Sintiencia en máquinas

Manuel de la Herrán Gascón
Primera versión: Dic. 2016
Actualizado: Ene. 2017

Este texto lo he creado a partir de los materiales que preparé para la charla que di en la Facultad de Filosofía de la Universidad de Santiago de Compostela el jueves 15 de diciembre de 2016 junto con Brian Tomasik, y que llevaba por título "Perspectivas y riesgos futuros de la consciencia artificial".

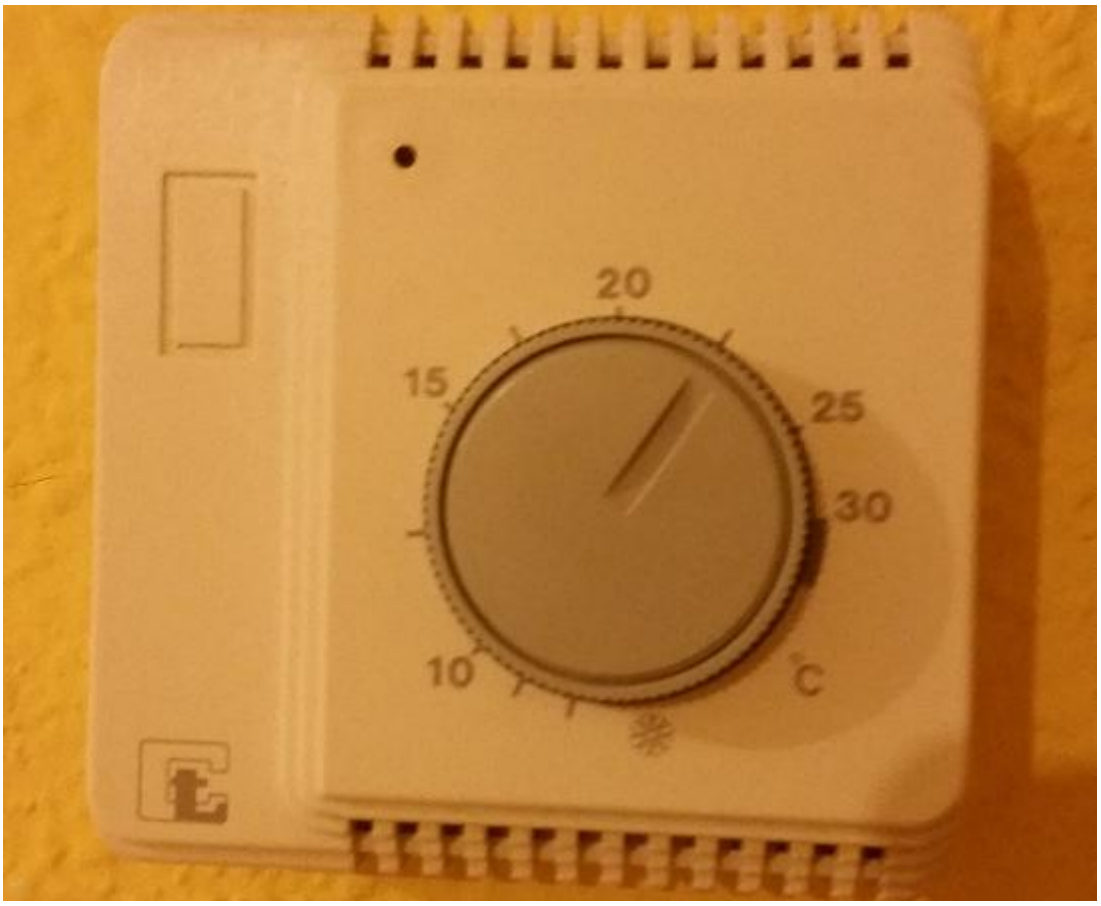


Fig.1 El termostato de mi casa

Este es el termostato de mi casa. Es la pieza "inteligente" que regula la temperatura mediante el control de una caldera. Si hace frío, enciende la caldera. Si hace demasiado calor, la apaga. Parece que tiene el objetivo de mantener la temperatura constante en cierto valor, como si no le gustara ni el excesivo frio ni el excesivo calor, igual que me ocurre a mí.

¿Estamos completamente seguros de que este termostato no es sintiente[1]?

¿Tal vez la pregunta está mal formulada?

Quizás la pregunta acerca de la sintiencia del termostato sea una pregunta mal formulada, de esas que es mejor no contestar. Si alguna vez le acusan de un asesinato que no cometió y le interrogan con preguntas de tipo Sí/No como esta:

[1] David Chalmers sugiere que incluso un termostato podría tener experiencias.

"¿Es cierto que Usted le mató por dinero?", sería mejor no contestar. La pregunta ya asume la culpa. Sí dice que sí, estaría admitiendo el crimen, y reconociendo que el móvil fue económico. Si dice que no, parece que está asumiendo la acusación implícita y que simplemente el motivo fue otro.

Tal vez la pregunta "¿Qué es lo que genera la consciencia? ¿bajo qué condiciones emerge?" está mal formulada y la consciencia no se genera, no emerge.

Tal vez el nivel de descripción de la realidad en el cual hablamos de la idea de un termostato concreto (el nivel de los objetos físicos identificables por nosotros) no sea el nivel de realidad adecuado para hacer preguntas en relación a la consciencia, siendo posible describir la realidad a otros niveles, como por ejemplo hablando de un difuso conjunto de átomos de metal y plástico, el nivel de los fenómenos cuánticos, o el universo en su conjunto. Tal vez sea en otro de estos niveles en el cual tenga sentido la pregunta acerca de la consciencia.

Este texto trata sobre la sintiencia (sentir, sufrir, disfrutar), la consciencia (percibir, ser consciente, darse cuenta), tener intereses (cosas que me benefician, cosas que me perjudican), la subjetividad (tener un yo, tener una opinión, ser alguien, tener un punto de vista), tener deseos o preferencias (querer, desear, anhelar, amar, cosas que me importan).

Tal vez cada uno de nosotros tenga una definición diferente para palabras como "consciencia" o "sintiencia", e incluso una teoría diferente acerca de su naturaleza. Sin embargo a veces parece que dos personas manejan conceptos enfrentados cuando en realidad están hablando de la misma cosa; y en otros casos usan las mismas palabras y teorías pero cuando profundizas en ello puedes ver que se están refiriendo a cosas diferentes.

Hay algo básico: *"Siento, luego existo"*[2]. Siento, luego tengo consciencia. Siento, luego soy sintiente. Siento, luego soy. Pero, ¿quiénes son los otros seres con consciencia / sintiencia? ¿Necesariamente humanos, como yo? ¿Animales? ¿Es necesario estar vivo? ¿Es necesario haber sido producido por una evolución? ¿Dicha evolución ha de ser una evolución natural? ¿Ha de ser biológica? ¿Basada en el carbono? ¿Es requisito ser húmedo?

Algunas definiciones, teorías e hipótesis sobre la sintiencia / consciencia son evidentes para unos pero no para otros, y viceversa. El objetivo de esta exposición es mostrar algunas de estas teorías e ideas y argumentar que no está justificado descartarlas completamente, incluso aunque parezcan descabelladas o anti-intuitivas. Y sembrar la duda, si no existía ya, en cuanto a que tal vez los termostatos sí puedan ser sintientes.

En cuanto a la consciencia y la sintiencia debemos reconocer que hay muchas cosas que desconocemos y por ello es un riesgo muy grande descartar las teorías que no son intuitivas para nosotros. En concreto es un riesgo muy grande en relación a la posible sintiencia de las máquinas.

Para algunos es evidente que las máquinas no pueden ser sintientes, y para otros es evidente que las máquinas sí pueden ser sintientes. Mi objetivo es que ambos grupos tengan en cuenta la hipótesis contraria, y profundizar en los aspectos que pueden determinar la sintiencia o no de las máquinas. Y hacer todo esto orientándolo a prevenir algunos riesgos morales que pueden ser de una magnitud enorme.

[2] Inspirado en el *cogito ergo sum* «pienso, luego existo» de Descartes (siglo XVII)

El mundo está siendo transformado por y para los humanos. Una máquina de propósito general tendría algunas ventajas si tuviera forma humana. Por ejemplo, podría conducir un coche. Podría evocar la ternura en nosotros. También podría atraernos sexualmente, incluso enamorarnos.

Podemos prever que al menos algunas máquinas o robots van a tener apariencia muy similar a la humana, tal vez indistinguible. A medida que las máquinas se parezcan más a nosotros, tanto en aspecto como en comportamiento, nos resultará más fácil empatizar con ellas. ¿Cómo reaccionaremos? Tal vez creando una relación afectiva con el robot, como en las películas *IA*, *Ex-machina* o *Her*. Tal vez aversivamente, como en *Yo, Robot*.

> ***Detective Del Spooner**: Los humanos tienen sueños. Hasta los perros tienen sueños, pero tú no. Tú eres solo una maquina. Una imitación de la vida. ¿Puede un robot escribir una sinfonía? ¿Puede un robot convertir... un lienzo en una obra maestra?*
> ***Sonny (robot)**: ¿Acaso podría Usted?*

Cuando nos presenten una afirmación en relación a lo que es la sintiencia, debemos pedir al interlocutor que aclare si se trata de una definición, o por el contrario, si se trata de una explicación / descripción de las condiciones requeridas para la sintiencia. Si es una definición, no hay nada que discutir. Cada uno llama a las cosas como quiere. Si se trata de una explicación o de las condiciones requeridas, entonces podemos debatir sobre ello.

Por ejemplo, alguien podría decir:

- *"Un ser consciente es aquel que opera un sistema simbólico donde un símbolo es él mismo" (auto-referencia)*
- *"Un ser consciente es aquel que maneja información de su entorno, en forma de mapa".*
- *"La consciencia es la popularidad de las ideas (contrapuestas) o agentes en el mercadillo de la mente[3]".*
- *"La consciencia es un personaje del teatro de la mente. Ser consciente consiste en representar un papel: en parecer (o creer) ser consciente" (eliminativismo).*

Si se trata de definiciones, no hay nada que objetar. Por ejemplo alguien podría definir "volador" de esta forma:

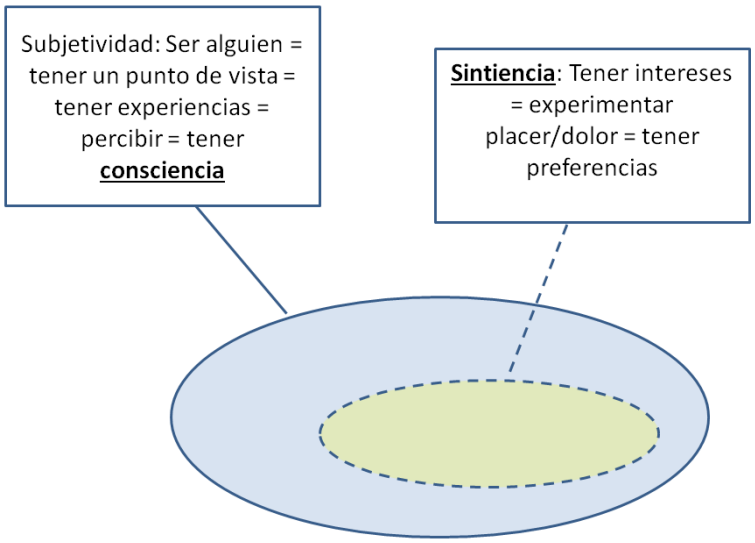
- *"Un ser volador es aquel que tiene plumas".*

Si se trata de una definición, el avestruz sería *volador* (según dicha definición). El avestruz no vuela, pero sería *volador*.

Mis definiciones de sintiencia y consciencia las realizo mediante sinónimos. Con ello trato de que el interlocutor identifique en su propia mente la idea de la que estoy hablando. Pero las palabras que uso se podrían intercambiar o podría usar otras diferentes, no hay problema en ello.

[3] Marvin Minsky entre otros propone que la sintiencia se produce en la interacción de diferentes agentes mentales en un "mercado" en el que compiten y se seleccionan soluciones a diferentes necesidades que pueden ser contrapuestas, como por ejemplo: tener sueño y hambre al mismo tiempo. La sintiencia se relaciona o identifica con la existencia de una representación neuronal del entorno, subjetiva y egocéntrica.

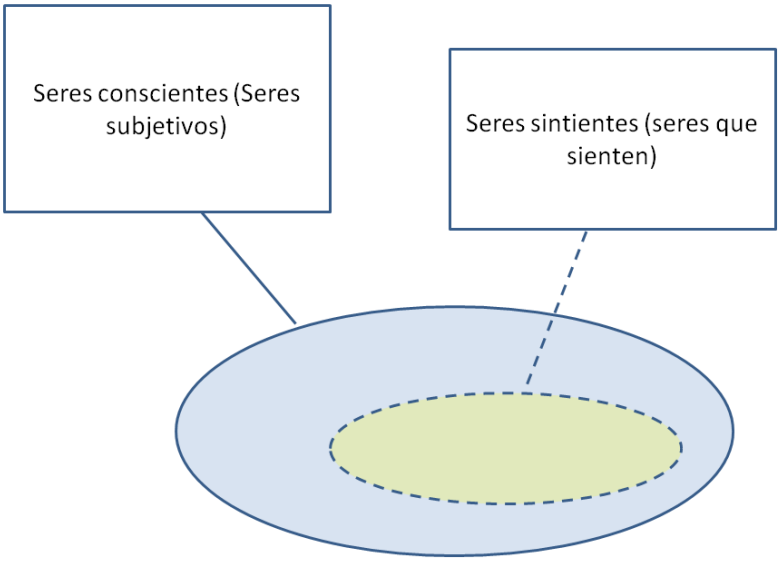
- Considero **Sintiencia** la capacidad de tener sensaciones placenteras o dolorosas, lo que implica tener preferencias e intereses (evitar el dolor, buscar placer).
- Considero **Subjetividad** la capacidad de experimentar. Dentro de experimentar incluyo la capacidad de sentir placer y dolor, pero también incluyo tener un punto de vista, ser alguien, percibir, tener **consciencia**.



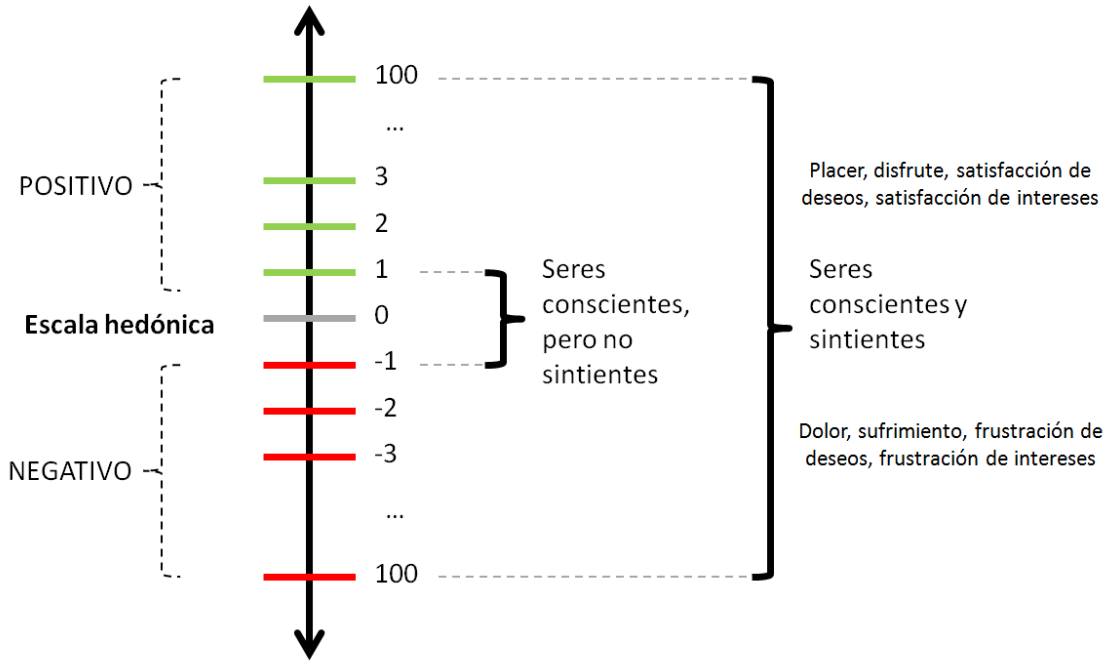
Con estas definiciones asumo que un ser que siente, es obligatoriamente consciente (todos los seres sintientes son a la vez, conscientes).

Adicionalmente asumo que es al menos teóricamente posible ser consciente sin experimentar placer ni dolor: que es posible ser consciente y no tener preferencias.

Las expresiones placer y dolor las empleo en un sentido amplio. No se refieren únicamente a placer y dolor físico sino también psicológico.



Otra forma de verlo es imaginar una escala hedónica de placer y dolor, donde por el motivo que sea, hay seres que se mantienen siempre junto a un valor cero, sin posibilidad de salir de él.



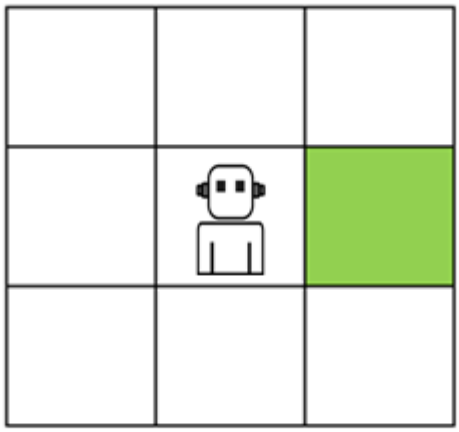
Los seres que sienten placer y dolor tienen relevancia moral (por sí mismos) ya que tienen intereses. En cambio los seres que no tienen preferencias, ni pueden tenerlas, no tienen relevancia moral (por sí mismos).

Si las máquinas tuvieran un punto de vista, pero no pudieran experimentar placer ni dolor, ni tener intereses, no tendrían relevancia moral. Pero si las máquinas pudieran experimentar sufrimiento sí tendrían relevancia moral.

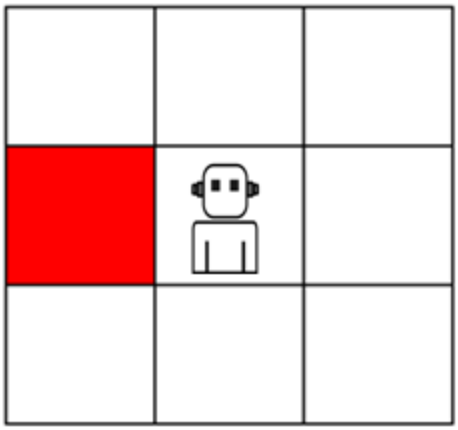
Siendo más precisos, no deberíamos asumir que todos los seres que sienten placer pueden sentir dolor y viceversa. Parece al menos teóricamente posible que existan seres que sólo pueden sentir dolor, o solo pueden sentir placer. Voy a definir tres capacidades y sus combinaciones para explicar estas ideas.

- **Hedón**: es aquel ser que puede experimentar cosas positivas (sentir placer)
- **Sufrón**: es aquel ser que puede experimentar cosas negativas (sentir dolor)
- **Perceptrón**: es aquel ser que puede percibir, tener un punto de vista, darse cuenta.

Todos los hedones y todos los sufrones son obligatoriamente perceptrones. Adicionalmente, estamos acostumbrados a que todos los hedones sean también sufrones y viceversa. Pero eso podría no ser siempre así. Podríamos imaginar seres, incluso mundos, donde solo exista el placer (sólo existan las experiencias positivas), solo exista el dolor (sólo existan las experiencias negativas), o incluso un mundo donde exista otro tipo de experiencias subjetivamente relevantes, que no podríamos calificar ni de positivas ni de negativas (otros ejes o dimensiones adicionales a la escala hedónica).

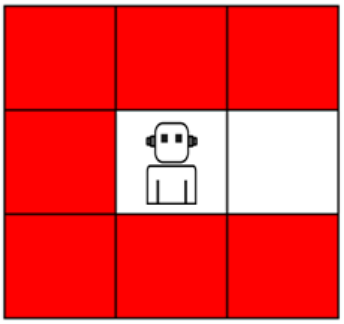


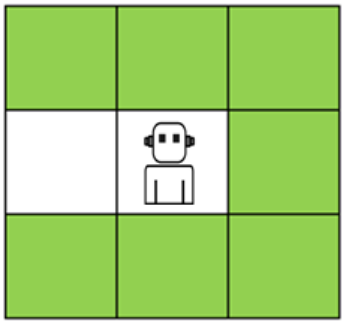
El placer y el dolor parecen ser útiles, y serlo de distintas formas, justificando la existencia de ambas experiencias. Imaginemos que me encuentro en el centro de un tablero, y que la casilla verde es algo bueno para mis genes. Mis genes me han programado para la supervivencia y la reproducción. Soy una máquina biológica creada por los genes con el objetivo de perpetuar dichos genes. Supongamos que la casilla verde fuera comida o una pareja sexual; algo positivo para mis genes. La programación genética me puede llevar a esa casilla mediante el placer.



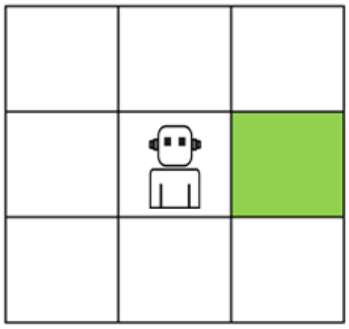
En cambio, si algo amenaza mi supervivencia, como el fuego o un depredador, el dolor parece útil para escapar o evitar cierta situación.

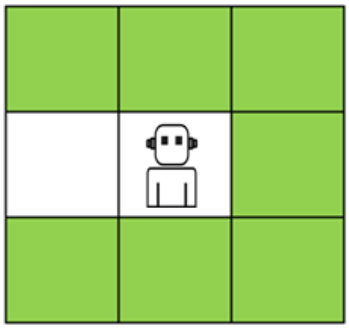
En resumen, el placer es práctico para motivar a "ir" y el dolor es práctico para motivar a "escapar". "Ir" mediante el dolor o "escapar" mediante el placer parece más complicado de conseguir.





Sin embargo, a todos nos gusta el placer y aborrecemos el dolor[4]. Sería complicado, pero no imposible, que todos los comportamientos estuvieran guiados por el placer, y convertir el dolor en algo innecesario, e inexistente. Para ello, lo que necesitamos es un sistema cognitivo que funcione así para "Ir" y para "Escapar":





Eso es precisamente lo que propone David Pearce en su proyecto abolicionista "The hedonistic Imperative": un comportamiento guiado por gradientes de bienestar, donde el dolor sea innecesario e inexistente.

Pero ¿realmente es necesario sentir para comportarse como si se sintiera? ¿Realmente es necesario sentir el miedo para provocar la reacción de salir corriendo? ¿Acaso no podría la naturaleza habernos programado con los comportamientos más adecuados para maximizar la supervivencia y reproducción de los genes, sin tener que experimentar absolutamente nada? ¿Podría un robot estar programado para comportarse como un animal, sin sentir tal como sentimos los animales?

David Chalmers propone que ciertos "zombies" son metafísicamente posibles (compatibles con nuestro conocimiento de la física): seres que son físicamente o se asemejan a las personas conscientes, pero que no son conscientes. Estos zombies gritan "¡Ay!" al pisar un clavo, pero no sienten nada en absoluto.

Hay dos posibles reacciones a la posible existencia de este tipo de zombies.

- **Son imposibles**: Cierto comportamiento va unido inevitablemente la sintiencia / consciencia

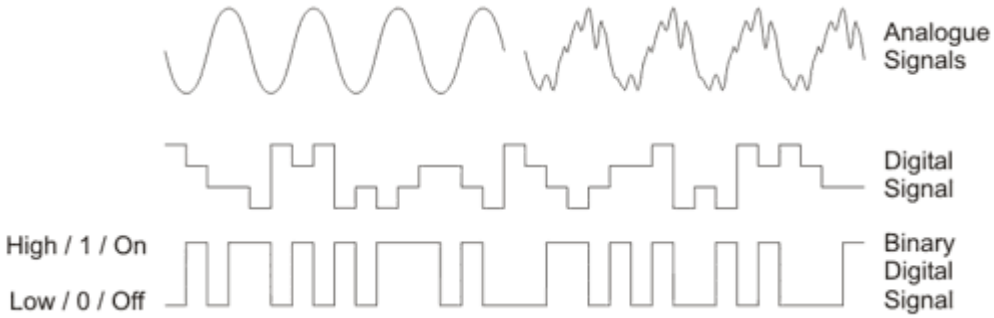
[4] Se podría objetar decir que hay ciertos dolores que producen satisfacción, al menos a ciertas personas. Dichos "dolores que gustan" quedarían incluidos dentro de lo que yo llamo "placer".

- **Son posibles**: Puede producirse ese comportamiento sin que exista sintiencia / consciencia. Este caso se ha empleado para argumentar en favor del dualismo.

Tanto si respondemos sí como si respondemos no, en ambos casos, los robots -las máquinas artificiales- podrían ser sintientes. En el primer caso, las máquinas podrían ser sintientes, simplemente, si adquirieran dicho comportamiento. En el segundo caso las máquinas podrían ser sintientes si incluyeran en su composición eso que hace a los animales sintientes.

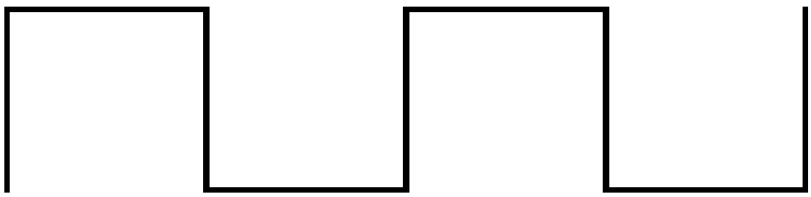
¿Qué es lo que hace a los animales sintientes? Se han mencionado diversos factores que pueden ser relevantes para la sintiencia:

- Naturales (no creados por el ser humano).
- Seres vivos, basados en el carbono, fruto de una evolución natural.
- Seres húmedos.
- "Realizaciones" en el mundo físico, y no simulaciones.
- Seres analógicos, no digitales
- Efectos cuánticos en nanotubos (Roger Penrose)
- Interacción con el Multiverso

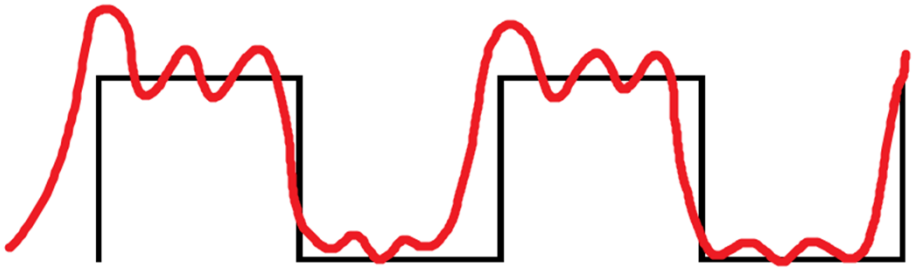


Fuente: mrcorfe.com

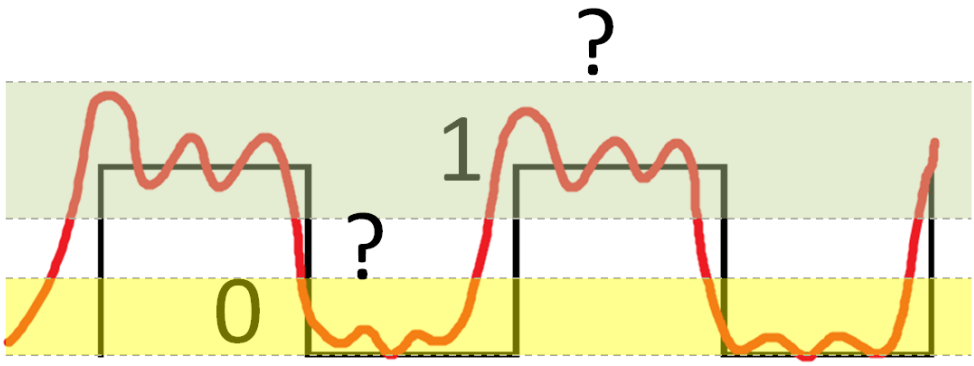
Nos han enseñado que la primera es una señal analógica y la segunda y tercera son señales digitales.



Sin embargo, las señales digitales como esta, físicamente no existen. Lo que existe es algo parecido a esto:

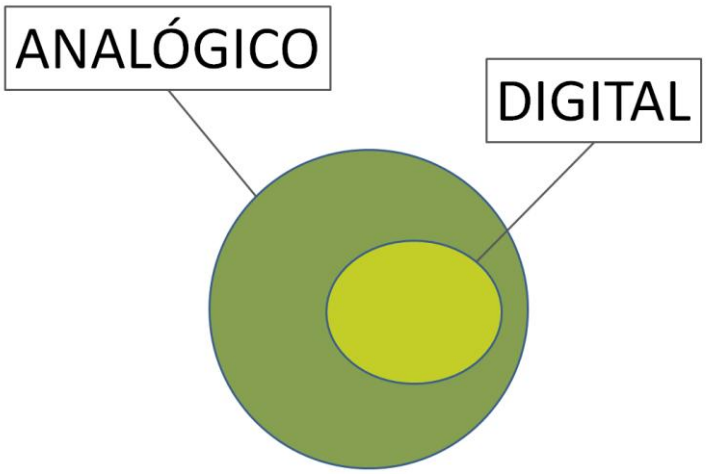


Y algún sistema que reacciona a dicha señal interpretándola, por ejemplo, de la siguiente forma:



Es decir, cuando ante una señal o magnitud física existe algún sistema que la interpreta (reacciona) mediante estados discretos, hablamos de señales digitales. Pero la señal sigue siendo analógica.

Nuestro universo es analógico (simplificando: tiene una precisión "infinita"). En cambio los ordenadores digitales discretizan (limitan) el tiempo y la información a una serie de valores predeterminados.



En definitiva, todo es analógico, y dentro de las cosas que son analógicas (todas), algunas son digitales. Esto ocurre hasta llegar al nivel cuántico, donde espacio (posición), tiempo (momento) y materia / energía / información podrían ser fenómenos discretos / digitales.

El enorme éxito de los computadores digitales es debido a que se aproximan a una máquina de propósito general. Pero el ser digitales conlleva el coste de perder algunas características muy útiles de los ordenadores analógicos. En cierto modo, parece como si el desarrollo de ordenadores, al ser digitales, se hubiera metido en un callejón sin salida (un máximo local). El uso de magnitudes físicas sin discretizar tiene inconvenientes, pero también algunas ventajas enormes. Los "computadores analógicos" son de propósito específico (menos flexibles), suelen sufrir averías y procesos de desgaste que provocan falta de precisión. Sin embargo, pueden ser extremadamente eficaces (por ejemplo, rápidos) para ciertas tareas obteniendo rendimientos radicalmente diferentes a los de los ordenadores digitales. Además sus fallos típicos pueden ser más aceptables, como un proceso progresivo de "desgaste", mientras que los fallos típicos de los sistemas digitales son más bien de la forma todo o nada.

Los algoritmos de ordenación que se emplean en computadoras digitales tienen una complejidad computacional exponencial o asimilable a exponencial. Esto quiere decir que el tiempo necesario para realizar la ordenación no se incrementa de forma más o menos constante en función del incremento del número de elementos que se necesita ordenar, sino que la situación se vuelve cada vez peor, y va empeorando cada vez más. La solución a esto ha sido emplear la fuerza bruta, aprovechando la Ley de Moore -que se ha venido cumpliendo desde 1965- según la cual las prestaciones de los computadores se duplican cada año, lo que supone un incremento de capacidad exponencial.

Sin embargo existe un algoritmo *analógico* cuya complejidad computacional no es exponencial: el tiempo necesario para realizar la ordenación se incrementa al incrementarse el número de elementos que se necesita ordenar, pero estos incrementos son constantes, directamente proporcionales al número de elementos. ¿Cómo funciona este algoritmo maravilloso?

El funcionamiento de esta máquina analógica de ordenar es el siguiente. Por cada elemento que se requiere ordenar, se corta un palito cuya longitud es proporcional al número que se desea ordenar y se escribe dicho número en el palito. Cuando se han cortado todos los palitos, se juntan en un haz, colocados de forma vertical, se sujetan firmemente, y se da con ellos un golpe sobre una superficie plana, por ejemplo, en la mesa. Después se baja la mano sobre el haz. El primer palito que toque nuestra mano será el del número más alto. Lo extraemos y anotamos el número. Después volvemos a bajar la mano, tocando y extrayendo el segundo número, y así sucesivamente.

Los computadores digitales han eclipsado a los analógicos, pero tal vez la extraordinarias ventajas de los ordenadores analógicos, como su precisión "infinita" o su capacidad de resolver eficientemente problemas como el de la ordenación pudieran ser un requisito para la sintiencia, ya que las máquinas para las cuales tenemos pruebas abrumadoras de su sintiencia (los animales en general) son máquinas analógicas.

1
11
21
1211
111221
312211
13112221
1113213211
31131211131221
13211311123113112211
11131221133112132113212221
3113112221232112111312211312113211
1321132132111213122112311311222113111221131221

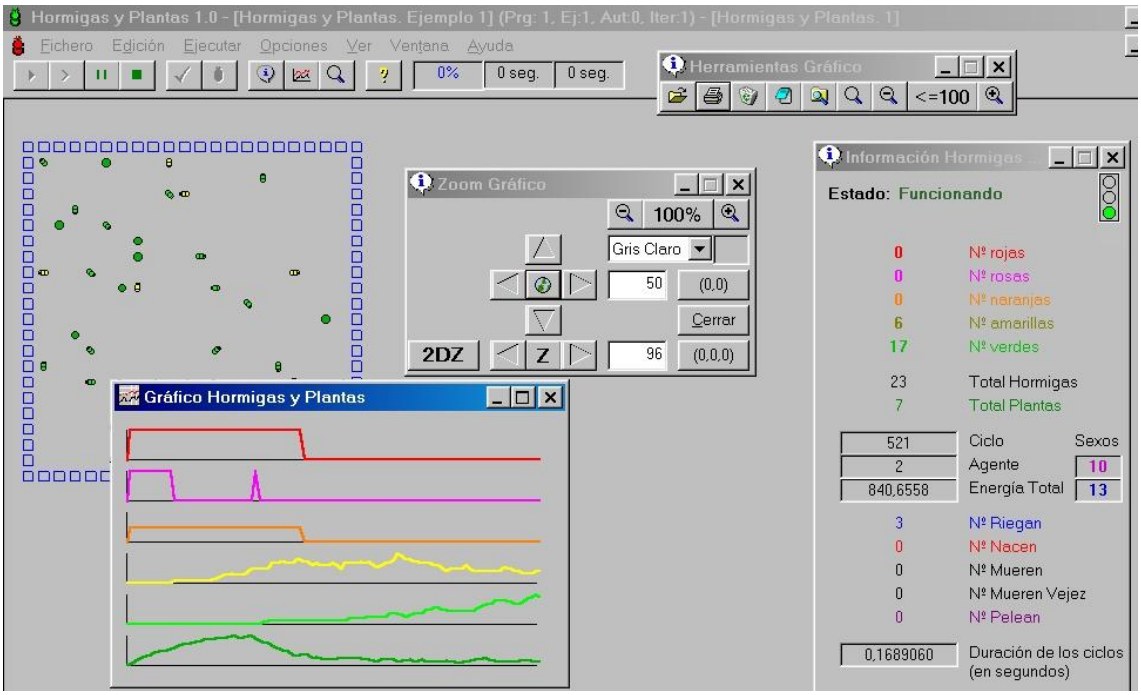
¿Cuál es el siguiente valor de esta serie? Puedes ver este vídeo o descargar este programa para descubrir la solución.

¿Puede un cerebro ser lo suficientemente complejo como para comprenderse a sí mismo? Parece difícil.

Hay fenómenos que parecen complejos y no lo son. Para explicar lo que no entendemos (por ejemplo, la sintiencia / consciencia) apelamos a otras cosas que tampoco entendemos bien (como los sistemas complejos, o los sistemas inteligentes).

La sofisticación de la complejidad se disuelve cuando entendemos su origen, como en el caso de la serie anterior. Es injusto decir que el cerebro humano, como no lo comprendemos, tiene suficiente complejidad como para crear sintiencia, y que otros cerebros más sencillos, como sí los comprendemos, no son sintientes. Si un ser extraterrestre super-inteligente pudiera entender perfectamente nuestro cerebro y anticipar su comportamiento ¿debería concluir que no somos sintientes?

El siguiente robot de charla fue creado con unas pocas reglas muy sencillas, pero fue capaz de "engañar" a varias personas que llegaron a creer que estaban hablando conmigo. Considerar que cierto "alguien" que se encuentra al otro lado del chat pueda ser un ser sintiente es una especulación razonable. Pero el hecho de entender cómo funcionan sus procesos mentales no es un argumento para rechazar su sintiencia.



En este programa que simula la evolución de unas hormigas con comportamientos altruistas, cooperativos y egoístas, reproducción sexual (recombinación) y mutaciones se puede observar cierta "evolución de las especies" adaptándose a su entorno. ¿Cuál es el argumento para pensar que una evolución natural puede producir seres sintientes, y en cambio no puede hacerlo una evolución artificial en simulaciones por computador?

Se escucha decir que la subjetividad / consciencia implica la existencia de un yo, de una identidad, y que los seres humanos tenemos una libertad, una voluntad, un libre albedrío. Hay una confusión habitual en este sentido: Libre no es lo mismo que impredecible.

- **Ser impredecible no implica ser libre**: si tuviéramos un sistema físico aleatorio, diríamos que es impredecible pero no diríamos que por ello es libre.
- **Ser libre no implica ser impredecible**: Puedo libremente decidir tomar siempre la misma decisión, y los demás podrán predecir correctamente mi futura acción.

A nivel "macro" podemos considerar que "todo tiene una causa" (genes, entorno), pero en este asunto realmente no importa si nuestro comportamiento está determinado o no, si es predecible o no, si es aleatorio o no. La libertad consiste en tener la *sensación* de ser libre. Se trata de tener la experiencia de que es uno mismo quien toma las decisiones. Lo que llamamos *libertad* es realmente *identidad*: "Soy libre cuando lo decido *yo*".

Habitualmente asociamos la palabra libertad a un espacio físico inmenso, sin límites ni obstáculos, y la falta de libertad a una cárcel. Pero a ninguno de nosotros nos gustaría que nos sacaran de la cárcel para dejarnos abandonados en la sabana africana a merced de los depredadores, o en la superficie de la luna sin oxígeno. La libertad está subordinada al concepto de interés. No es más libre aquel a quien se le ofrecen más alternativas, sino aquel que puede elegir aquello que más desea. Ser libre consiste ante todo en poder satisfacer los propios deseos e intereses.

El deseo de libertad es uno de los deseos más fuertes que existen, y ese es el motivo por el que debemos respetarlo: en general los individuos quieren ser libres, quieren tener la sensación de que toman sus propias decisiones, y en general quedarían frustrados si no se les permite hacerlo.

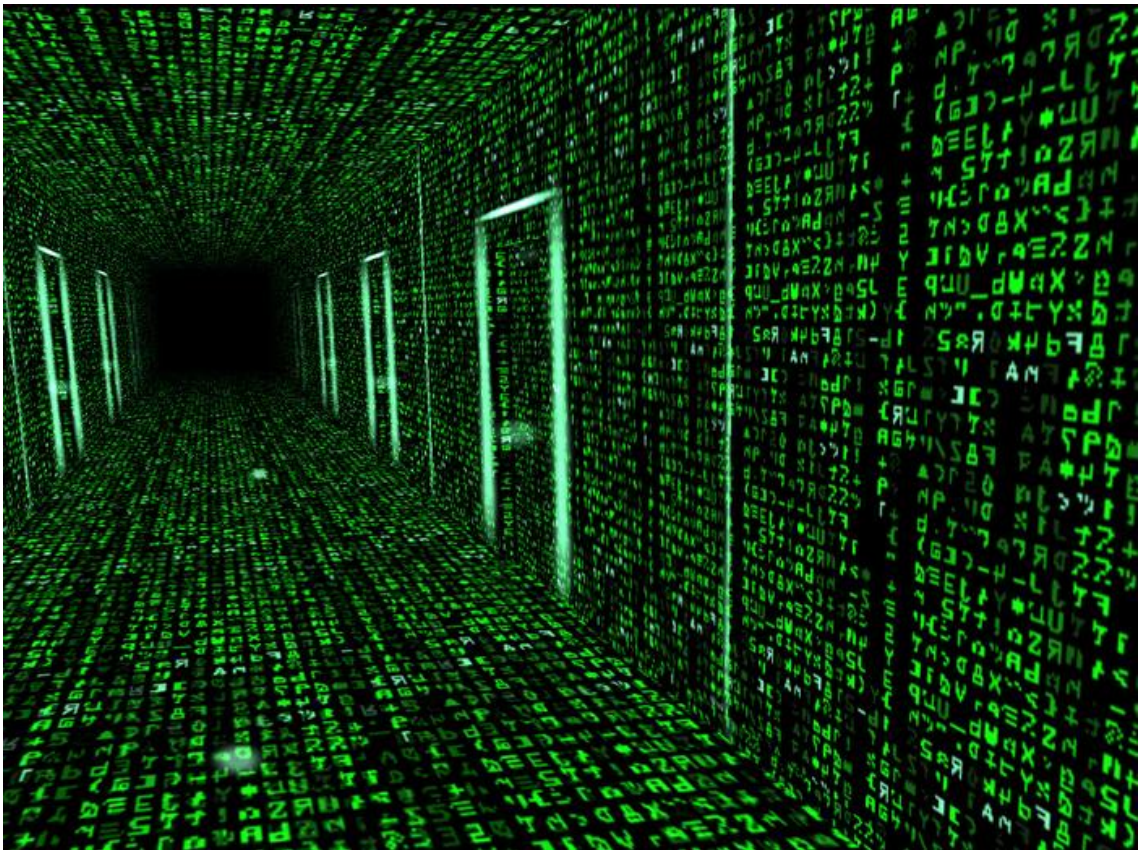
Hay quien se pregunta: *"Si estamos determinados por genes + ambiente ¿qué hacemos con todos los penados? ¿Abrimos todas las cárceles?".* La respuesta es sencilla: ¿cuál era la finalidad de esa pena de cárcel? Según las distintas teorías la finalidad puede ser la sanción (compensación, reacción), la prevención, enmienda, readaptación, etc. Bien, pues dicha finalidad de la pena no se ve afectada por el hecho de no existir la libertad en el sentido habitual de la palabra, y se puede aplicar igualmente a humanos y a robots.

Ahora voy a comentar una serie de "ideas curiosas" relacionadas con los conceptos de "identidad" y "realidad", y relevantes para establecer la sintiencia o no de las máquinas.

Todos mis recuerdos podrían ser falsos. Todo mi pasado hasta hace exactamente medio segundo podría haber sido implantado en mi mente, en mi memoria, como ocurre en la película Total Recall / Desafío Total

Podría haber una anestesia que no eliminase el dolor, sino el recuerdo del dolor. Esta anestesia es mencionada en una de las novelas de ficción de Robert Anson Heinlein, y en el ensayo "El peor de los males" de Thomas Dormandy.

Tal vez vivimos en una simulación, como sugiere Nick Bostrom.



Nuestro mundo, como el mundo de las hormigas que hemos visto antes, podría detenerse durante años (botón "*pause*"), y después reanudarse (botón "*play*"); y podríamos aumentar o disminuir proporcionalmente el tamaño de los objetos del

mundo mediante el botón de "*zoom*", manteniendo las proporciones, sin que nadie note nada raro.

El barco invencible ganaba todos los combates navales. Como es de suponer, era muy apreciado y cada vez que se reemplazaba alguna pieza, éstas no se destruían, sino que se guardaban en un almacén. Por desgracia, los enemigos entraron cierto día en el almacén y robaron todas las piezas, con tal mala suerte que se habían reemplazado tantas partes del barco invencible, que existían piezas suficientes como para crear un nuevo barco invencible. Y eso hicieron, y este nuevo barco invencible parecía incluso más auténtico. Cuando se enfrentaron en el mar los dos barcos invencibles ¿quién ganó? ¿Qué pasaría si hiciéramos esto con nuestras neuronas?

Tal como muestras innumerables relatos de ciencia ficción sobre el teletransporte: si copiara, una por una, todas mis neuronas y otras células de mi cuerpo, formando un nuevo cuerpo ¿cuál de los dos sería *yo*?

Si sustituyera, una por una, todas mis neuronas naturales, por dispositivos artificiales que externamente se comportasen como neuronas ¿perdería mi capacidad de sentir?

Si no sustituyera, sino ampliara mi capacidad neuronal con dispositivos artificiales ¿podría hacerlo de forma que superaran de tal manera a los biológicos que podría incluso desconectar la parte biológica sin notar gran diferencia?

Si pudiera conectar directamente mi cerebro humano con el de otro animal, uniendo nuestras neuronas ¿seríamos un individuo o dos?

Si pudiera separar en mi cerebro humano las neuronas, creando varios grupos de neuronas bastante separados entre ellos, como en cierto modo ya ocurre al tener dos hemisferios ¿seríamos un individuo o varios? Los delfines duermen cada vez con un hemisferio ¿dentro del delfín hay uno o dos individuos?

Existen muchas alternativas en relación a la sintiencia que podríamos resumir en las siguientes, creando una especie de mapa o ejes:

A: ¿Quién siente? Yo, los de mi especie, animales con un sistema nervioso central, materia
B: ¿Qué se siente? Placer, dolor…
C: ¿Cuánto se es capaz de sentir? Discreto, continuo, máximo...
D: ¿Cuándo se siente? Pasado, Presente, Futuro
E: ¿Dónde se produce la experiencia sintiente?
F: ¿Por qué se produce la experiencia sintiente? ¿Cómo es posible la experiencia sintiente? ¿Cuál es su naturaleza?
G: ¿La sintiencia se crea o es pre-existente?
H: ¿Para qué se produce la experiencia sintiente? ¿Tiene alguna finalidad?
I: ¿Cuánta experiencia sintiente hay? ¿Cuál es su disponibilidad?

La hipótesis de la identidad separada o "individualidad cerrada" podría ser una ilusión, y en cambio existir una subjetividad única como la que describí en mi libro "Arena Sensible" (2005) o el "Individualismo Abierto" de Daniel Kolak en "I Am You" (2004).

Todos los días, a todas horas, tenemos la experiencia de que las cosas caen hacia abajo. Todos los días, a todas horas, obtenemos evidencias de que las cosas caen hacia abajo. Y sabemos que no es verdad. Sabemos que el concepto "abajo" está mal planteado. La pregunta "¿las cosas caen hacia abajo?" es una de esas preguntas mal planteadas que es mejor no responder de forma "Sí / No".

La realidad puede ser como un cuadro tapado por un lienzo. El "yo" un agujero en la tela, y el "tu" es otro agujero.

Al menos a nivel *descriptivo* podemos hablar de que existe la materia, las ideas y las experiencias. Son tres tipos de cosas o tres tipos de realidades.

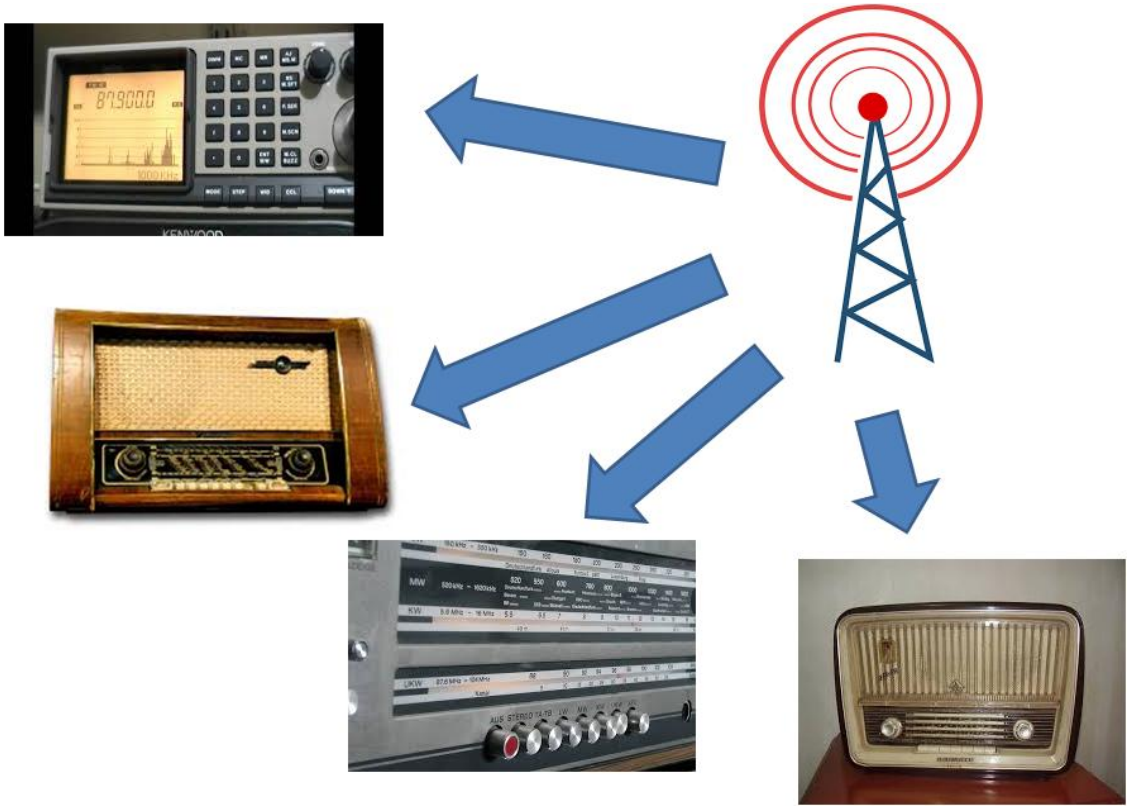
Curiosamente, coinciden con los tres tipos de trabajos que existen: manipular materia (procesar cosas materiales), manipular ideas (consultoría, servicios), y

manipular experiencias (actividad comercial, política, educativa, manipular personas).

Hay materialistas / reduccionistas / monistas que dicen: "Yo sólo creo en la materia. Sólo puedo estar seguro de lo que puedo tocar: las cosas, lo material, lo tangible". Pero lo "tangible" se refiere a la *experiencia* de *tocar*. Está hablando de experiencias, no de materia. Estas personas que dicen que todo es material, en realidad están diciendo que todo es experiencia.

Estás en casa. Sobre la mesa hay una manzana. Señalas primero con el dedo la manzana y a continuación señalas con el dedo tu propia cabeza y dices: "¿De qué forma el cerebro se las arregla para representar esta manzana en mi mente?". Ese planteamiento está totalmente equivocado. Esa cosa que señalo no es ninguna manzana real: es la representación que mi mente tiene de la manzana. Ahora señalo a mi cabeza. Esa no es mi cabeza. Esa es la representación mental que mi mente tiene de mi cabeza. Ahora salgo de la casa. Veo el campo, los árboles, la luna y las estrellas. Abro los brazos, miro a mi alrededor y digo: "todo esto, es mi mente". Eso es correcto. Por si hubiera alguna duda, las personas que sufren una amputación habitualmente se quejan de dolor en el miembro que ya no existe. Lo que llamamos "mi pie derecho" es una representación mental en el cerebro. El zapato también.

Habitualmente se considera que la sintiencia emerge de la materia. Pero también pudiera ocurrir que la sintiencia se recibiera, invocara o sintonizara, como ocurre con los receptores de radio. La configuración material de nuestros cuerpos, al igual que la configuración material de los receptores de radio, podría, no generar sintiencia, sino estar sintonizados con ella, de forma que en función de la composición interna se recibiera una u otra señal sintiente.



En vez de emerger ("hacia arriba"), la sintiencia podría ir hacia abajo, podría sintonizarse, como un receptor de radio que escucha en cierta frecuencia. Una

muela con caries podría ser una configuración material que resuena o se alinea o se sintoniza con una señal que está emitiendo constantemente y que genera ese dolor en lo que aparentemente es un individuo.

La hipótesis del platonismo sintiente considera que las experiencias podrían existir por sí mismas, con independencia de los seres sentientes que las experimentan. Incluso aunque la probabilidad del platonismo sentiente fuera extremadamente pequeña, mientras exista una probabilidad mayor que cero, y teniendo en cuenta que no está muy claro de dónde viene la sentiencia, podríamos pensarlo dos veces antes de ignorar esta idea, ya que en caso de ser cierta, sus implicaciones en cuanto a prevenir el sufrimiento serían inmensas.

En la hipótesis de la simulación de Nick Bostrom, ciertamente, no es necesario simular todo el universo físico, sino únicamente el universo tangible (aquella parte que alguien va a percibir). Pero es que, de hecho, no es necesario simular absolutamente nada material del universo; únicamente es necesario simular las experiencias subjetivas. Y ésta forma de ver el argumento de la simulación sí parece ser más propicia al platonismo sentiente y al monismo inmersionista, ya que de la misma forma que nosotros creemos que empleamos materia (ordenadores) para simular materia (aviones y puentes), tal vez en otro mundo platónico alguien esté empleando sintiencia para simular sintiencia.

Supongamos que tenemos varios cubos de agua con diferentes temperaturas. Imaginemos que existen unas esponjas de plástico que entran y salen de los cubos.

En esta metáfora las esponjas mojadas son seres sentientes, y el agua (a diferentes temperaturas) son las experiencias. Cada esponja junto con el agua contenida en ella conforman un todo coherente, de forma que se identifica la temperatura del agua contenida dentro de la esponja como la experiencia específica que experimenta dicha esponja.

Las altas temperaturas corresponden con experiencias positivas (diferentes satisfacciones o placeres) y las bajas temperaturas con experiencias negativas (diferentes frustraciones o dolores). En cada esponja, cuanto más alta sea la temperatura, tanto más placentera será la sensación, y cuanto más baja, más dolorosa; existiendo una temperatura intermedia en la que se produce un estado de indiferencia, sin placer ni dolor significativo.

Es muy importante señalar que en esta metáfora el agua representa las experiencias, y estas experiencias (agua) existen independientemente de las esponjas. Por ejemplo, un cubo etiquetado como "dolor de muelas" podría contener agua azul a 4 grados centígrados, y otro etiquetado como "cosquillas agradables" podría contener agua naranja a 27 grados centígrados. Al introducir la esponja en un cubo, la esponja adquiere dicha experiencia (dolor de muelas o cosquillas) y las esponjas (seres) pueden identificar sus propias experiencias y las de otros seres en función del tipo de agua que contienen. Pero el agua existe independientemente de las esponjas.

Las esponjas entran y salen de los cubos. Al cambiar de cubo, la esponja adquiere un agua diferente, de la misma forma que los seres experimentamos diferentes cosas a lo largo del tiempo. En un proyecto para reducir o eliminar el sufrimiento, y según esta metáfora, existe el riesgo de que podríamos estar enfocados en aliviar el sufrimiento llevando esponjas progresivamente desde los cubos más fríos hacia los cubos más calientes. Y suponiendo que existe una mayoría de cubos de agua fría, y que las esponjas son capaces de reproducirse, podríamos dedicarnos a promover evitar la reproducción de las esponjas, como forma de evitar que existan nuevas esponjas en cubos de agua fría. Incluso podríamos plantearnos que el mejor mundo posible es aquel en el que no existen esponjas.

Efectivamente, en ambos casos podríamos comprobar que hay esponjas individuales y muy concretas que han mejorado su bienestar, y que ahora se encuentran en cubos más calientes. O que ya no hay ninguna esponja en ciertos cubos especialmente fríos. El problema es que dichos cubos de agua fría seguirían existiendo, y si la metáfora fuera cierta, no habríamos solucionado absolutamente nada. Si esta metáfora representase verdaderamente la naturaleza de las experiencias, estaríamos perdiendo el tiempo llevando unas esponjas de unos cubos a otros, o evitando que las esponjas se reprodujeran. Lo que deberíamos hacer es aumentar la temperatura de los cubos fríos. Al hacerlo, ciertas experiencias negativas dejarían de existir, o quedarían aliviadas, instantáneamente, para todos, y para siempre.

Así como la sentiencia de los robots y de las simulaciones podrían generar una catástrofe moral, deberíamos también tener en cuenta la posibilidad del platonismo sentiente. Aunque nos parezca improbable, sus consecuencias serían gigantescas.

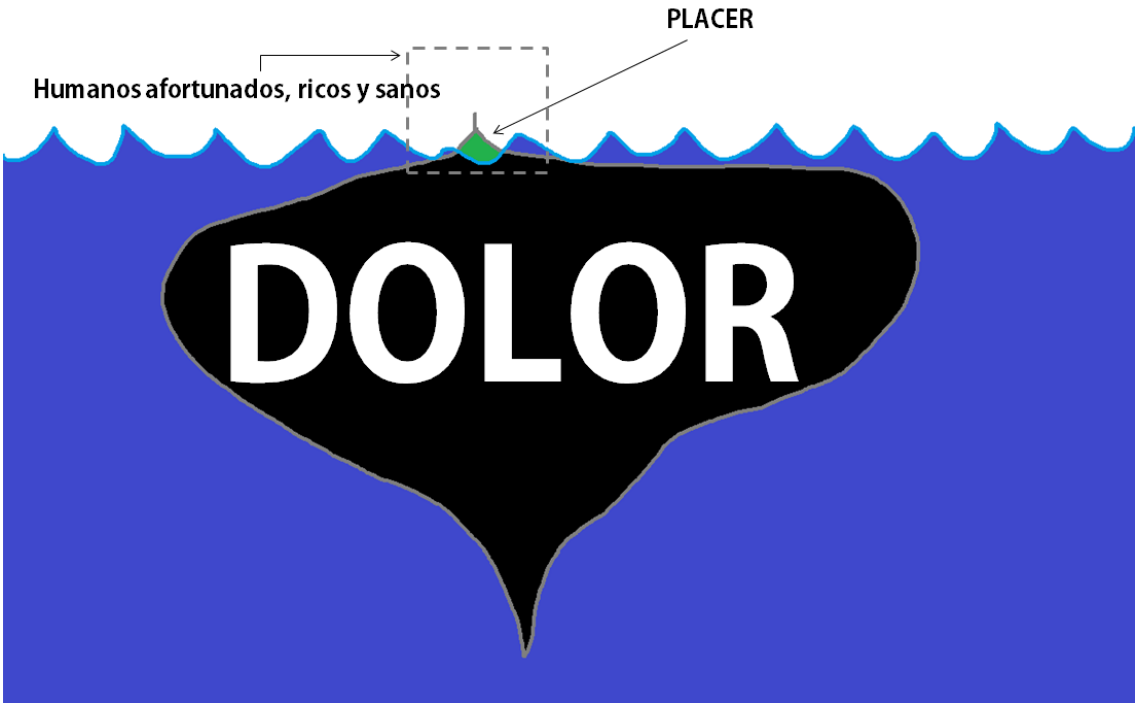
Cuando explicamos cómo reconocemos a otro ser sentiente, hablamos de criterios conductuales, evolutivos y fisiológicos, pero lo hacemos "a posteriori" de una manera muy egocéntrica. Partiendo de la evidencia máxima "yo siento", parece que realizamos una interpolación entre individuos que tienen aspecto similar y/o comportamiento similar al propio ("Si se parece a mí y se comporta como yo, sentirá como yo"); también parece que hacemos una interpolación entre especies ("Si ha sido creado como yo, sentirá como yo") valorando si tiene el mismo origen (evolutivo) y si es genéticamente próximo a mí. Adicionalmente y dado que la sentiencia tiene una utilidad evolutiva, parece que establecemos -injustificadamente- que tener una utilidad evolutiva es un criterio necesario para existir sentiencia.

Si observáramos un grupo de seres humanos matando un cerdo para después comérselo, podríamos tratar de defender al cerdo argumentando que el animal siente: "mira la expresión de pánico sus ojos", "escucha sus desgarradores gritos", "fíjate como se retuerce de dolor"... pero esto podría ser como decir, para argumentar que un animal es volador: "fíjate que huesos tan ligeros", "qué forma tan aerodinámica" y "qué majestuosas alas". El hecho de que esa descripción encaje con la de un ser volador no quiere decir que todos los seres voladores deban tener dichos atributos. Los aviones son aerodinámicos pero no tienen plumas ni son especialmente ligeros.

Esta forma egocéntrica de reconocer la sintiencia es muy peligrosa para aquellos que no se parecen a nosotros, como los robots, los insectos, las simulaciones, los átomos y quién sabe qué otras cosas.

Si alguien me habla español puedo entender lo que dice y saber si es inteligente. Pero si me habla en otro idioma que desconozco, no puedo saber si me dice una genialidad o una banalidad. Es injusto desconsiderar y despreciar radicalmente aquello que no entendemos. Existe la idea muy extendida de que sólo los animales somos sintientes, y sin embargo no entendemos bien lo que es la sintiencia. Mientras tanto creamos robots y simulaciones de vida por computador cada vez más complejas. Esto es moralmente muy atrevido.

Otra forma de verlo: Si yo soy egoísta (analogía de sintiente) y quiero saber quién más es egoísta en un grupo de personas, puedo buscar en otros comportamientos que en mi caso van asociados al egoísmo. O también: si soy heterosexual y quiero saber quién más es heterosexual, puedo buscar en otros comportamientos que en mi caso van asociados a ser heterosexual. Pero esto no quiere decir que no existan otras personas, egoístas o heterosexuales (analogía de sintientes) que simplemente lo oculten o lo manifiesten de formas que yo no entiendo.



Por si esto fuera poco, los que escribimos y leemos sobre temas como los tratados aquí somos algunos de los individuos más afortunados (con mayor calidad de vida), miembros de la especie más afortunada -la humana-, y en el momento de mayor

bienestar de la historia. Esto nos puede hacer olvidar la asimetría entre placer y dolor: hay mucho más dolor que placer. Hemos logrado controlar hasta tal punto el dolor físico que el dolor psicológico nos parece muy relevante, cuando la mayoría de los animales de otras especies, y la mayoría de los seres humanos de otras épocas, han experimentado grandes dosis de dolor físico en muchas de las etapas de sus vidas. El riesgo de extender el sufrimiento y un gran sufrimiento, en máquinas y simulaciones es muy elevado.

En conclusión:

- No existe una idea clara de lo que es la sintiencia / consciencia: ni de lo que la genera / invoca, ni de las condiciones necesarias para que suceda.
- Desde distintas perspectivas filosóficas, las máquinas pueden ser sintientes. Lo que cambia son las condiciones que se suponen necesarias para que lo sean.
- La creación de máquinas o simulaciones sintientes puede provocar una catástrofe moral de una magnitud astronómica.
- No debemos descartar hipótesis simplemente porque nos parecen descabelladas o anti-intuitivas. Sus implicaciones morales podrían ser extraordinarias.