inclina a su favor, no son pocos los conectivistas que han decidido abandonar la propia idea de un sistema simbólico representativo, formando un tercer buffet de defensa de los ordenadores: los llamados *eliminativistas*, que en cierto modo constituyen un subgrupo dentro de los conectivistas. Los eliminativistas sostienen que, en realidad, las redes neurales no pretenden ser estructuras simbólicas, sino que sus procesos son sencillamente una prolongación, en términos de información no simbólica, de las relaciones con el medio. De hecho, no hay nada en la estructura de una red neural que aspire a ser un *símbolo*, supuesta representación de algo exterior a la propia red. ¿Están tirando la toalla? ¿Al negar que las máquinas desarrollen procesos simbólicos, no están actuando en contra de sus defendidos, aceptando que los ordenadores son incapaces de hacer lo que hacemos nosotros los humanos? No, porque la idea de los eliminativistas es que, en el fondo, el tipo de relación que desarrollan los cerebros humanos con el medio tampoco es simbólica ni intencional. La idea de que en nuestras cabezas pululan símbolos que misteriosamente apuntan hacia algo distinto de ellos mismos sería un rancio prejuicio, heredado de la psicología popular, que ha de ser abandonado[5]. ¿Que las redes neurales no son capaces de desarrollar una semántica? No importa: en realidad tampoco los seres humanos utilizan símbolos, ni tienen procesos intencionales, ni creencias, ni deseos, ... Todos estos son conceptos obsoletos, como en tiempos lo fueron el éter o el flogisto, y la ciencia pronto demostrará que carecen de referente en la realidad. No se trata de si las redes neurales son capaces de hacer lo que hacemos nosotros; la cuestión, para los eliminativistas, es que nosotros mismos no somos capaces de hacer algo distinto de lo que hacen las redes neurales.

Parece que el alegato de la defensa llega aquí demasiado lejos y que cualquier juez prudente lo desestimaría por inverosímil: la cuestión de si los seres humanos son capaces de manejar símbolos intencionales o de detentar creencias y deseos está fuera de discusión. Parece por lo tanto que el argumento de la habitación china es inexpugnable, y que el único modo de dejarlo de lado pasaría por renunciar a explicar aquello que caracteriza más propiamente nuestra vida mental: la conciencia y la intencionalidad.

2. *Criterios de atribución de conciencia intencional: el problema de* las otras mentes

El argumento de la habitación china demuestra que los ordenadores, tal y como los estamos diseñando, no pueden pensar como nosotros. Pero nuestro mayor problema no es ese: la cuestión que quiero plantear ahora es que, tal y como está diseñado el experimento, excluye también al resto de los humanos de la categoría de sujetos mentales intencionales. El criterio de asignación de mentalidad es tan rígido que sólo lo puede superar uno mismo.

Analicemos el argumento de Searle desde otra perspectiva: para que se sostenga es preciso que utilicemos dos lenguajes distintos: el inglés y el chino. Las instrucciones que recibe Searle están en inglés; la historia que se le ofrece a través de la rendija de la habitación, así como las preguntas que se le formulan con respecto a ella, están en

[5] Cfr. P. M. Churchland. "Eliminative Materialism and the Propositional Attitudes". En *Journal of Philosophy* 78 (1981). pp. 67-90.

chino. Supongamos que nosotros somos capaces de manejar ambos idiomas, y que mantenemos una conversación diferente en cada uno de ellos. En primer lugar, mantenemos una conversación por escrito, en chino, a través de las ranuras de la habitación. Como no queremos establecer ningún prejuicio acerca de la identidad del sujeto con el que mantenemos esta primera conversación, lo llamaremos "X". A continuación, mantenemos una segunda conversación, esta vez oral y en inglés, con Searle en persona. Hay por lo tanto dos conversaciones: una con X, por escrito, en chino y otra con Searle, oralmente, en inglés[6]. Preguntemos a ambos sujetos si comprenden el chino; evidentemente, surgirá una contradicción:

> X (en chino): ¿Qué si hablo chino? ¿Me está usted tomando el pelo? ¿Y qué cree que hemos estado haciendo todo este tiempo?
>
> Searle (en inglés): No entiendo ni una sola palabra en chino. Para mí la escritura china no es más que un montón de garabatos sin sentido.

Todo el peso del argumento reside en que, según Searle, debemos concederle mayor credibilidad a su discurso en inglés que al emitido por X en chino. Pero, ¿por qué habríamos de hacerle caso? La cuestión fundamental es: ¿cuál es *el criterio* con el que podemos otorgar una mayor credibilidad al discurso de uno sobre el del otro? Consideremos los distintos candidatos:

a) *El lenguaje.* Opción descartada, pues en el uso de esa habilidad X y Searle están empatados por principio: gracias al libro de instrucciones, X puede asegurarnos que comprende la historia, indignarse ante nuestra desconfianza, jurar exaltado que sabe perfectamente de lo que está hablando, como lo haría el propio Searle si dudáramos acerca de su capacidad de comprensión de la lengua inglesa.

b) *La conducta.* Searle puede demostrar que sabe lo que significa el término "hamburguesa" señalando una hamburguesa, pero X no puede señalar nada: sus posibilidades de relación con el mundo están limitadas a la producción de lenguaje. Esta opción es fácilmente descartable mediante la llamada "respuesta del robot" (*The Robot Replay*). Se trataría de dar un poco más de trabajo a Searle mientras está en la habitación: además de producir ideogramas chinos, tendría el papel de generar instrucciones que dirigirían el movimiento de un robot. Por supuesto, él no sabría lo que significan las instrucciones que genera pues, como los caracteres que escribe, sólo serían consecuencia de la aplicación de reglas formales sintácticas. Al preguntarle a X si sabe lo que significa "hamburguesa" en chino (como quiera que eso se escriba), emitiría el texto "Evidentemente, sé lo que es una hamburguesa: es eso", mientras que el robot señalaría una hamburguesa en su entorno. Searle sólo precisaría dos libros de instrucciones: uno para manejar caracteres chinos y otro para manejar el robot.

c) *La apariencia física.* En el caso de la conversación inglesa con Searle estamos ante una figura humana, cuya similitud corporal con nuestro propio cuerpo nos lo hace parecer más cercano. Ante esa similitud corporal, realizamos lo que los fenomenólogos husserlianos llaman una "transferencia de sentido": si yo tengo una conciencia intencional, y tengo un cuerpo con determinadas características, es lógico pensar que este sujeto que tengo delante, con un cuerpo parecido al mío, tenga una conciencia inten-

[6] La idea de esta doble conversación aparece en el comentario de R. Wilensky ("Computers, Cognition and Philosophy"), al artículo de Searle de 1980, en el mismo número de *Behavioral and Brain Sciences*, pp. 449-50.

cional similar. Pero lo que la tesis de la inteligencia artificial fuerte está poniendo en entredicho es precisamente la necesidad de una similitud física a la hora de aplicar esa transferencia de sentido. No se trata de construir ordenadores que se parezcan físicamente a los seres humanos, sino máquinas capaces de pensar como nosotros. Por eso, en ningún caso puede la apariencia física ser el criterio mediante el cual otorgamos una mayor credibilidad a Searle que a X.

d) *El cerebro.* Aquí está, según Searle, la diferencia. La conciencia es una cualidad del cerebro, que es producida causalmente por él. Los chips de los ordenadores carecen de esa capacidad causal: imitan el resultado de la inteligencia humana, pero no son ni remotamente capaces de generar una conciencia similar a la nuestra. Y esta diferencia, según él, es *una cuestión empírica*[7]. Si se trata de una cuestión empírica, entonces no hay más que mostrar el *dato empírico* que demuestra que un cerebro es consciente, mientras que un ordenador no lo es. El problema es que ni siquiera sabemos lo que diferencia a un cerebro consciente de otro que no lo es. ¿Cuál es exactamente el mecanismo causal que produce la conciencia? Searle ha repetido en varios lugares que la subjetividad es un hecho objetivo desde el punto de vista biológico. Y acepta sin tapujos que se trata de un *hecho* objetivo que carece de ningún *dato* objetivo que lo corrobore. Pero en tal caso, ¿qué nos permite asegurar que el cerebro de Searle es capaz de comprensión? Por el momento, más allá de sus aseveraciones subjetivas, expresadas en el lenguaje, y de sus efectos en la conducta, carecemos de una prueba empírica de la existencia de la consciencia intencional. El argumento de la habitación china se construye sobre la idea de que podemos abrir la habitación y comprobar que en ella sólo hay computación sintáctica. Pero igualmente puede decirse que, ante un interlocutor humano, podemos abrir su cerebro y comprobar que sólo hay transmisiones sinápticas. Hoy en día sabemos cómo se forman las moléculas a partir de los átomos, las células a partir del ADN, los cerebros a partir de las neuronas, pero no sabemos cómo aparece la conciencia a partir de los cerebros. Sólo sabemos que, a nivel celular, hay intercambio de información: las neuronas se limitan a intercambiar ciegamente signos que no comprenden, como hace Searle en la habitación china. El paso del intercambio de información entre las neuronas a la formación de una conciencia unificada capaz de comprender el lenguaje como algo que refiere al mundo es, por el momento, un misterio.

Searle está convencido de que no se trata de un misterio, sino de un mero problema. Es decir, que la cuestión está formulada de modo que tiene una posible solución y que además esa solución llegará pronto: en el momento en que la neurociencia llegue a encontrar el dato físico del cerebro que está relacionado causalmente con la aparición de la conciencia intencional. ¿Pero tiene realmente sentido esta esperanza? Si no tenemos más datos para certificar la apariencia de conciencia intencional que los criterios objetivos que hemos indicado (lenguaje, conducta, apariencia, ...), y todos ellos pueden ser cumplidos por una simulación computacional, ¿cómo sabremos que el dato físico en cuestión está correlacionado con la conciencia intencional intrínseca, y no con la mera computación? La articulación de las perspectivas objetiva y subjetiva, de la tercera y la primera persona, parece imposible por mucho que Searle crea haberlo conseguido. Su esfuerzo por lograr esa síntesis es loable, pues se enfrenta a la cuestión en lugar de dejarla de lado, como ocurrió con el conductismo o con gran parte del fun-

[7] J. Searle. "Minds, Brains and Programs". p. 455.

cionalismo. Pero no parece en realidad que su teoría de la conciencia escape a las consecuencias de una epistemología de primera persona, pues la apelación a la experiencia subjetiva sería el único apoyo que queda para mantener el argumento de la habitación china[8]. De ahí que la única posibilidad de salvar el argumento sea despejar la incógnita X.

e) *X, en realidad, es Searle.* Según Searle, él mismo es X; es *una misma consciencia* la que sostiene las dos conversaciones. No se trataría de dos mentes enfrentadas, sino de una misma mente que en chino no sabe lo que dice y en inglés sí. Esta apelación a la experiencia de primera persona salvaría el argumento de no ser porque nadie ha sostenido nunca que Searle, en tanto que Searle, comprenda chino en este experimento. Evidentemente, no se trata de si Searle entiende el chino, sino de si hay una propiedad emergente en el sistema que permita *a* X comprender chino[9]. Es preciso darse cuenta de hasta qué punto la simulación perfecta de un interlocutor exigiría la creación de una nueva *persona*: para que el experimento tuviera éxito, X habría de poseer un pasado que recordar, un carácter propio, una manera de ser, un conjunto de opiniones generales acerca del mundo, etc., etc. De lo contrario, el interlocutor chino se daría cuenta rápidamente de que en realidad no está hablando con *nadie*. Y ninguna de esas características tiene porqué coincidir con las que presenta Searle como persona. De modo que la apelación a la primera persona como criterio no se sostiene, pues es la propia identidad entre Searle y X lo que está en cuestión.

Resumiendo: el argumento de Searle para que le creamos a él y no a X es que, según él, X no sabe a qué se refieren los signos que maneja, mientras que Searle sí posee una intencionalidad genuina gracias a la cual comprende el lenguaje que utiliza. Pero, desgraciadamente, Searle no tiene ni puede tener modo alguno de demostrarnos a nosotros la existencia de su propia conciencia intencional: ni el lenguaje, ni la conducta, ni la apariencia física ni el análisis empírico del cerebro pueden ayudarnos a decidir si hemos de hacerle caso a él o a X, dado que carecemos de un criterio objetivo para determinar en qué momento hace aparición en un sistema la conciencia intencional. Planteado así el problema, el caso de la inteligencia artificial tiene que ser sobreseído, por falta de pruebas.

En realidad, el problema que encuentra Searle a la hora de atribuir intencionalidad y conciencia a X es el mismo que podemos encontrar nosotros a la hora de atribuir al propio Searle esa intencionalidad y esa conciencia intrínsecas que dice poseer. No se trata ya de demostrar si los ordenadores tienen o no una mente intencional. La habitación china nos introduce en un atolladero mucho más preocupante: apelando a la infalibilidad de la primera persona, el argumento nos obliga a encerrarnos en el solipsismo, por mucho que Searle quiera escapar de esa posición. De este modo reaparece el gran problema que ha perseguido como una sombra a la filosofía del siglo XX cada vez que se ha ocupado de la subjetividad: el problema de las "otras mentes". Al concebir la

[8] Acerca de la dificultad de encajar la subjetividad ontológica en la objetividad epistemológica de la ciencia cfr. J. A. Guerrero del Amo. "Problemas epistemológicos subyacentes a la teoría de la mente de Searle". En *Logos. Anales del Seminario de Metafísica* 3 (2ª Época) (2001). pp. 297-316.

[9] Es la respuesta conocida como *Systems Replay*: la persona con la que mantuvimos la conversación en chino no es la misma persona que aquella con la que mantuvimos la conversación en inglés, de modo que no hay ninguna contradicción entre ambas respuestas. En inglés hemos hablado con Searle, mientras que en chino hemos hablado con la habitación como sistema, del cual Searle sólo es una parte. Cfr. D. C. Dennett. "The Milk of Human Intentionality", comentario al artículo de Searle de 1980, en el mismo número de *Behavioral and Brain Sciences*, pp. 428-30 y *Consciousness Explained*. Boston: Little, Brown & Co., 1991. pp. 435-40.

conciencia como una cualidad intrínseca, sólo accesible desde la privilegiada perspectiva de primera persona, estamos predestinados a sospechar acerca de la existencia de la conciencia ajena, sea en una máquina o en un ser humano, y siempre nos queda la posibilidad de denunciar, tras el interlocutor, los mecanismos de un autómata.

Sólo hay un modo de escapar del atolladero chino: redefinir nuestros términos fundamentales de modo que sean epistemológicamente accesibles. Sólo así podremos demostrar si en un sistema –ya sea un ordenador o un cerebro biológico– se está produciendo de hecho una genuina comprensión. Los conceptos de *conciencia e intencionalidad intrínsecas*, tal y como están definidos por Searle, conducen hacia una posición solipsista. Dado que dependen de la epistemología de la primera persona, carecemos por principio de un criterio de tercera persona para demostrar su presencia. Suponer que esa situación es meramente provisional es sólo un modo de enmascarar el problema. Por el contrario, la única salida a la habitación china pasa por una profunda reformulación de nuestros conceptos de conciencia e intencionalidad, y por una reflexión detenida acerca de qué entendemos como perspectivas de primera y tercera persona.

Jesús Navarro
Departamento de Metafísica y Corrientes
Actuales de la Filosofía, Ética y Filosofía Política
Facultad de Filosofía
Universidad de Sevilla
41018 Sevilla
jnr@us.es