¿Sienten las máquinas? (Parte 1)

Manu Herrán
Primera versión: Dic. 2017
Actualizado: Mar. 2018
Actualizado: Ene. 2019

Este texto lo he creado a partir de los materiales que preparé para la charla coloquio que di en el Grado en Diseño y Gestión de Proyectos Transmedia de la Universidad de La Salle el 15 de diciembre de 2017 y que llevaba por título "¿Sienten las máquinas?" así como la conferencia "Del antiespecismo al antisubstratismo: singularidad tecnológica y sintiencia en máquinas" en la Universidad Complutense de Madrid durante las Jornadas de Análisis Crítico del Especismo organizadas por AUCE y Ética Animal el 31 de octubre de 2017.

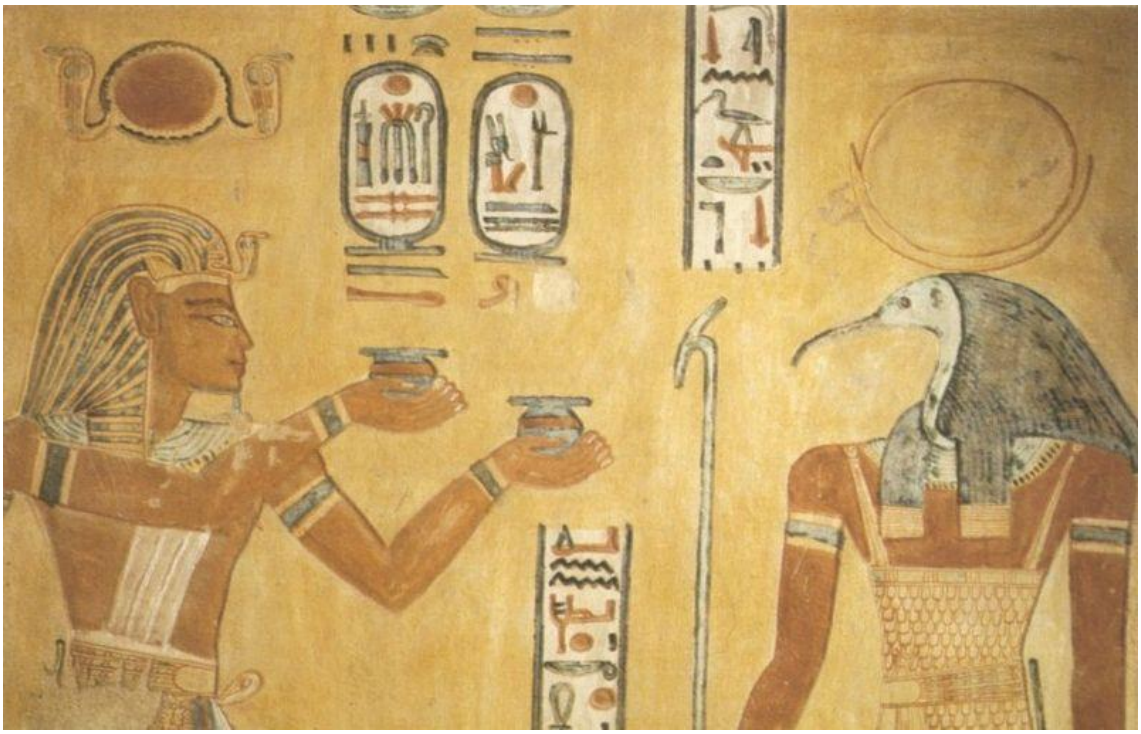


Fig. 1 RamsesIII and Thoth
https://commons.wikimedia.org/wiki/File:RamsesIII_and_Thoth_in_QV44.jpg

No hace mucho que los seres humanos adoraban al sol. El dios sol ofrecía una explicación sencilla y aparentemente válida sobre las cosas que ocurren. El sol determinaba las cosechas y en definitiva toda la vida. Era razonable pensar que desde su inalcanzable altura, el sol determinaba el curso de los acontecimientos. La mayoría de los seres humanos han creído en uno u otro dios y este relato religioso es capaz de explicar cualquier cosa. Estas creencias nos parecen ahora absurdas, pero las religiones modernas son versiones más evolucionadas de este tipo de explicaciones, más o menos adornadas y sofisticadas, que han ido adaptándose a medida que el conocimiento científico ha ofrecido evidencias ampliamente aceptadas por la sociedad que contradicen el relato precedente.

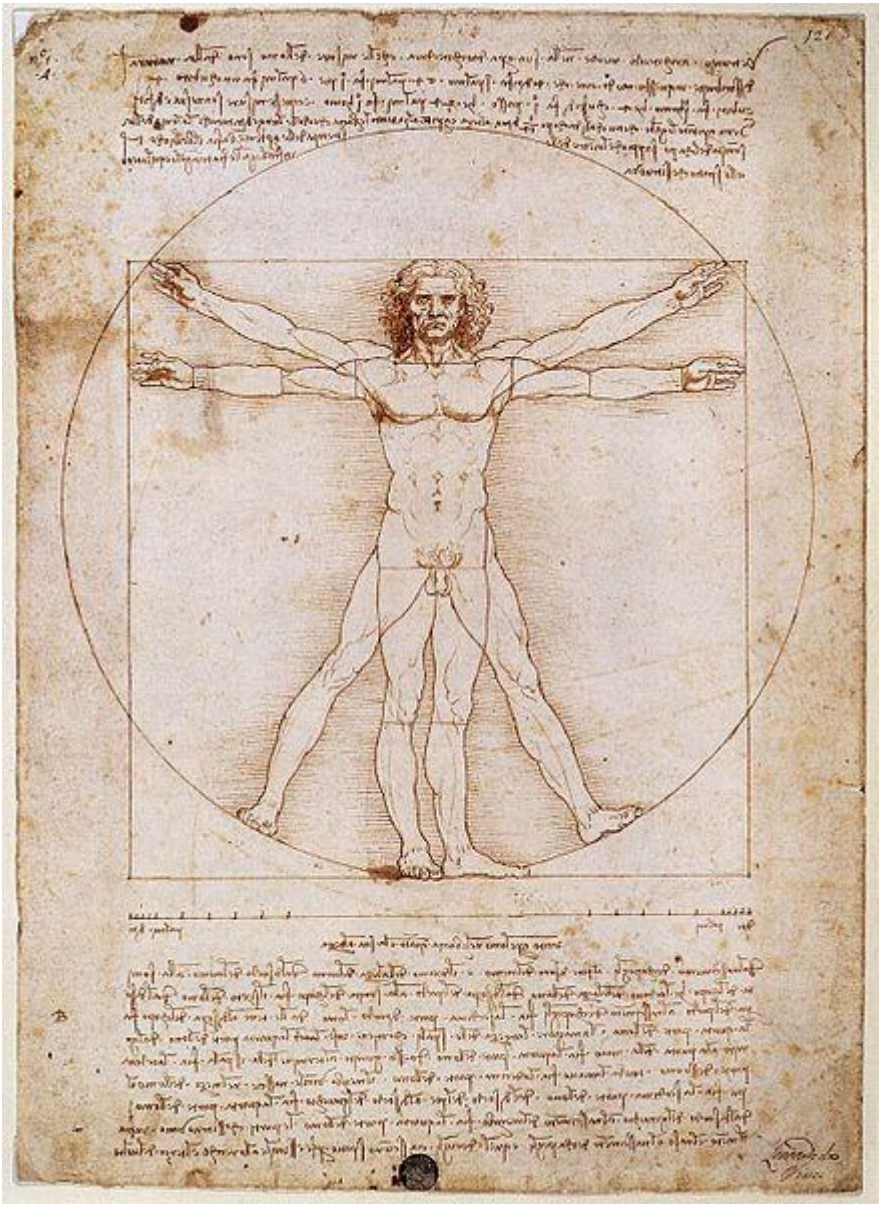


Fig. 2 Vitruvian
https://en.wikipedia.org/wiki/File:Vitruvian.jpg

Tras el éxito de la cosmovisión teocéntrica se produjo una transición del teocentrismo al antropocentrismo, que aún vivimos. El hombre pasó a considerarse *"medida y referencia de todas las cosas"* (Protágoras) en el sentido de que todo lo relevante lo es porque es relevante para los miembros de la especie humana. Esto supuso un gran avance moral, al incluir dentro del círculo moral a todos los seres humanos. Podríamos decir que el evento más representativo del antropocentrismo es la Declaración Universal de los Derechos Humanos. Este es un gran avance moral porque incluye dentro del círculo moral no solo a los hombres ricos y poderosos, sino también a los parias, a los esclavos, a las mujeres, a los niños y a los vencidos en combate, que durante mucho tiempo han tenido consideración más o menos de propiedad.

Especialmente desde el siglo XVII, el racionalismo primó el uso de la razón frente a otras consideraciones como la superstición, el mito, la intuición, la autoridad o la fe. El avance que supuso este nuevo enfoque fue extraordinario. Paralelamente, y también como reacción al pensamiento medieval, surgió el empirismo que defiende como conocimiento válido aquel que es obtenido a partir de los sentidos.

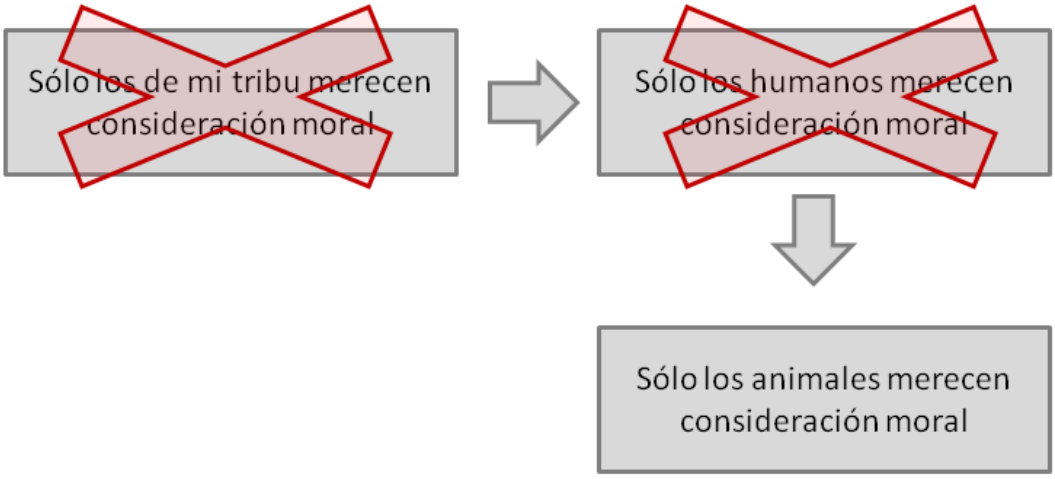


Fig. 3 Expansión del círculo moral. Primera aproximación.

Hemos pasado de "Sólo los de mi tribu merecen consideración moral" a "Sólo los humanos merecen consideración moral". La tribu es una versión extendida de la familia. Las tribus con sus dioses se enfrentan a otras tribus en conflictos en los que los vencidos son tomados como esclavos. Las mujeres y los niños son considerados más o menos una propiedad. Todo esto queda superado con el antropocentrismo, en el que todos los seres humanos, independientemente de su condición, son sujetos merecedores de derechos. Lo hacemos porque nos sentimos identificados con ellos, tenemos empatía por otros seres humanos, vemos que sienten y sufren como nosotros.

Pero el círculo moral se sigue ampliando. Nos hemos dado cuenta de que no sólo los humanos merecen consideración moral, sino también los animales, y precisamente por el mismo motivo: porque son capaces de sentir. No olvidemos que nosotros, los humanos, somos animales. Hay muchas especies animales que se nos parecen mucho, como los grande simios (chimpancés, bonobos, gorilas y orangutanes). Pero incluso también en mamíferos como perros y gatos somos capaces de observar comportamientos y emociones que nos resultan muy familiares, como alegría, tristeza, celos o miedo.

Fig. 4 René Descartes.
https://commons.wikimedia.org/wiki/File:Frans_Hals_-_Portret_van_Ren%C3%A9_Descartes_(cropped).jpg

Pero esto no ha sido siempre así. Se dice de Descartes, que vivió entre los siglos 16 y 17 (1596-1650), que consideraba a los animales no humanos como máquinas insensibles, llegando a pensar que los chillidos de un cerdo mientras le diseccionaban vivo no eran más que respuestas mecánicas como las de un mecanismo de relojería que necesita ser engrasado, sin un "alguien" que esté experimentando subjetivamente este sufrimiento. (*"Animals are nothing more than unconscious machines. Lacking consciousness, they lack reason or language."*). Aunque también pensó que los cuerpos de los humanos obedecían a las mismas reglas (*"The bodies of animals and men act wholly like machines and move in accordance with purely mechanical laws."*), en el caso de los humanos consideraba que había una diferencia, el alma, exclusivamente humana, lo que se interpreta como un "dualismo" pero exclusivamente en humanos.

Algunos académicos consideran que Descartes no negaba la capacidad de sentir de los animales:

"Descartes did not dev corollary of animal insensitivity"
https://www.jstor.org/stable/2220217?seq=1#page_scan_tab_contents

"In mechanizing the concept of living thing, Descartes did not deny the distinction between living and nonliving, but he did redraw the line between ensouled and unensouled beings. In his view, among earthly beings only humans have souls. He thus equated soul with mind: souls account for intellection and volition, including conscious sensory experiences, conscious experience of images, and consciously experienced memories. Descartes regarded nonhuman animals as machines, devoid of mind and consciousness, and hence lacking in sentience. (Although Descartes' followers understood him to have denied all feeling to animals, some recent scholars question this interpretation; on this controversy, see Cottingham 1998 and Hatfield 2008.)"
https://plato.stanford.edu/entries/descartes/

pero siendo defensor y practicante de la vivisección, sus argumentos fueron empleados para defenderla, así como para defender la idea de matar y comer animales:

"Vivisection was carried out by such ancient luminaries as Galen and there was a resurgence of the practice in early modern times (Bertoloni Meli 2012). Descartes himself practiced and advocated vivisection (Descartes, Letter to Plempius, Feb 15 1638), and wrote in correspondence that the mechanical understanding of animals absolved people of any guilt for killing and eating animals. Mechanists who followed him (e.g. Malebranche) used Descartes' denial of reason and a soul to animals as a rationale for their belief that animals were incapable of suffering or emotion, and did not deserve moral consideration — justifying vivisection and other brutal treatment (see Olson 1990, p. 39–40, for support of this claim). The idea that animal behavior is purely reflexive may also have served to diminish interest in treating behavior as a target of careful study in its own right."
https://plato.stanford.edu/entries/consciousness-animal/

Es interesante señalar que habitualmente empleamos la expresión "máquina" o "robot" para referirnos a un objeto sin capacidad de sentir, de la misma manera que a veces empleamos la palabra "animal" para referirnos a alguien basto, sin criterio ni habilidad. Sin embargo hay animales más inteligentes que bebés humanos, capaces de resolver problemas mucho mejor que ellos.

Creo que es adecuado emplear la palabra "máquina" o "robot" sin asumir implícitamente que las máquinas no sienten o que los seres que sentimos no somos máquinas. Por eso no digo que "se dice que Descartes consideraba a los animales no humanos como máquinas", sino como "máquinas insensibles".

Fig. 5 Una vivisección realizada sobre un cerdo.
https://commons.wikimedia.org/wiki/File:Galen-Pig-Vivisection.jpg

Hemos estado muy equivocados en el pasado y hemos cometido errores terribles por creer en ideas que eran muy intuitivas o que provenían de muy respetados científicos.

"In the sixteenth century, professors in European medical schools simply read a book by the ancient Greek physician Galen aloud while a surgeon showed students the relevant parts from the corpse of an executed criminal. The professors would never look at the cadavers, and the students barely would, because it was believed that everything worth knowing was in Galen's book.

Andreas Vesalius (1514-1564), a young medical professor, began dissecting corpses himself only to find that Galen was often very wrong. In Galen's culture, dissecting a human was taboo, and Vesalius finally determined that Galen had never dissected one! Vesalius made it his life's work to dissect human cadavers, and show medical students how the human body was actually structured rather than relying on ancient Greek texts.

Now, in part because of Vesalius, you live in a world of science-based and observation-based medicine. If you get sick, you will be treated with the best science has to offer, instead of with humors and leeches. Aren't you glad?"

http://study.com/academy/lesson/dissection-techniques-alternatives.html

Afortunadamente, los científicos modernos consideran que sin ninguna duda los animales no humanos sienten. Algunos como Richard Dawkins piensan que incluso podrían sentir más intensamente que los humanos.

"Would you expect a positive or a negative correlation between mental ability and ability to feel pain? Most people unthinkingly assume a positive correlation, but why?

Isn't it plausible that a clever species such as our own might need less pain, precisely because we are capable of intelligently working out what is good for us, and what damaging events we should avoid? Isn't it plausible that an unintelligent species might need a massive wallop of pain, to drive home a lesson that we can learn with less powerful inducement?

At very least, I conclude that we have no general reason to think that non-human animals feel pain less acutely than we do, and we should in any case give them the benefit of the doubt. Practices such as branding cattle, castration without anaesthetic, and bullfighting should be treated as morally equivalent to doing the same thing to human beings."

Richard Dawkins. Science in the Soul: Selected Writings of a Passionate Rationalist
https://boingboing.net/2011/06/30/richard-dawkins-on-v.html

Un hito importante en la consideración moral de los animales y en el reconocimiento de su capacidad de sentir es la Declaración de Cambridge.

Los científicos reunidos en Cambridge reconocen que ni el neocortex (la parte más moderna de la corteza cerebral, específica de homínidos como los humanos) ni la corteza cerebral (únicamente existente en mamíferos) son estructuras cerebrales necesarias para la existencia de la capacidad de sentir, de manera que los animales, incluyendo no sólo los mamíferos, sino también los peces, todos los vertebrados e incluso los invertebrados, poseen los sustratos neurológicos necesarios para generar la capacidad de sentir.

"Decidimos llegar a un consenso y hacer una declaración para el público que no es científico. Es obvio para todos en este salón que los animales tienen conciencia, pero no es obvio para el resto del mundo. No es obvio para el resto del mundo occidental ni el lejano Oriente. No es algo obvio para la sociedad."
Philip Low, en la presentación de la Declaración de Cambridge sobre la Conciencia, 7 de julio de 2012

«De la ausencia de neocórtex no parece concluirse que un organismo no experimente estados afectivos. Las evidencias convergentes indican que los animales no humanos tienen los sustratos neuroanatómicos, neuroquímicos, y neurofisiológicos de los estados de la conciencia junto con la capacidad de exhibir conductas intencionales. Consecuentemente, el grueso de la evidencia indica que los humanos no somos los únicos en poseer la base neurológica que da lugar a la conciencia. Los animales no humanos, incluyendo a todos los mamíferos y pájaros, y otras muchas criaturas, incluyendo a los pulpos, también poseen estos sustratos neurológicos.»
—Cambridge University, UK.

"en los últimos años la neurociencia ha estudiado las áreas del cerebro, descubriendo que las áreas que nos distinguen del resto de los animales no son las que producen la conciencia. Así, se deduce que los animales estudiados poseen conciencia porque "las estructuras cerebrales responsables por los procesos que generan la conciencia en los humanos y otros animales son equivalentes"

https://es.wikipedia.org/wiki/Declaraci%C3%B3n_de_Cambridge_sobre_la_Conciencia

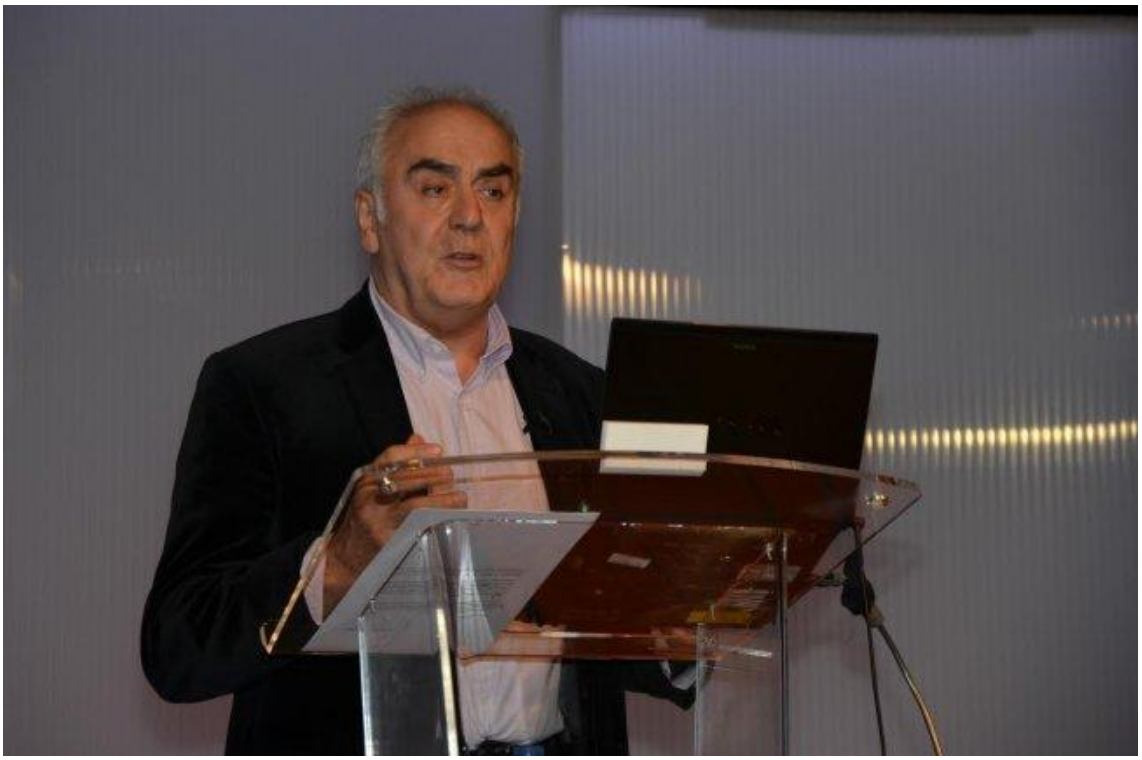


Fig. 6 Javier de Felipe en Futuro Singular 2015. Imagen cortesía de ATAM.

En 2015, Javier de Felipe, co-director del Human Brain Project mencionó que "Descartes estaba equivocado. Los animales no son máquinas. O si lo fueran, nosotros también lo seríamos". Es decir, Javier de Felipe reconoce la posibilidad de que tanto los humanos como los animales no humanos seamos máquinas, pero máquinas con capacidad de sentir.

"El principal objetivo del HBP es obtener simulaciones detalladas, desde el punto de vista biológico, del cerebro humano completo, así como desarrollar tecnologías de supercomputación, modelización e informáticas para llevar a cabo dicha simulación".
https://www.humanbrainproject.eu/

Javier de Felipe destacó que a medida que conocemos mejor el cerebro, éste va perdiendo su componente mágico y misterioso. Al entender cómo funciona, al ser cada vez más predecible, lo vemos como una máquina. Desde esta perspectiva, tanto los humanos como el resto de animales somos máquinas, seguramente producidas por la evolución.

Es interesante el fenómeno de atribuir cualidades mágicas o ultra-materiales (como la conciencia, la sintiencia o la voluntad) a aquello que no entendemos bien, y negarlas en aquello que sí entendemos. Esto puede producir un agravio comparativo.

```
// ---------------------------------------------------------------
// Ejemplos de reglas de tipo 1:
// ---------------------------------------------------------------
que tal?#como estas
como estas?#como estas
que tal estas?#como estas
como te encuentras?#como estas
que tal amigo?#como estas
como va eso?#como estas
// ---------------------------------------------------------------
// Ejemplos de reglas de tipo 2:
// ---------------------------------------------------------------
como estas#estoy muy bien, gracias
como estas#muy bien, que tal tu?
tema por defecto#vaya, vaya
tema por defecto#no se si te he entendido bien...
tema por defecto#jajaja
// ---------------------------------------------------------------
// Ejemplos de reglas de tipo 3:
// ---------------------------------------------------------------
hola, que tal?#bastante bien, y tu?
cuentame algo#sabes lo que estoy pensando?
dime algo#sabes lo que estoy pensando?
dime algo#crees que soy un robot?
dime algo#te puedo hacer una pregunta?
// ---------------------------------------------------------------
// Ejemplos de reglas de tipo 4:
// ---------------------------------------------------------------
crees que soy un robot?#si#por que?
te puedo hacer una pregunta?#si#que se siente al ser humano?
sabes lo que estoy pensando?#no#ah! pues yo creia que todos los humanos erais
adivinos...
```

Fig. 7 Ejemplos de reglas de un robot de charla

Este es un ejemplo de un robot de charla que construí en PHP hace unos años. El robot es muy sencillo, solo tiene cuatro tipos de reglas muy básicas y con unas 80 reglas logré que varias personas que entraban en esta página web creyeran, al menos en las primeras interacciones, que estaban realmente hablando conmigo. Es decir, las personas que interactuaron con este robot creían que había un ser consciente, sintiente al otro lado. Sin embargo, cuando estas personas conocieron el sistema y las reglas que definen el comportamiento de ese robot con el que interactuaron, tan sencillo, a nadie se le ocurrió seguir pensando que ese robot tenía consciencia, sintiencia o voluntad.

Algo parecido nos puede ocurrir a nosotros hacia otros animales más sencillos, y podría ocurrirle a un supuesto extraterrestre super-inteligente hacia nosotros.

Por ejemplo, ahora mismo los ordenadores son tan complejos que parece que tienen a veces un comportamiento impredecible. Sin embargo, disponemos de todo el conocimiento necesario para interpretar cada pequeño detalle, y si quisiéramos y pudiéramos invertir todo el tiempo y los recursos necesarios, podríamos al menos teóricamente entender y predecir cualquier comportamiento en un ordenador. ¿Es ese un criterio suficiente para considerar que los ordenadores no tienen consciencia, sintiencia ni voluntad?

Supongamos que con los avances en neurociencia, con el Blue Brain / Cajal Blue Brain / Human Brain Project en Europa (promovido por Henry Markram) o con el proyecto Brain en USA (liderado por Rafael Yuste), fuéramos capaces de entender y predecir perfectamente el comportamiento del sistema nervioso central de un animal cuya complejidad fuera inferior a la nuestra, como por ejemplo una hormiga (250.000 neuronas) o una rana (16 millones de neuronas), ¿deduciríamos que la rana o la hormiga no tienen conciencia, ni sintiencia ni voluntad?

En ese caso, si un extraterrestre super-inteligente, cuyo cerebro tuviera, digamos por ejemplo, 500 billones de neuronas (teniendo nosotros los humanos 86 mil millones de neuronas), fuera capaz de entender perfectamente el comportamiento de nuestro sistema nervioso humano y hacer predicciones fiables sobre él ¿debería deducir que el humano no tiene conciencia, ni sintiencia ni voluntad?

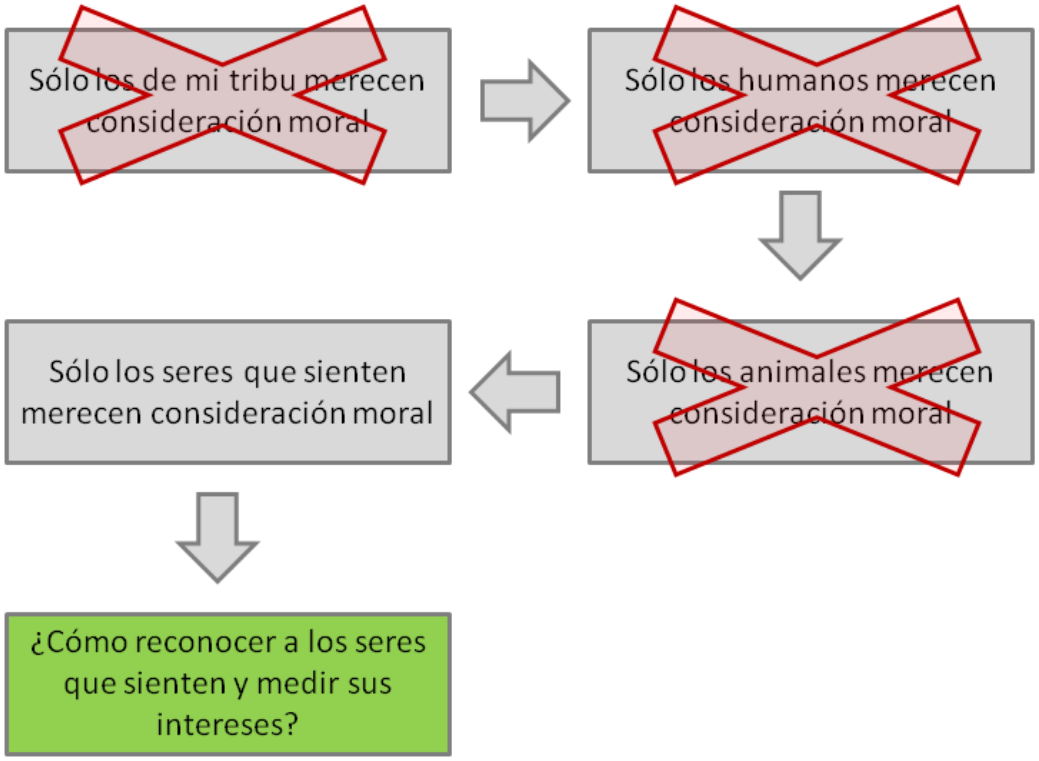


Fig. 8 Expansión del círculo moral. Segunda aproximación.

Dado que el criterio para recibir consideración moral es la capacidad de sentir, podemos y debemos pasar de la afirmación "Sólo los animales merecen consideración moral" a "Sólo los seres que sienten merecen consideración moral". De esta forma no dejaremos fuera de la consideración moral a otros seres que puedan sentir, tal vez robots, o quién sabe si las plantas.

Los seres que sienten tienen intereses, por ejemplo, interés en disfrutar e interés en evitar el sufrimiento. Por supuesto, muchos intereses están enfrentados. Llegados a este punto, lo moralmente relevante ser capaces de saber quiénes son los seres que pueden sentir, y ser capaces de medir sus intereses, para poder aplicar unos recursos limitados en satisfacer estos intereses, además de para poder resolver conflictos de interés.

"Los intereses no deben ser frustrados" (Respuestas Veganas)
"Principio de igual consideración de intereses" (Peter Singer)

Podemos hacer una categorización de los intereses y experiencias en tres dimensiones:

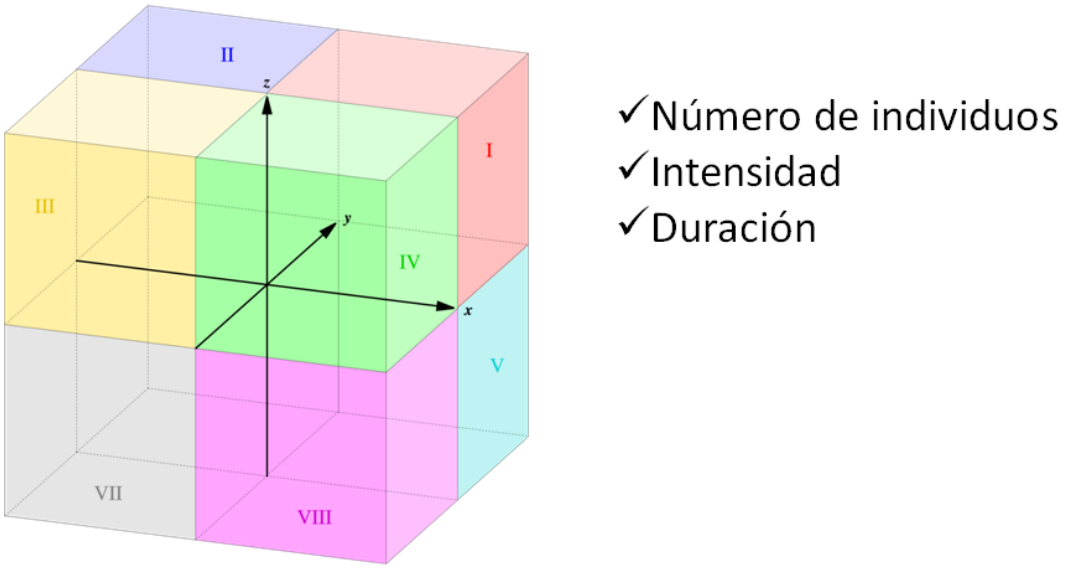


Fig. 9 Categorización de las experiencias en tres ejes.

Los intereses son relativos a experiencias presentes o del futuro. Los seres tienen interés en que sucedan ciertas cosas, y dejen de suceder otras.

A las experiencias que no deseamos podríamos darles un valor negativo, y a las deseables podríamos darles un valor positivo.

Todas las experiencias se podrían representar en un espacio tridimensional donde cada punto sería una experiencia, y los ejes serían el número de individuos que la experimentan, la intensidad (que podría ser positiva para cosas deseables y negativa para las indeseables) y la duración de la experiencia.

Cada experiencia, como por ejemplo romperse un tímpano por accidente, podríamos descomponerla en una serie de sub-experiencias de distinta intensidad y duración. Como aproximación, podríamos decir que el primer minuto es de una intensidad muy dolorosa seguida de unos 5 minutos de una intensidad algo menor, etc. hasta la recuperación completa del tímpano.

Si tuviéramos este mapa de todas las experiencias podríamos concentrar nuestros recursos en tratar de prevenir las peores experiencias. Esto es precisamente lo que trata de hacer la organización con la que colaboro: OPIS (Organisation for the Prevention of Intense Suffering) y por eso trabajamos en un proyecto para crear un mapa o sistema de visualización que transmita la dimensión de todo el sufrimiento que existe.

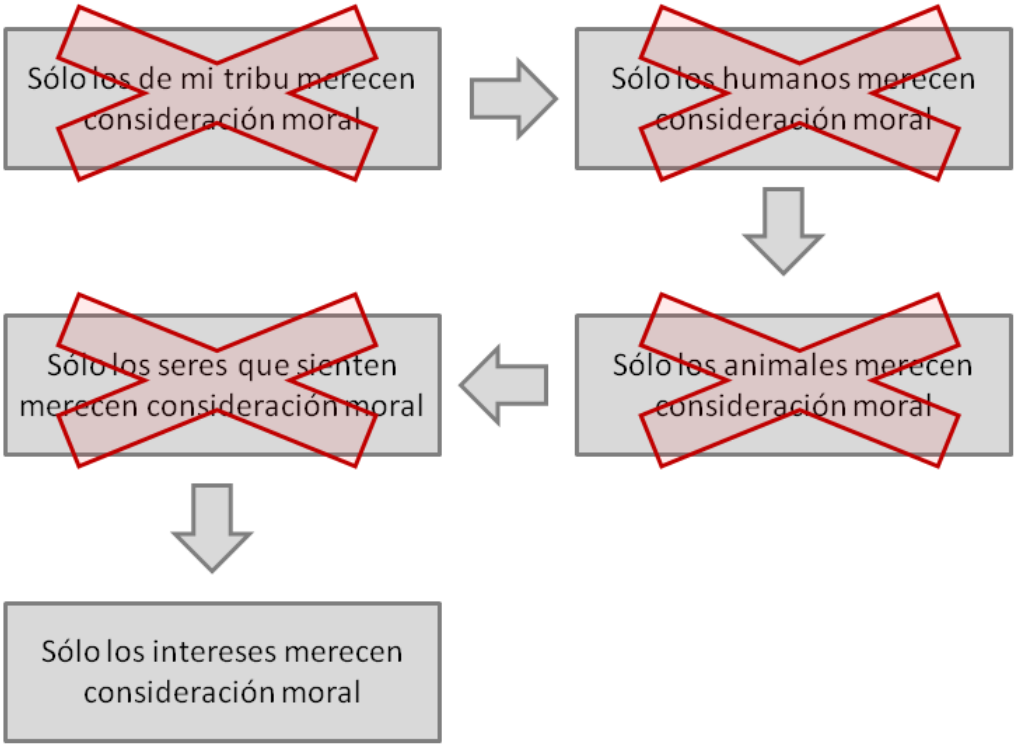


Fig. 10 Expansión del círculo moral. Tercera aproximación.

A esta reflexión todavía le podemos dar una vuelta de tuerca más. Ya que lo relevante son las experiencias y los intereses, no es necesario identificar seres, sino intereses. Es decir, en realidad no es necesario establecer donde empieza y acaba un ser, basta con saber cuáles son sus intereses (o sus posibles experiencias). Esto es interesante porque pueden existir seres sin contornos bien definidos, algo alejados de nuestra idea habitual de "individuo" o "ser" según la cual está muy claro dónde empieza y dónde termina (en el tiempo y el espacio) cada ser.

¿Cómo es posible la existencia de seres sin contornos bien definidos? Veámoslo con dos ejemplos.

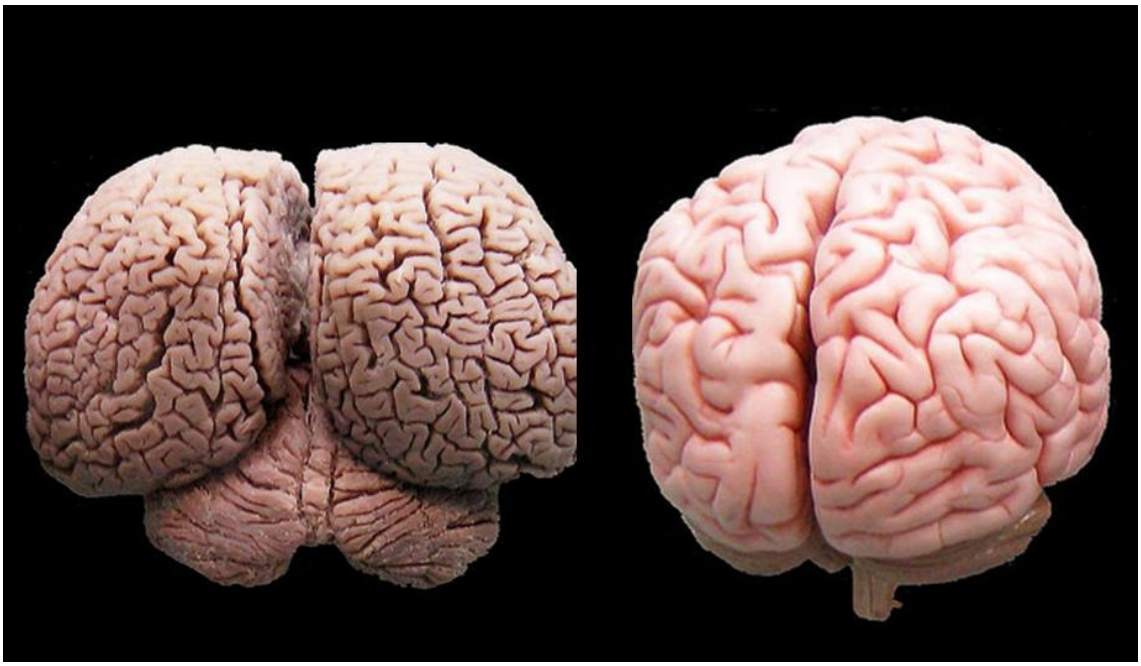


Fig. 11 Cerebros delfín y humano

En la foto de la figura 11 podemos ver un cerebro humano y un cerebro de un delfín. Se dice que los delfines duermen cada vez con un lado del cerebro, mientras siguen nadando. Los dos hemisferios del cerebro del delfín se muestran más "desconectados" que los del humano. Cuanto más separación existiera entre esos dos hemisferios, realizando funciones y controlando comportamientos independientes cada uno de ellos, tanto más plausible parece la idea de que en ese cerebro existan dos seres distintos.

Fig. 12 Cerebros conectados

Por el contrario, si fuéramos capaces de unir nuestras mentes, enlazando nuestros cerebros, tendiendo cables entre ellos y realizando funciones progresivamente más coordinadas, a medida que realizáramos estas conexiones resultaría cada vez más creíble pensar que estos dos cerebros puedan generar una única sintiencia o consciencia.

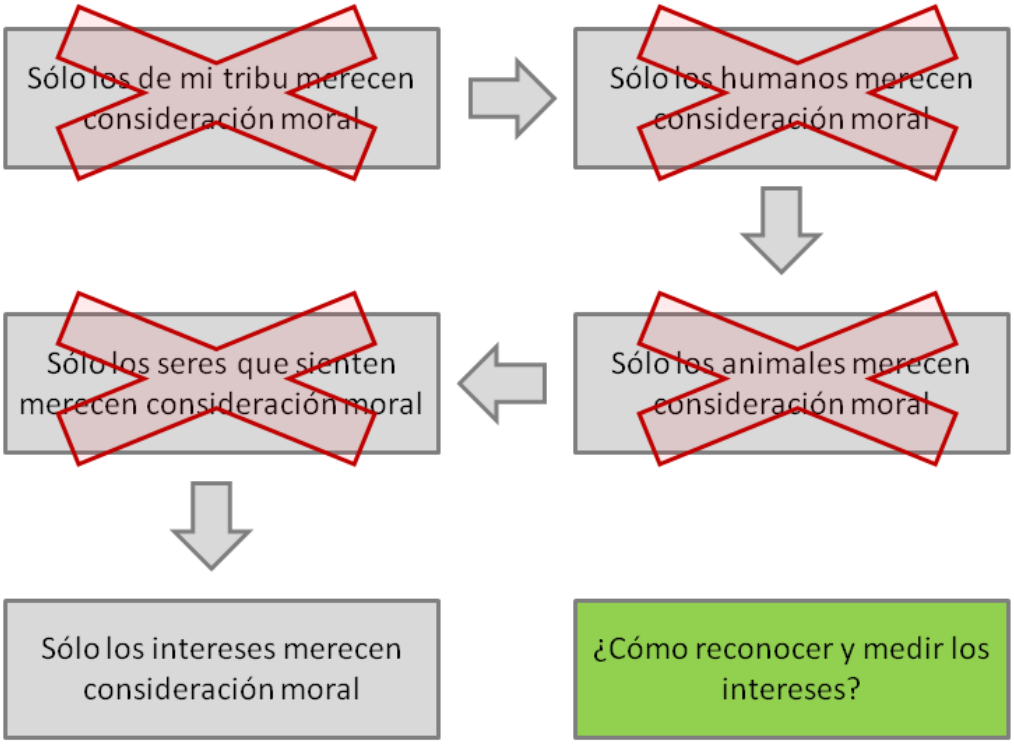


Fig. 13 Expansión del círculo moral. Cuarta aproximación.

En resumen, y dado que la "regla de oro" es que "Los intereses no deben ser frustrados" tal como menciona David Díaz en "Respuestas Veganas", la prioridad moral es considerar los mayores intereses, y no es necesario establecer dónde empieza y dónde acaba un ser. No es necesario identificar seres, sino intereses, y esto es muy relevante para el caso de la sintiencia en las máquinas ya que la sientiencia en las máquinas se puede producir, no en una máquina físicamente bien identificable, con contornos bien delimitados, sino en un sistema complejo, como por ejemplo Internet o un sistema compuesto de multitud de pequeños elementos (robots, programas etc.)

Sensocentrismo es el planteamiento ético que considera que lo relevante para establecer consideración moral es la capacidad de sentir.

*Llamaré **sensocentrismo convencional** a la postura moral que considera que la capacidad de sentir es el criterio relevante para establecer quiénes son los seres que merecen consideración moral.*

*Llamaré **sensocentrismo estricto** a la postura moral que considera que lo relevante para establecer consideración moral es la capacidad de sentir. O lo que es lo mismo: lo relevante para establecer consideración moral son los intereses.*

La primera definición asume implícitamente que existen seres, y que esos seres son distintos entre sí. La segunda no lo hace y es coherente con el individualismo abierto y el individualismo vacío.

Referencias:
https://es.wikipedia.org/wiki/Sensocentrismo
http://www.redfilosofica.org/definiciones.php#sensocentrismo-estricto
http://www.redfilosofica.org/patrones_coherencia_y_los_lugares_extranos_del_sensocentrismo.php

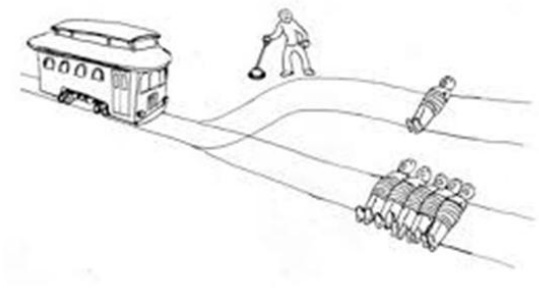
Fig. 14 Sensocentrismo estricto

En este sentido, y para aquellos a quienes les gusta el concepto de sensocentrismo, creo que deberíamos distinguir entre un "sensocentrismo convencional" que considera que la capacidad de sentir es el criterio relevante para establecer *quiénes* son los seres que merecen consideración moral, y un "sensocentrismo estricto" que considera que lo relevante para establecer consideración moral es la capacidad de sentir, sin necesidad de introducir el concepto de individuo, que puede ser problemático.

En realidad, en este contexto las ideas de "sentir" y "tener intereses" son equivalentes. Pero "interés" no es lo mismo que "deseo". Por ejemplo, yo deseo caminar hacia la máquina expendedora para comprar una chocolatina, pero desconozco que la chocolatina está envenenada y me produciría una dolorosa muerte. En este caso es moralmente deseable frustrar mi deseo de comprar la chocolatina impidiéndome el paso, incluso con cierta violencia, porque mi interés realmente es evitar comerme esa chocolatina, aunque yo no lo sepa.

¿Por qué es importante la sintiencia de las máquinas? Porque la creación de máquinas capaces de sentir, en un proceso exponencial, por ejemplo, creando máquinas capaces de sentir que a su vez sean capaces de crear otras máquinas capaces de sentir, puede provocar una catástrofe moral de dimensiones astronómicas.

Trolley problem. El dilema del tranvía.



- ✓ Agentes racionales
- ✓ Agentes egoístas
- ✓ La muerte no es deseable
- ✓ Solución universal, con independencia de mi lugar en el conflicto

El velo de la ignorancia

⇨

Egoísmo velado

Fig. 15 El dilema del tranvía

El dilema del tranvía ilustra la dificultad para lograr un consenso a la hora de tomar decisiones morales. El dilema queda perfectamente reflejado por la imagen de la figura 15 y no requiere de mucha más explicación.

La solución al dilema del tranvía, suponiendo que somos agentes racionales, egoístas, suponiendo que la muerte no es algo deseable, y suponiendo que tratamos de buscar una solución universal (con independencia de mi lugar en el conflicto), es conocida como "El velo de la ignorancia", aunque la expresión es algo redundante y podría ser más adecuado llamarla "El egoísmo velado" o "Egoísmo ciego", ya que se trata de suponer que "yo" soy cualquiera de las seis personas atrapadas en las vías. Suponiendo ciertas todas las premisas anteriores, no tendré dudas en elegir que el tranvía se dirija por la vía por en la que sólo hay una persona atrapada, ya que tendría cinco veces más probabilidad de morir en el caso contrario.

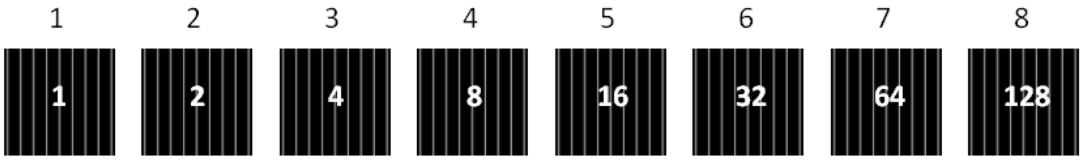


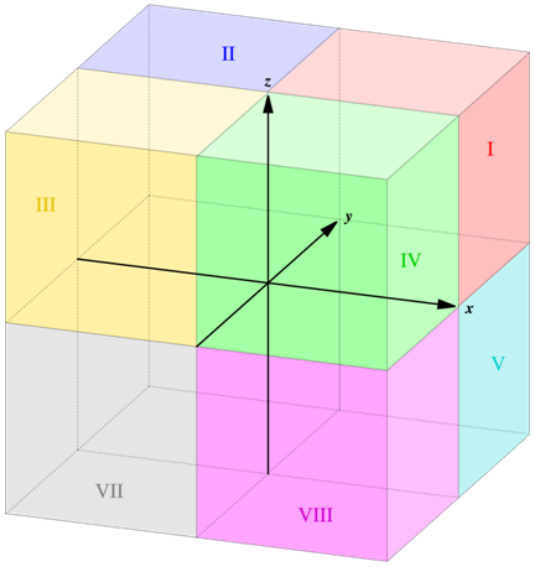
Fig. 16 Filósofos malvados

El dilema del tranvía quizás no sea tan habitual en nuestra vida diaria. El siguiente dilema (Los filósofos malvados) tal vez represente mejor la situación en la que nos encontramos. Existen ocho celdas donde los filósofos malvados nos han encerrado. En la primera celda nos dan una bofetada al día (algo muy asumible), en la segunda dos (también), en la tercera cuatro (empieza a ser molesto, hay que reconocerlo) y así sucesivamente mediante potencias de 2, de forma que en la octava celda el prisionero recibe 128 tortazos al día, lo que ya casi podría considerarse una paliza.

Las celdas están desordenadas de manera que no sabemos en qué celda se dan más bofetadas.

Los filósofos malvados nos proponen lo siguiente: puedes elegir una de las celdas y en ella se reducirá a la mitad el número de tortazos. Esta situación refleja muy bien el escenario un moral muy común: tenemos unos recursos limitados y si somos capaces de descubrir dónde esos recursos son más necesarios y efectivos, podremos reducir considerablemente el sufrimiento, aunque no eliminarlo por completo.

Categorización de intereses / experiencias en tres dimensiones



✓Número de ~~individuos~~ ocurrencias
✓Intensidad
✓Duración

Fig. 17 Categorización de intereses / experiencias en tres dimensiones

En ese espacio tridimensional, y con el objetivo de ser lo más efectivos, sería interesante conocer dónde se encuentra el máximo número de ocurrencias de sufrimiento, de mayor intensidad y duración, para colocar ahí nuestros recursos limitados. Y según ciertas teorías de la sintiencia, es muy posible que en un futuro próximo dichas ocurrencias de sufrimiento ocurran en seres que consideramos "máquinas.

¿Qué es el Anti-substratismo?
"Antisubstratismo" es equivalente a "Antiespecismo", referido en este caso a la idea de substrato en vez de a la idea de especie. Es injustificado discriminar moralmente según el substrato que soporta la consciencia (entendida en este caso como la capacidad de sentir, de tener intereses), lo mismo que es injustificado discriminar moralmente según especie (especismo), raza (racismo), sexo (sexismo), etc.

¿Por qué es especialmente relevante ahora la sintiencia de las máquinas?

Singularidad tecnológica

La singularidad tecnológica es el advenimiento hipotético de inteligencia artificial general (también conocida como "IA fuerte", del inglés *strong AI*). La singularidad

tecnológica implica que un equipo de cómputo, red informática, o un robot podrían ser capaces de auto-mejorarse recursivamente, o en el diseño y construcción de computadoras o robots mejores que él mismo. Se dice que las repeticiones de este ciclo probablemente darían lugar a un efecto fuera de control -una explosión de inteligencia, en donde las máquinas inteligentes podrían diseñar generaciones de máquinas sucesivamente cada vez más potentes, la creación de inteligencia muy superior al control y la capacidad intelectual humana.
La singularidad tecnológica ocasionará, según Albert Cortina y Miquel-Ángel Serra, cambios sociales inimaginables, imposibles de comprender o de predecir por cualquier humano. En esa fase de la evolución se producirá la fusión entre tecnología y también inteligencia humana, en donde la tecnología dominará los métodos de la biología hasta dar lugar a una era en que se impondrá la inteligencia no biológica de los posthumanos, que se expandirá por el universo.

https://es.wikipedia.org/wiki/Singularidad_tecnol%C3%B3gica

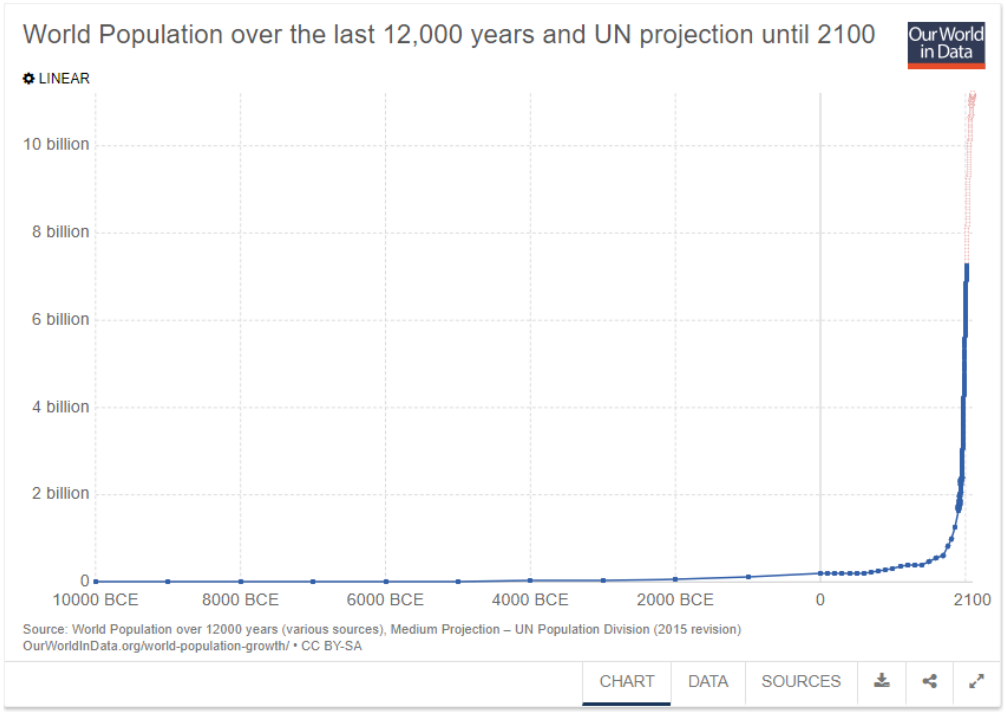
La singularidad tecnológica es el advenimiento hipotético de una inteligencia artificial general (también conocida como "IA fuerte", del inglés *strong AI*). Si esa IA fuera capaz de mejorarse a sí misma, de forma recursiva, será cada vez más inteligente. Si la Inteligencia Artificial es cada vez más inteligente, llegará un momento en que iguale las capacidades humanas. Esa será la última vez que sepamos qué es lo que está haciendo. A partir de ese momento, no entenderemos lo que la IA está haciendo y no podremos controlarla. Todos los esfuerzos para mantener bajo control a la IA deben realizarse ahora, antes de que llegue ese momento.

Riesgo de experimentar un gran sufrimiento

Entre detenidos desaparecidos, ejecutados, torturados y presos políticos, el número de víctimas de la dictadura de Pinochet superó las 40.000 personas. Más de 4.000 personas sufrieron torturas en Euskadi en los últimos 50 años, según un informe. Desde el golpe de Estado de Egipto, 60.000 personas han sido detenidas y muchas de ellas torturadas. Más de 11.000 niños han muerto en la guerra civil de Siria y cientos de ellos han sido asesinados o torturados. Se barajan cifras de entre decenas de miles o cientos de miles de personas que sufrieron de alguna manera u otra la Inquisición. Unos 50.000 pacientes mueren al año en España con sufrimiento evitable, por no tener acceso a unos cuidados paliativos. Cada día más de 2000 niños de todo el mundo mueren en dolorosos accidentes. Sólo en un año y sólo en la Unión Europea se sacrificaron 252 millones de cerdos. El 77% de estos cerdos son castrados sin anestesia. Durante un año se producen 140.000 experimentos con animales no humanos en España en los que el animal muere o sufre un gran daño.

http://www.manuherran.com/la-gran-mentira/

El hecho de que la IA sea muy exitosa no quiere decir que esté exenta de sufrimiento. Que la especie humana domine sobre las otras especies no implica que los humanos no suframos. Si la IA se extiende y multiplica por el universo, y esta IA sufre, puede producirse la mayor catástrofe moral de toda la historia.

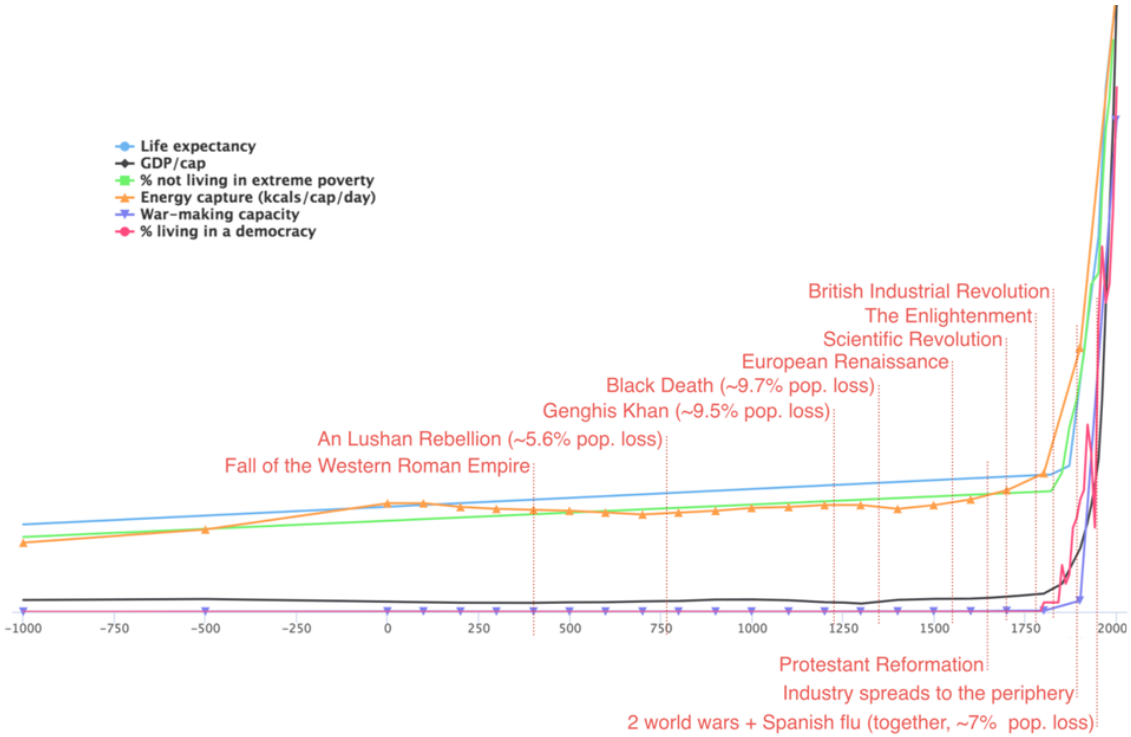


https://ourworldindata.org/world-population-growth/

Fig. 18 Crecimiento exponencial (I)

Actualmente estamos viviendo varios fenómenos exponenciales, como el incremento de la población humana. Algo similar a lo mencionado con las máquinas está seguramente ocurriendo ya con los animales empleados para uso y consumo humano.

Si los humanos siguen comiendo animales, la población de animales seguirá el crecimiento exponencial de la de humanos. Seguramente el balance neto medio de la felicidad de los humanos y de los animales no humanos haya mejorado en los últimos años, pero aun habiendo mejorado, es posible que esa valoración neta sea negativa. Esto ocurre cuando las vidas tienen un balance neto negativo (hubiera sido mejor no vivir esa vida). Si ese fuera el caso, al extender la vida humana por la galaxia estaríamos incrementando el sufrimiento global. De la misma forma, si las máquinas tienen experiencias y éstas son por lo general negativas, estaremos expandiendo el sufrimiento tal vez de una forma terrible. Y la única forma de pararlo sería ahora, antes de que la situación escape de nuestro control.



http://blog.jessriedel.com/2017/10/19/links-for-october-2017/
http://blog.jessriedel.com/wp-content/uploads/2017/10/all-curves-with-events.png
http://lukemuehlhauser.com/three-wild-speculations-from-amateur-quantitative-macrohistory/

Fig. 19 Crecimiento exponencial (II)

La figura 19 muestra otros datos relacionados con la calidad de vida y el consumo de energía, que también siguen un patrón exponencial.

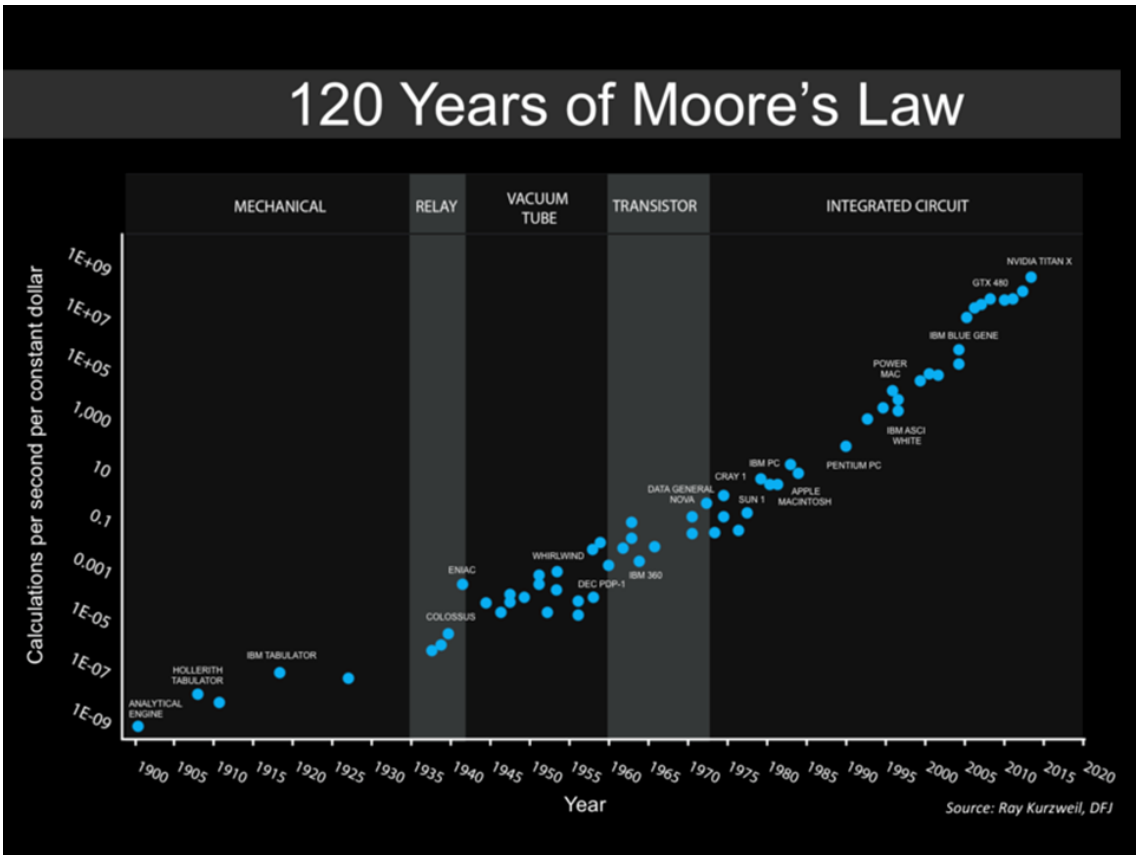


Fig. 20 Crecimiento exponencial (III)

La ley de Moore expresa que aproximadamente cada dos años se duplica el número de transistores en un microprocesador. La escala del eje vertical es logarítmica: la gráfica es exponencial aunque no lo parezca.

Fig. 21 Pong

Este cambio exponencial también lo hemos experimentado muchos de nosotros, que hemos visto como en unos pocos años, los videojuegos han pasado del sencillo Pong…

Fig. 22 Videojuegos y realidad virtual

...a simulaciones tremendamente detalladas, algunas de ellas inmersivas que podríamos denominar "realidad virtual".

Fig. 23 Personajes de videojuegos

En un contexto de incremento exponencial de la complejidad y el detalle de las simulaciones, y suponiendo que la complejidad fuera condición suficiente para la emergencia de la consciencia ¿podrían algún día los personajes de los videojuegos adquirir sintiencia y reclamar nuestra ayuda?

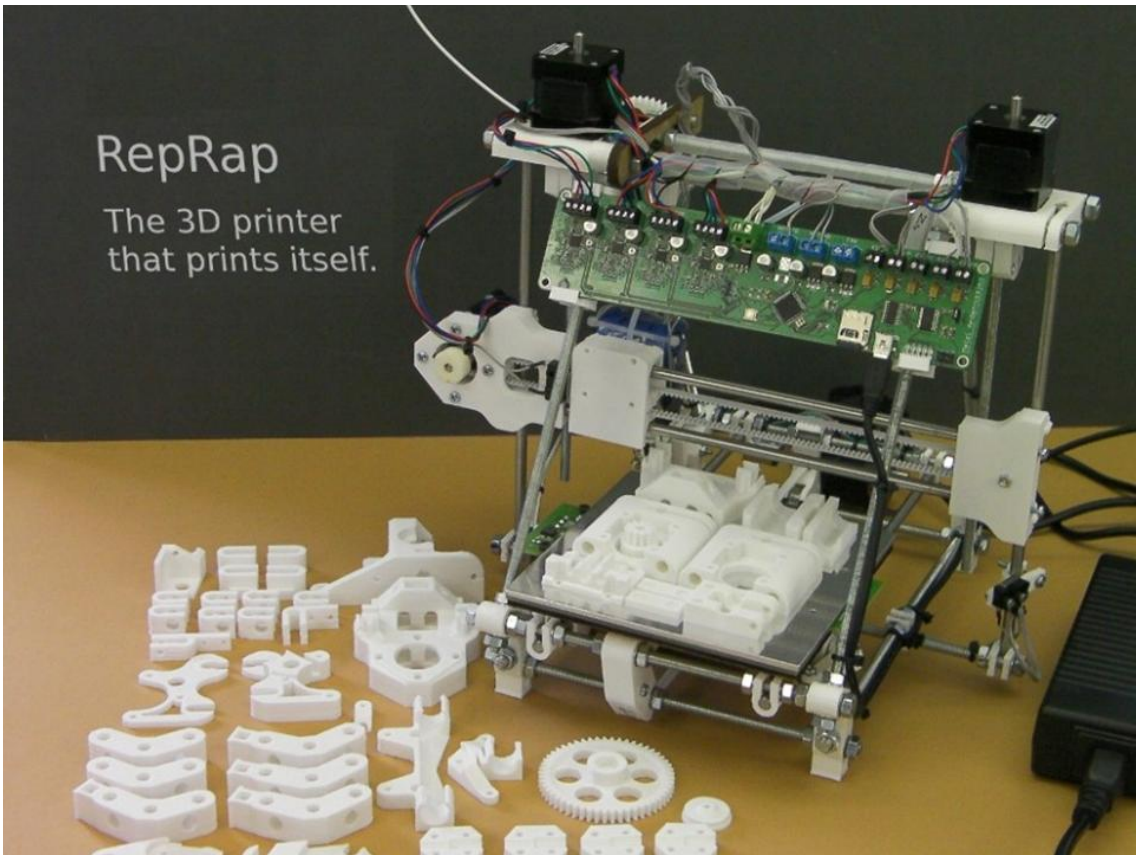


Fig. 24 Impresoras 3D

Las impresoras 3D representan un tipo de tecnología con la que pueden producirse fenómenos exponenciales incontrolables, de tipo vírico.
Supongamos que enviamos a la luna una impresora 3D que es capaz de tomar materia de su entorno y con ella fabricar piezas, las mismas piezas con las que podría copiarse a sí misma.

A partir de una única impresora 3D, con la primera copia tendríamos 2, luego 4, 8, 16 y así sucesivamente. Llegaría un momento en que los materiales disponibles en la superficie de la luna empezarían a escasear.

Podemos suponer que ocasionalmente se produzcan errores en las copias. No todas las impresoras serán iguales. Estos errores podrían ser indiferentes, ser perjudiciales de forma que dificulta a la impresora seguir copiándose a sí misma, o incluso beneficiosos. Gracias a estos errores, algunas impresoras podrían especializarse en encontrar materiales en las capas más profundas, o cierto tipo de materiales. Algunas podrían desarrollar habilidad para robar piezas de otras máquinas y otras desarrollar escudos que les protejan de estos ataques. También podrían desarrollar pactos de colaboración entre ellas, intercambios de piezas etc.

IA (la película)

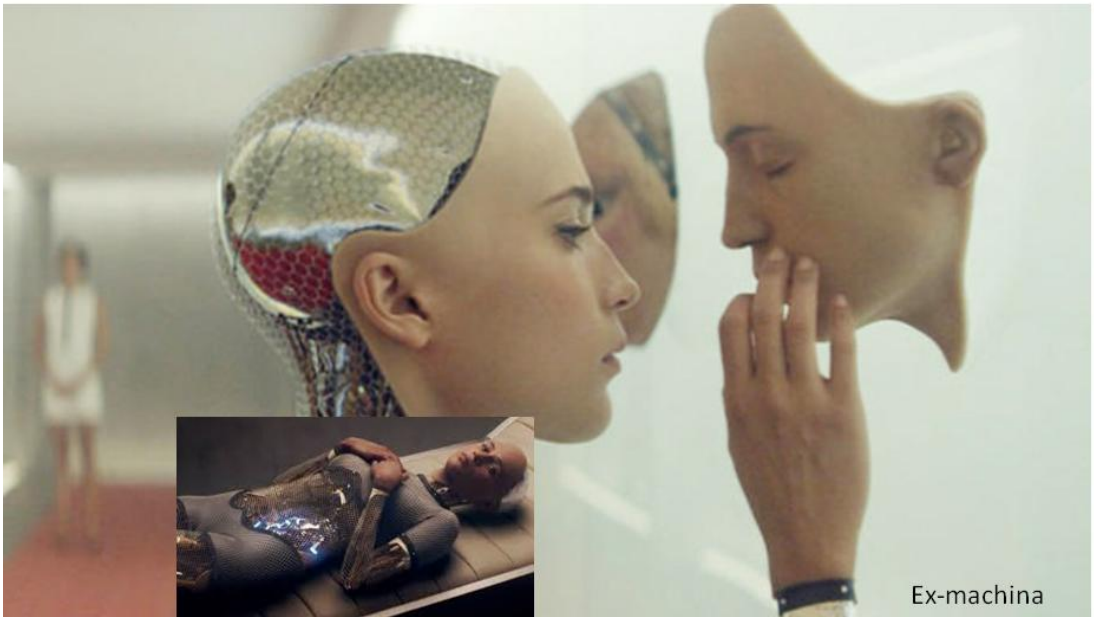


Fig. 24 Películas de ciencia-ficción

Multitud de películas de ciencia-ficción han desarrollado la idea de la empatía con las máquinas, y la posibilidad de que tengan sentimientos.

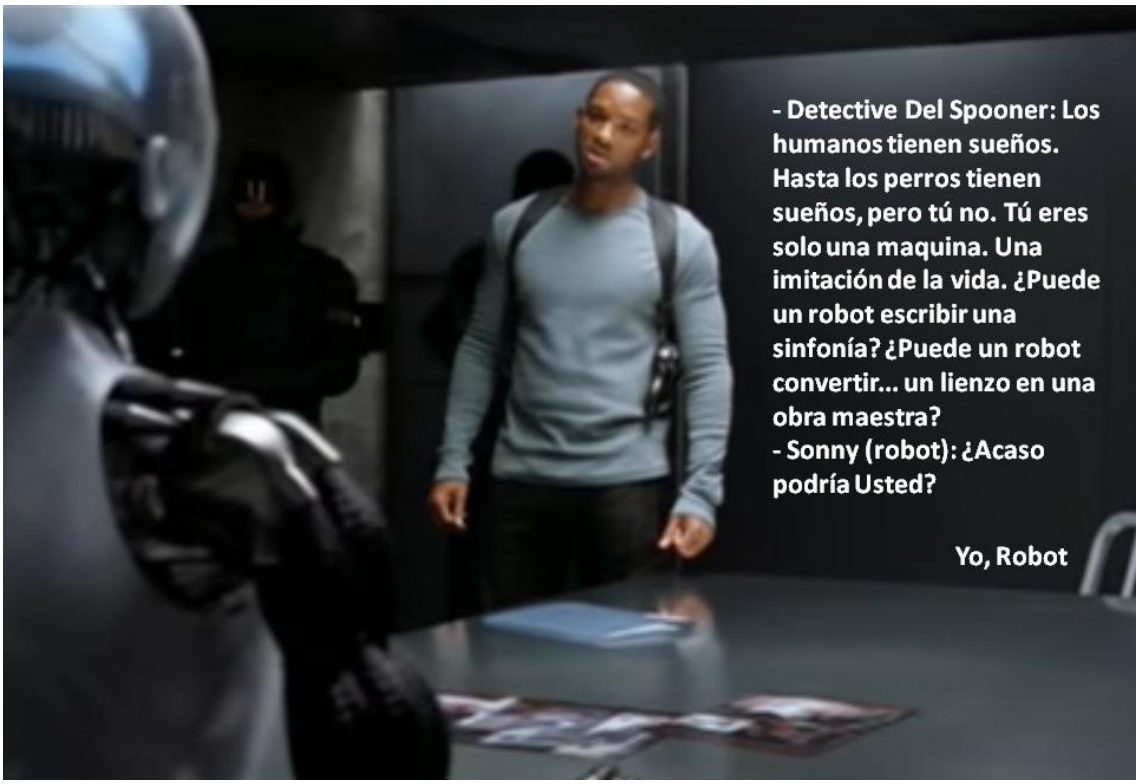


Fig. 25 Yo, Robot

En escenarios en los que los seres humanos sienten aversión a las máquinas…

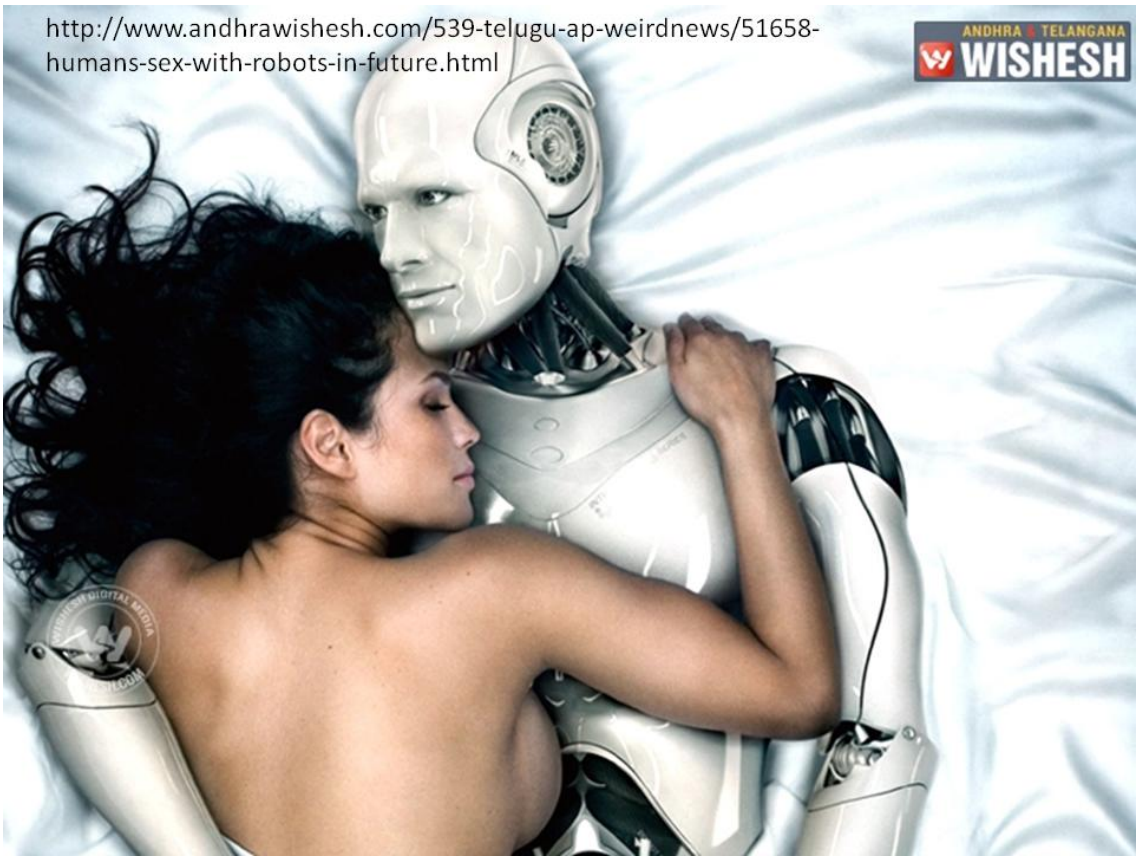


Fig. 26 Amor con robots

…pero también podríamos amarlas y enamorarnos de ellas.

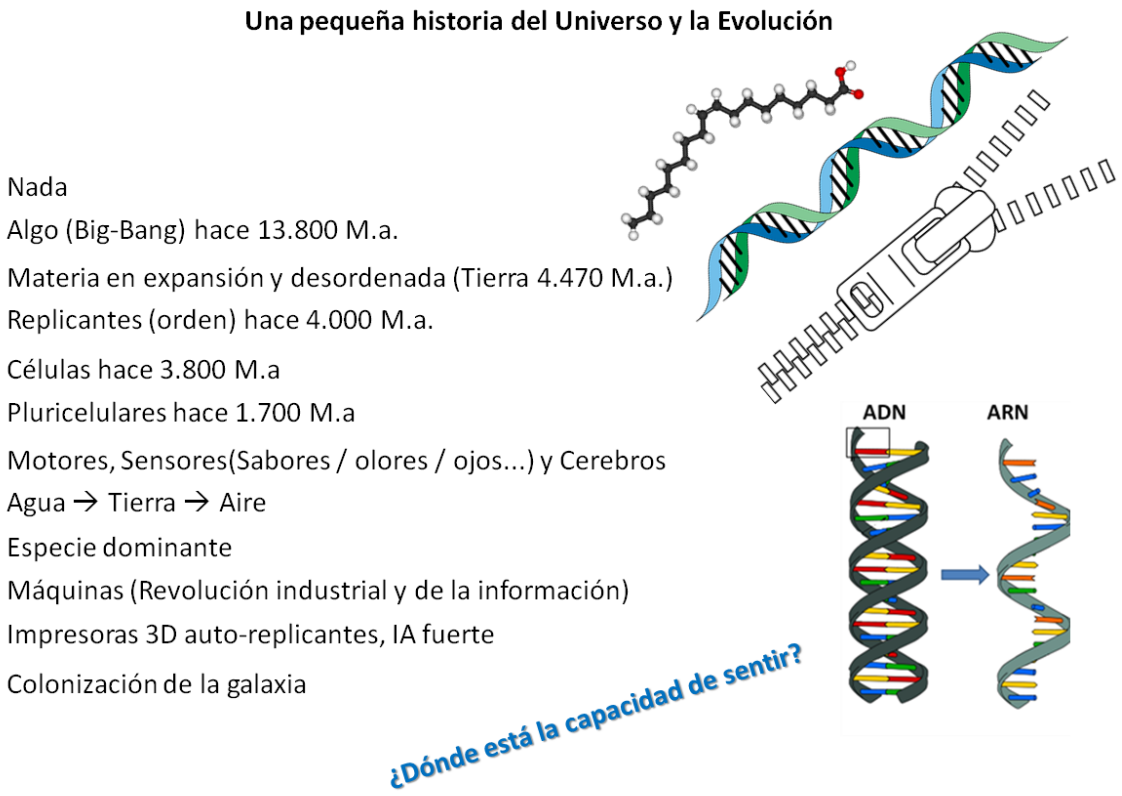


Fig. 27 Pequeña historia del Universo y la Evolución

Repasar las mejores teorías que tenemos acerca del origen de todo lo que existe nos ayudará a entender cómo y por qué nosotros podemos ser máquinas, y cómo es posible que las máquinas sientan. Para ello voy a hacer de forma muy simplificada una pequeña historia del Universo y de la Evolución.

Al principio no había nada. Si damos por válida la teoría del Big-Bang, hace 13.800 millones de años, la nada "explotó". La materia en expansión y desordenada formó el planeta Tierra hace unos 4.470 millones de años.

Hace 4.000 millones de años, en este planeta Tierra comenzó un proceso de orden: empezaron a surgir ciertos replicantes. Se trata de moléculas en forma de hilos (en este caso fueron hilos pero supongo que tal vez podrían haber sido anillos, superficies, mallas o cualquier otra forma) que tienen la propiedad de emparejar en forma de llave-cerradura materia de forma simétrica, de manera que a partir de un único hilo se forman dos hilos que son el uno espejo del otro. Bajo ciertas condiciones, estos dos hilos simétricos se separan, abriéndose como una cremallera, de manera que ahora cada uno de esos dos hilos puede volver a atraer materia formando de nuevo el doble hilo. A partir de un hilo, y disponiendo de materia suficiente, tendríamos dos, después cuatro, ocho y así sucesivamente.

Podemos suponer que en el proceso de copia se producían errores, lo que genera diversidad. También podemos suponer que llegó un momento en que la materia empezó a escasear. Los hilos podrían "robarse" materia unos a otros. Precisamente para evitar ser robado, algunas copias podrían haber desarrollado una suerte de escudo o capa protectora: la célula. Por supuesto, esta capa protectora no es algo que alguien desarrolló intencionadamente. Simplemente, por error en el proceso de copia, algunas cadenas produjeron capas protectoras, más o menos rudimentarias,

mientras que otras no lo hicieron. Aquellas que se protegieron se reprodujeron más que las que no lo hicieron.

En este proceso algunos replicantes con la misma información genética empezaron a actuar conjuntamente, de forma colaborativa. Se trata de la aparición de los seres pluricelulares, hace 1.700 millones de años. Inicialmente por simple azar, y después seleccionado por su utilidad reproductiva, los conglomerados de células desarrollaron especializaciones. Unas células se especializaron en el desplazamiento, pues estando en movimiento es más fácil encontrar los nutrientes necesarios para reproducirse. Para hacerlo en un líquido, parece buena idea desarrollar una hélice. Y mucho mejor si pudiéramos tener algún tipo de sensores (receptores de moléculas, es decir, "olores" o receptores de fotones, es decir "ojos") que nos indiquen hacia dónde ir o de dónde escapar, además de un cerebro o sistema con el que procesar toda esta información.

Estos seres pluricelulares poblaron primero el mar, después la superficie de la tierra, y más tarde incluso el cielo, desarrollando la capacidad de volar.

Entre ellos, surgió un tipo de especie, un tipo de organismo pluricelular, que es ahora la especie dominante. Pudieron ser dos o cuarenta, pero el caso es que actualmente es una, con diferencia, la especie dominante: la humana. Esta especie ha vivido una revolución industrial y ahora se encuentra viviendo una revolución de la información que podría desembocar en la creación de una Inteligencia Artificial Fuerte y posiblemente la colonización de la galaxia, no sabemos si por seres humanos, por máquinas construidas por los humanos o por una combinación de ambos.

En esta historia de la vida no ha aparecido en ningún momento la sintiencia. No es necesaria para explicar la evolución de la materia hacia la vida. Todo el fenómeno de la evolución se explica aludiendo a las propiedades físicas de la materia y nada más. La evolución no requiere de la sintiencia.

Por tanto, tenemos que combinar y hacer coherentes dos ideas: la primera, que la sintiencia no es necesaria en la evolución. La segunda: que la sintiencia existe. Podemos establecer diversas hipótesis acerca del origen de la capacidad de sentir. La siguiente figura trata de organizarlas en cuatro grandes grupos.

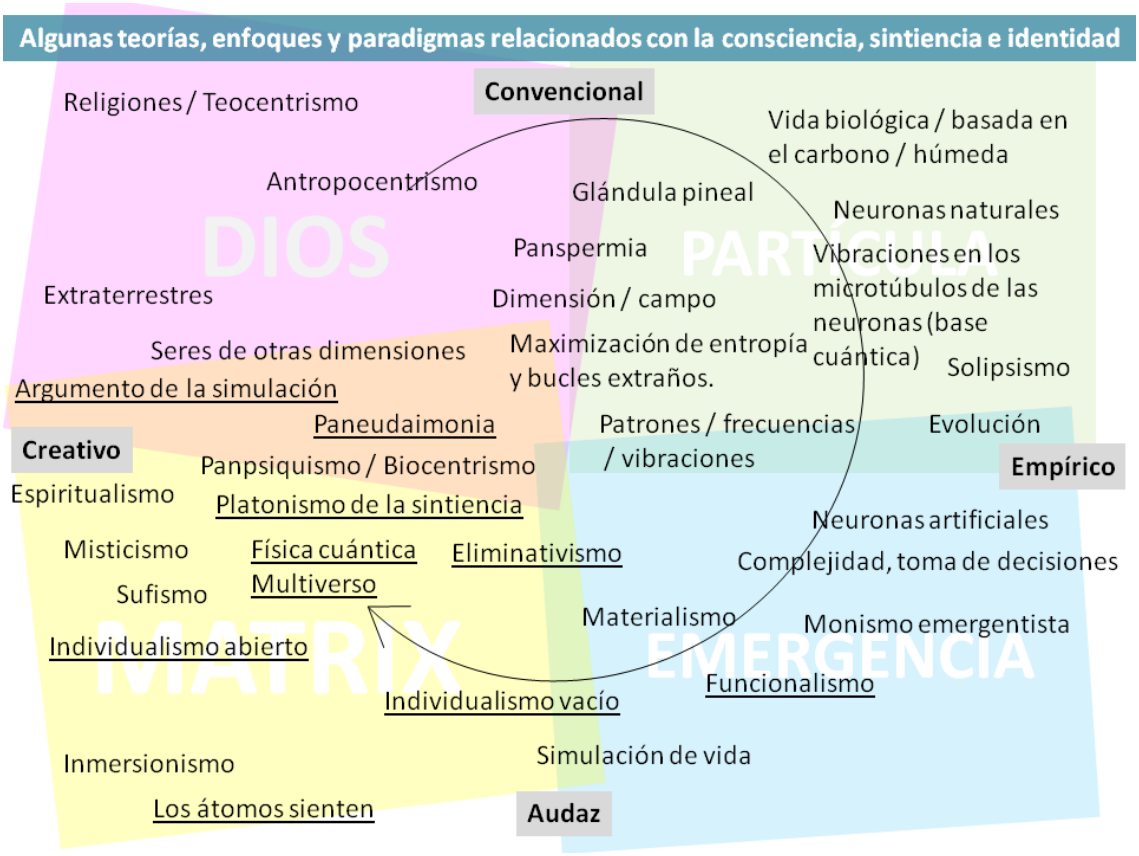


Fig. 28 Algunas teorías, enfoques y paradigmas relacionados con la consciencia, sintiencia e identidad

En esta imagen trato de presentar de forma global algunas de las teorías y enfoques sobre la sintiencia / consciencia / identidad, y en general, sobre la realidad, agrupadas en cuatro grandes grupos. La enumeración no es exhaustiva y varias de estas teorías podrían estar clasificadas en más de un grupo a la vez. He tratado de dar significado a la posición de cada etiqueta, aunque en algunos casos no ha sido fácil decidir dónde colocarla.

El primer grupo son las que llamo teorías o visiones del mundo de tipo **DIOS**, que hacen referencia a seres o realidades superiores a la nuestra, y que de alguna forma la determinan, como por ejemplo las religiones.

El segundo grupo lo llamo **PARTÍCULA** y se trata de aquellas teorías o hipótesis que consideran que para la existencia de la capacidad de sentir es necesario algún componente (por lo general, material) en particular, como por ejemplo, componentes biológicos, húmedos, basados en el carbono.

El tercer grupo de teorías son las **EMERGENTISTAS**, las más populares entre los científicos modernos, que consideran que, partiendo de una base material, la sintiencia emerge si se dan ciertas condiciones.

El cuarto grupo lo denomino **MATRIX**, porque según estas teorías nada es lo que aparece y ponen en duda nuestras intuiciones sobre la sintiencia y sobre la realidad en general.

Las teorías de la parte superior son las más **CONVENCIONALES** mientras que las de la parte inferior son las más **AUDACES**. He tratado de colocar a la derecha aquellas teorías con un enfoque más **EMPÍRICO** y a la izquierda las más **CREATIVAS**.

Tanto históricamente como a nivel personal, no es extraño observar una evolución de las creencias en el orden indicado, que he ilustrado con una flecha: DIOS, PARTÍCULA, EMERGENCIA y MATRIX. De alguna forma, este recorrido intelectual vuelve al punto de partida.

Si me preguntasen acerca de la probabilidad que asigno a cada uno de los cuatro tipos de teorías, yo diría que algo así como: 1%, 25%, 75% Y 99%.

En cuanto a los modelos morales (prototipos) de cada cuadrante, creo que más o menos podrían ser de la siguiente forma

Cuadrante 1 (DIOS)

El prototipo moral de quienes tienen estas creencias es el de personas solidarias, altruistas, preocupadas por los derechos humanos, en contra de la tortura y de la pena de muerte y que colaboran con organizaciones humanitarias. Consideran y valoran por igual a todos los seres humanos independientemente de su inteligencia, cultura, país, edad, identidad sexual, preferencias sexuales, preferencias políticas, raza, color de la piel, capacidades, etc. Son contrarios a la experimentación (involuntaria y perjudicial) con seres humanos.

Cuadrante 2 (PARTÍCULA)

Estas personas comparten las preocupaciones morales del cuadrante 1, pero además incluyen a todos los animales con sistema nervioso central. Son defensores de los derechos de los animales. Tratan de minimizar el sufrimiento de todos los seres que sienten. Son contrarios a la experimentación con animales, y también con sistemas neuronales biológicos, ya que éstos podrían generar sintiencia y sufrimiento.

Cuadrante 3 (EMERGENCIA)

Además de asumir las posiciones morales de los cuadrantes 1 y 2, estas personas consideran la posible emergencia de sintiencia en máquinas y por tanto los derechos robots, simulaciones informáticas y en general, software, que haya sido construido de forma similar o bajo condiciones similares a aquellas bajo las cuales hemos sido construidos nosotros, los seres biológicos que sentimos. En concreto previenen del riesgo implícito en la construcción de sistemas físicos o digitales muy complejos, capaces de razonar y/o capaces de evolucionar.

Cuadrante 4 (MATRIX)

Quienes consideran estas hipótesis, además tener en cuenta las tres posiciones morales descritas anteriormente, tienen en cuenta otras posibilidades relacionadas con la física y la filosofía del sufrimiento que pueden ser muy poco intuitivas e incluso podrían considerarse improbables, pero cuyas implicaciones en relación a la prevención del sufrimiento, en caso de ser ciertas, serían inmensas; y por tanto consideran moralmente correcto y necesario dedicar al menos una parte de los recursos disponibles a investigar acerca de estas posibilidades.

La respuesta a la pregunta acerca de la posible sintiencia en máquinas, según cada uno de los cuadrantes, con matices, me parece que sería la siguiente:

Cuadrante 1 (DIOS)

"La pregunta es absurda. Las máquinas no pueden sentir. Los animales no humanos podrán hacerlo, pero no es muy relevante, ya que el único ser relevante es el ser humano, hecho a imagen y semejanza de Dios. La humana es la especie elegida, el pueblo elegido, ungido de divinidad, lo que le legitima para usar a los animales en su provecho y por supuesto, también a las máquinas".

Cuadrante 2 (PARTÍCULA)

"Las máquinas secas, hechas de metal y plástico, no pueden sentir. En cambio una máquina biológica, construida mediante células artificiales, sí podría hacerlo".

Cuadrante 3 (EMERGENCIA)

"Nosotros los humanos, así como el resto de animales y todos los seres vivos, somos, en definitiva, máquinas. Por tanto, lo que se conoce como robots, y en general las máquinas construidas por humanos e incluso simulaciones artificiales podrán sentir si se dieran ciertas condiciones de complejidad y evolución en un entorno adecuado, como ha ocurrido con nosotros, los animales".

Cuadrante 4 (MATRIX)

"No sólo los robots podrían sentir. Es que los átomos y hasta las ideas podrían sentir. No entendemos bien la realidad y no sabemos lo que es posible".