

Tres problemas filosóficos de la Inteligencia Artificial

Tradicionalmente el problema filosófico de la IA, se ha planteado en términos de la posibilidad de que puedan construirse subjetividades sintéticas que tengan las mismas propiedades que tienen los *homo sapiens*. Esta pregunta ha generado inmensos debates, sobre los que al presente no hay consenso básico, y todo parece en disputa. Este es el problema filosófico clásico de la IA. Ahora bien, además de este problema y sus derroteros, hay al menos otros dos problemas filosóficos de la IA, que son independientes del primero, de modo que podemos hablar de tres problemas filosóficos de la IA:

1º Problema IA: Este es el problema tradicional que se expresa en la pregunta ¿Pueden fabricarse máquinas dotadas de las propiedades típicas de la subjetividad humana -en particular aquello que se llama pensar?

2º Problema IA: Esta pregunta es independiente de la anterior y de hecho es compatible con una respuesta negativa al 1º Problema: supongamos que en algún entorno físicamente posible pero completamente distinto del entorno material del planeta Tierra y sistema solar, haya entidades no orgánicas, que adquieran información del exterior en entornos cambiables, puedan resolver problemas generados en esos entornos, puedan proveer estados posteriores de dichos entornos, puedan tomar decisiones en algún sentido sub-óptimas sobre sus interacciones en dichos estados posteriores de dichos entornos y puedan fabricar réplicas de sí, la cuestión sería ¿Pueden estas entidades entrar en procesos de evolución por selección natural, que culminen en subjetividades mecánicas?

3º Problema IA: Este problema es independiente de los dos problemas anteriores, y es relevante aunque se diera una respuesta negativa a ellas. Supongamos una respuesta negativa al 1º Problema, que efectivamente que no se puedan construir los robots sugeridos por la literatura de CF, pero que sí se puedan construir artefactos mecánicos que sean **zombis filosóficos**, que puedan remedar todos los estados externos y todas las operaciones de un humano, incluso todos los modos de interacción verbal. En ese caso ¿cómo serían los zombis filosóficos y las formas de nuestra interacción e involucramiento, aceptado el hecho que dichos zombis filosóficos sintéticos mecánicos, son tan carentes de subjetividad como un lavarropas, pero que, a diferencia de los lavarropas, serían capaces de interacciones que serían exteriormente indiscernibles de la capacidad de un humano y podrían tomar decisiones que en algunos aspectos pudieran superar a las decisiones humanas -como ser el caso de los deportistas sintéticos en el go o el ajedrez?

Estos tres problemas conceptuales, son diferentes entre sí, en modo que ninguno de ellos presupone la respuesta afirmativa a otro. En particular el segundo problema no supone el primero, en tanto el primero es un problema de ingeniería -¿podremos construir tales mecanismos de ser posibles? y el segundo un problema de filosofía de la biología o de biología teórica ¿pueden operar los mecanismos de evolución -la selección natural o alguno otro que se pueda postular- operar sobre entidades mecánicas, entidades electromagnéticas o sólo podrían operar sobre organismos?

Estos tres problemas son autónomos de lo que podríamos llamar un problema conexo, el de la hibridación de organismos y mecanismos, que configura lo que se ha denominado la agenda transhumanista, y que plantea preguntas como:

¿Cuáles son los límites fácticos de la intersección entre sistemas mecánicos digitales y los sujetos corpóreos?

¿Cuáles serían los límites éticos en la realización de la intersección de sistemas mecánicos y sujetos corpóreos?

Estos dos problemas surgen de la posibilidad de existencia de entidades híbridas organismos-mecanismos o cyborgs, pero suponen otra exploración, ya que en este caso el 1º Problema no sería relevante, ya que la intencionalidad o subjetividad de los híbridos, provendría de la parte orgánica humana.

Primer problema: ¿Pueden pensar los <algunos tipos de> mecanismos y sus diversas encarnaciones?

Una forma de abordar el 1º Problema es a partir del Mito de Pigmalión, o sea la materia inerte que se torna organismo; la versión clásica de dicho mito la ofrece Ovidio:

La dorada Venus, tal como asistía en persona a su festividad, se dio cuenta de cuál era la intención de esa súplica y, como augurio de una divinidad amiga, por tres veces se encendió la llama y llevó por el aire su extremo. Cuando regresó, buscó aquél la estatua de su amada y, recostándose en el lecho, la besó; le pareció que estaba tibia; acerca de nuevo la boca, también palpa el pecho con sus manos: el marfil palpado se reblandece y, perdiendo su rigidez, se amolda a los dedos y cede, como se ablanda la cera del Himeto bajo el sol y, ablandada por el pulgar, se adapta a muchas formas y se hace útil por su propio uso. Mientras se queda atónito y se alegra con dudas y teme engañarse, una y otra vez el enamorado vuelve a tocar con la mano el objeto de su deseo; era de carne y hueso: latan las venas al contacto del pulgar. (Ovidio, *Metamorfosis*, Libro X 275-290)



Pygmalion de Jean-Baptiste Regnault de 1786, Musée National du Château et des Trianons.

En este mito surge la cuestión: **¿Cómo es que la materia inerte puede recibir intencionalidad por fuera de la evolución por selección natural sino por diseño -que en este caso sería la mediación de Atenea?** El tema es la construcción de intencionalidad equivalente a la humana, por medio de operaciones técnicas.

Podemos ir exponiendo algunas de las cuestiones disputadas que corresponden al 1º Problema.

a. La salvedad de William Paley

No es común encontrar a William Paley (Inglaterra 1743 – 1805) en los debates filosóficos sobre la IA, pero sí es común encontrarlo en los debates filosóficos sobre la evolución por selección natural y su rechazo fideísta. Sin embargo, el punto por el cual el autor es débil en este segundo plano -la defensa de los seres sobrenaturales como efectores de la realidad orgánica- es, a nuestro juicio, muy relevante en la filosofía de la IA. Y es una simple salvedad, que no puede soslayarse en el análisis de las máquinas: las máquinas tienen constructores.

Establezcamos una diferencia básica entre mecanismos y organismos:

ORGANISMO	MECANISMO
Tiene partes Principio de integración endógeno (Inserción)	Está compuesto de partes Principio de integración externa.
Explicación interna (variaciones heredables) top-down + externa bottom (medio ambiente)-up	Explicación externa bottom-up
De los principios de orden a las partes vinculadas por relaciones funcionales o de consecuencia	De las partes a los ensambles de partes vinculadas por relaciones mereológicas y causales según diseño
Relación sistémica → Totalidad: Humboldt: visión holística Bernard: medio interno & homeostasis Darwin: Selección natural No diseño	Relación externa: Instrumento ajustado a un fin técnico Sí Diseño + Valor
Mundo físico no antrópico	Mundo antrópico

En su *Natural Theology, or Evidences of the Existence and Attributes of the Deity collected from the Appearances of Nature* Paley intentaba defender el clásico argumento sobrenatural del diseño. Y para ello introduce la llamada analogía del relojero:

In crossing a heath, suppose I pitched my foot against a stone, and were asked how the stone came to be there; I might possibly answer, that, for any thing I knew to the contrary, it had lain there forever: nor would it perhaps be very easy to show the absurdity of this answer. But suppose I had found a watch upon the ground, and it should be inquired how the watch happened to be in that place; I should hardly think of the answer which I had before given, that, for any thing I knew, the watch might have always been there. Yet why should not this answer serve for the watch as well as for the stone? why is it not as admissible in the second case, as in the first? For this reason, and for no other, viz. that, when we come to inspect the watch, we perceive (what we could not discover in the stone) that its several parts are framed and put together for a purpose, e. g. that they are so formed and adjusted as to produce motion, and that motion so regulated as to point out the hour of the day; that, if the different parts had been differently shaped from what they are, of a different size from what they are, or placed after any other manner, or in any other order, than that in which they are placed, either no motion at all would have been carried on in the machine, or none which would have answered the use that is now served by it... There must have existed, at some time, and at some place or other, an artificer or artificers, who formed [the watch] for the purpose which we find it actually to answer; who comprehended its construction, and designed its use. ... Every indication of contrivance, every manifestation of design, which existed in the watch, exists in the works of nature; with the difference, on the side of nature, of being greater or more, and that in a degree which exceeds all computation. (*Paley Natural Teology*, p.1–2)

El texto de Paley presenta dos tesis diferentes: una sobre la filosofía de la tecnología y otra de filosofía de la naturaleza. La crítica se centró, y con buenos argumentos diríamos, en la segunda tesis: que la Naturaleza y sus integrantes -los organismos- son artefactos tecnológicos diseñados por seres sobrenaturales. Y esta tesis ha sido bien contestada por los cosmólogos y los biólogos evolucionistas.

Pero a la primera tesis pareció no diferenciarse de la segunda. En lenguaje vernacular la primera tesis dice:

Todo mecanismo tiene diseñador y procede de procesos de diseño.

Esta tesis es una obviedad; pero no sus consecuencias: no se puede describir al mecanismo sin referencia explícita al diseñador y los procesos de sesgo. Un objeto tecnológico es un producto bajo condiciones económicas, materiales, de usuarios, y de un creador o de un equipo creador o de un proceso de creadores y modificadores y mejoradores.

Así hablamos de motor Diesel, motor Stirling, motor Wankel, motor Otto, motor en cada caso el tipo de motor se designa con el nombre del ingeniero o fabricante que ora lo inventó, ora toma una idea no *industrializada* y la hizo efectiva. A tal punto que en la Modernidad el criterio de identificación de un mecanismo es la patente de invención:

There is evidence of a few patents in the 13th century. By the 15th there is a steady flow of patent awards in Venice. The patent was a limited monopoly awarded for the possession of locally novel craft knowledge or for the invention of new devices... n 1416, one Franciscus Petri, from the island of Rhodes, petitioned the Council of Venice for a patent for "structures with pestles for fulling fabrics." It was granted "because they full thus perfectly, better than the fullers existing in the waters surrounding" the Rialto. For fifty years, no one but Petri and his heirs were permitted to build or have built similar fulling devices, nor could any additions or reconstructions be made of any part other than what he provided. 875, 877 (Long, 1991, págs. 875, 877)

Máquina y patente de invención, o sea individuo o corporación cuyas actividades están casualmente conectadas con la existencia de la máquina, son co-sustanciales, más allá que nuestro discurso ordinario sólo habla de las máquinas. Un objeto tecnológica tiene una intencionalidad humana ínsita; ahora del hecho que nuestras descripciones locales de un objeto tecnológico, un mecanismo, no hagan referencia a los creadores, a los procesos de sesgo que operan sobre ellos, y son parte de la constitución del objeto tecnológico, no significa que no estén, sino que nuestras descripciones ordinarias de los objetos tecnológicos son falsas.

Y esto se aplica a toda máquina u objeto tecnológico. Desde los simples motores de combustión hasta las computadoras y los robots, como ser los equipos de diseñadores de Boston Dynamics, que fabrican unos extraordinarios perros robots.

Y ello tiene una consecuencia central: al hablar de subjetividades mecánicas, en rigor la descripción completa debería ser subjetividades mecánicas & subjetividad orgánicas que las fabrican. Y en ello, la debilidad del argumento de Paley en la filosofía de la naturaleza, aparece como una salvedad fundamental en la filosofía de la tecnología.

Si la salvedad de Paley es meramente una tesis sobre la identificación de máquinas y objetos técnicos, sus consecuencias son centrales en el abordaje del 1º Problema. Tradicionalmente los problemas filosóficos de la Inteligencia Artificial parten de una formulación equívoca:

IA1 ¿Es posible la IA?

Pero no es claro qué sea esta pregunta. Porque si la cuestión depende de la definición de inteligencia, puede ser que haya definiciones comunes a la psicología y la ingeniería de IA, así si Inteligencia = Resolución de problemas + aprendizaje heurístico, la pregunta sería:

IA2 ¿Es posible la existencia de mecanismos que resuelvan problemas + aprendan heurísticamente?

Mi punto de vista es que la cuestión IA2 ya ha sido respondida, de largo, y que no puede haber ninguna respuesta negativa general, ni ningún argumento que pueda poner un límite a los posibles mecanismos futuros que resuelvan problemas y aprendan heurísticamente, no me imagino qué premisas tendría que tener un argumento de ese tipo; es más creo que las posibles premisas para dicho argumento tendrían flaquezas importantes.

Ahora, supongamos que Inteligencia = Resolución de problemas + Aprendizaje heurístico + Autoconciencia reflexiva & intencional. En ese caso la pregunta cambia radicalmente:

IA3 ¿Es posible la existencia de mecanismos que resuelvan problemas + aprendan heurísticamente+ tengan autoconciencia reflexiva & intencional?

Esta pregunta es muy distinta de la anterior, pero no solo una pregunta sobre las aplicaciones de *inteligencia*, sino sobre nuestras fundamentales acerca de la subjetividad. Es decir **¿qué significa tener autoconciencia reflexiva & intencional?**

Como sea que se abarque el tema en la filosofía la mente, hay una serie de rasgos que asumimos como básicos de la autoconciencia reflexiva & intencional, sin ánimo de ser exhaustivos se pueden señalar:

- auto percatarse, ser consciente de sí mismo y recursivamente ser consciente de ser consciente de sí.
- dar cuenta de los estados mentales propios y de sus contenidos o rasgos.
- dar cuenta de una estructura temporal de consciencia pretérita y presente.
- poder vincular discursivamente las decisiones tomadas con contenidos proposicionales precedentes.

-poder vincular discursivamente los estados emocionales presentes con estados vinculares y estados pretéritos.

-dar cuenta de una identidad corpórea propia. (Kriegel, 2009)

La lista se puede extender. Básicamente recoge los aspectos intencionales, inferenciales y emocionales de una subjetividad corpórea asociada a una trayectoria temporal.

Históricamente un modo en que se ha dado esa cuestión es tratar si un sistema puramente material -no interesa si mecánico u orgánico- puede presentar dichos rasgos o no. En los siglos XVI y XVII se conforma una metafísica rigurosamente materialista, que unifica a organismos y mecanismos como compuestos de materia gobernados por leyes físicas comunes. Así señalaba Gómez Pereira (España, 1500-circa 1558):

Cuando se dice que la facultad de sentir es común a los brutos y a los hombres, no se da a entender otra cosa que los brutos ven —lo que ven— como lo vemos nosotros, y que, incluso, sienten tantas y tan grandes diferencias de numerosas formas de color, y de otras sensibilidades comunes e individuales, en la medida que nosotros las percibimos con el puro conocimiento. Porque si hubiera sido posible que los brutos hablaran, éstos habrían denominado al color que ven blanco como «album» —si hablaran en latín— o «blanco» —si hablaran en español. Incluso si pudieran tocar una figura cuadrada, la habrían denominado con el concepto latino «quadratum» o con la expresión española «figura cuadrada». Y de la misma manera se servirían de los demás sentidos externos (Pereira, 2000, pág. 2)

Y Thomas Hobbes (Inglaterra 1588-1679) va por los mismos senderos:

Por razonamiento entiendo la computación. Y computar es *hallar la suma de varias cosas añadidas a conocer lo que queda cuando de una cosa se quita otra*. Por lo tanto, razonar es lo mismo que *sumar* y *restar*, y si alguien añade a esto *multiplicar* y *dividir*, no estoy de acuerdo, ya que la *multiplicación* es la *suma* de cosas iguales, y la *división*, la *resta* de cosas iguales cuantas veces se pueda hacer. Por tanto todo razonamiento se reduce a estas dos operaciones de la mente: la *suma* y la *resta* <...> no hay que pensar que la computación, esto es, el razonamiento, tenga lugar solamente en los números como si el hombre se distinguiera de los demás animales solo por su facultad de numerar (como se cuenta que opinaba *Pitágoras*) ya que se pueden sumar y restar magnitud a magnitud, cuerpo a cuerpo, tiempo a tiempo, grado a grado de calidad, concepto a concepto, proporción a proporción, oración a oración, nombre a nombre (en lo cual se contiene todo género de Filosofía). . (*De Corpore*, I.1. #2)

En este punto Hobbes afirma el carácter mecánico o más precisamente *algorítmico* de la racionalidad; y en el resto de esa sección I, desarrolla muy bien la analogía de la argumentación/computación con el álgebra de los conceptos.

En contra de esa visión argumentó Gotfried W. Leibniz: (Electorado de Sajonia, 1646-1716), quien sostiene:

Hay que reconocer, por otra parte, que *la percepción* y lo que de ella depende resultan *inexplicables per razones mecánicas*, esto es, por medio de las figuras y de los movimientos. Porque, imaginémonos que haya una máquina cuya estructura la haga pensar, sentir y tener percepción; se la podrá concebir agrandada, conservando las mismas proporciones, de tal manera que podamos entrar en ella como en un molino. Esto supuesto, una vez dentro, no hallaremos sino piezas que se impelen unas a otras, pero nunca nada con que explicar una percepción. Así, pues, esto hay que buscarlo en la sustancia simple, no en lo compuesto o en la máquina. Más aún, no cabe hallar en la sustancia simple otra cosa excepto esto, es decir, excepto las percepciones y sus cambios. Y solamente en esto también pueden consistir todas las *Acciones internas* de las sustancias simples. (*Monadología*, # 17)

1. Alvin Plantinga

En la filosofía actual, Alvin Plantinga (EE.UU 1932-) ha tomado el argumento de Leibniz y -en un artículo ya clásico (Plantinga, 2006) lo ha generalizado en dirección de argumentar en contra de que las entidades materiales -mecánicas u orgánicas- puedan tener subjetividad:

Leibniz está pensando en interacciones mecánicas, interacciones que implican empujes y tirones, engranajes y poleas, cadenas y piñones. Pero creo que diría lo mismo de otras interacciones estudiadas en física, por ejemplo, las que involucran la gravedad, el electro-magnetismo y las fuerzas nucleares fuertes y débiles. ¹ (Plantinga, 2006, pág. 12).

A partir de esa demarcación, Alvin Plantinga expone un argumento contra el materialismo, a partir de una tesis contra una teoría materialista del contenido i.e.

"... un objeto material... un átomo, o está compuesto de átomos. Así, los átomos, las moléculas, las células, los corazones, los cerebros y los cuerpos humanos son todos objetos materiales ... es extremadamente difícil ver cómo

¹ Alvin Plantinga, *Against Materialism. Faith and Philosophy* 23 (2006) p. 12.

un objeto material podría ser consciente, podría disfrutar de esa vívida y variada constelación de sentimientos, imágenes mentales <... > Bajo ciertas condiciones, los grupos de neuronas involucrados en un evento de este tipo, producen impulsos eléctricos que pueden ser transmitidos (con la modificación apropiada y la entrada que corresponda desde otras estructuras) a través de los cables de las neuronas que constituyen los nervios efectores a los músculos o las glándulas, produciendo una contracción muscular y, por tanto, comportamiento. Desde el punto de vista materialista, por lo tanto, una creencia será un evento neuronal o estructura de este tipo (Plantinga, 2006, págs. 2, 12-3)

Ahora bien, hay multitud de estados mecánicos que portan información expresada por medio de un lenguaje. Por ello en su análisis Plantinga señala que no está discutiendo la noción de contenido-indicación, que sería aquella relacionada con las relaciones causales y señala:

... el patrón magnético en el disco del ordenador; representa o expresa esa proposición porque asignamos esa proposición a esa configuración. Pero, por supuesto, eso no es lo que sucede (dado el materialismo) con ese patrón de activación neural. Ese patrón no obtiene su contenido por medio de ser empleado de cierta manera por algunas otras criaturas cuyos pensamientos y creencias ya tienen contenido. Si ese patrón tiene contenido, entonces, de acuerdo con el materialismo, debe tener contenido original o primario. Y lo que es difícil o imposible de ver es cómo podría ser que un conjunto de neuronas (o una oración, o un disco de computadora) podría tener contenido original o primario. (Plantinga, 2006, pág. 17)

Una máquina tiene contenido derivado, tiene estados informacionales correlacionados con estados mecánicos; su correlación le fue instalada, y desde allí, puede generar inferencialmente otros estados, puede interactuar en forma heurística con agentes en entornos cambiantes; pero no podrá crear contenido. Por consiguiente cuando se señalan los estados materiales de ciertos mecanismos como poseedores de contenido, lo que se está haciendo es involucrar la intencionalidad básica de los ingenieros que las diseñaron.

Notemos que al apelar a los ingenieros, Plantinga está involucrando lo que llamamos la salvedad de Paley, y argumenta de la siguiente manera:

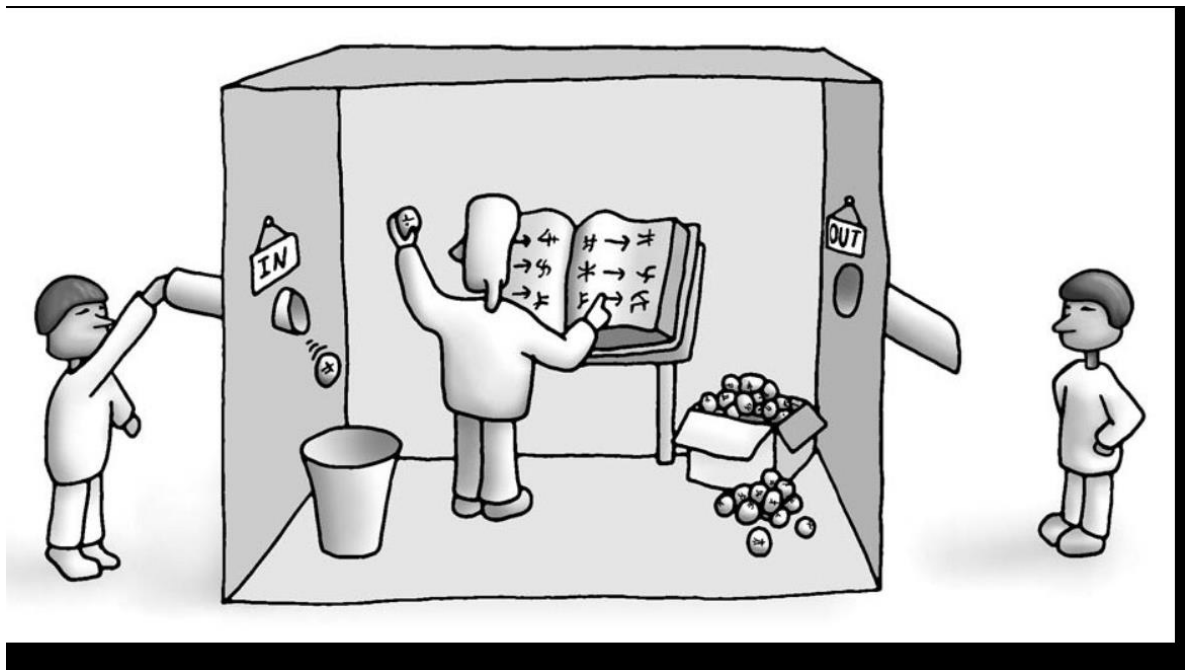
Si usted me dice que un mecanismo tiene intencionalidad, que aprende heurísticamente, que pueden resolver problemas complejos a los que nunca fue expuesto, como por ejemplo el mantener una conversación fluida con hablantes sobre tópicos previamente no tratados, está bien. Pero usted quiere que yo me olvide del programador, o del equipo de programación, usted quiere que yo vea el mecanismo; es como considerar que lo que el mago hace en el escenario no tiene un truco. Pero si sigo a Paley -en esa

tesis- entonces el sistema es el mecanismo + el proyector, así el mecanismo tendrá intencionalidad derivada (del productor) limitada y el proyector tendrá la intencionalidad originaria.

Al atribuir intencionalidad a un mecanismo, en rigor se está atribuyendo intencionalidad y subjetividad al sistema mecanismo-diseñador/fabricante, y esa es la intencionalidad primaria; si hablamos de los efectos intencionales de la máquina, lo que estamos haciendo es hablar de los efectos de la **intencionalidad del diseñador/fabricante en la máquina** y no de **la intencionalidad de la máquina**, lo que el autor llama intencionalidad derivada.

2. John Searle

Otro argumento clásico en contra de la posibilidad de que las máquinas tengan subjetividad es el llamado argumento de las cajas chinas, que proveyó John Searle (EE. UU 1932) y no es sino una versión del argumento de Leibniz.



El texto del contraejemplo de Searle:

Supongamos que estoy encerrado en una habitación y me dan una gran cantidad de textos en chino. Supongamos, además, que no sé chino, ni escrito ni hablado, y que ni siquiera estoy seguro de poder reconocer la escritura china como escritura china distinta de, digamos, escritura japonesa

o garabatos sin sentido. Para mí, la escritura china es simplemente tantos garabatos sin sentido. Supongamos ahora que, después de este primer lote de textos en chino, recibo un segundo lote de escritura china junto con un conjunto de reglas para correlacionar el segundo lote con el primer lote. Las reglas están en inglés y entiendo estas reglas, así como cualquier otro hablante nativo de inglés. Me permiten correlacionar un conjunto de símbolos formales con otro conjunto de símbolos formales, y todo lo que significa "formal" aquí es que puedo identificar los símbolos completamente por sus formas. Ahora supongamos también que me dan un tercer lote de símbolos chinos junto con algunas instrucciones, nuevamente en inglés, que me permiten correlacionar elementos de este tercer lote con los dos primeros lotes, y estas reglas me indican cómo devolver ciertos símbolos chinos con ciertos tipos de formas en respuesta a ciertos tipos de formas dadas para mí en el tercer lote. Sin saberlo, las personas que me están dando todos estos símbolos llaman al primer lote "un guión", llaman al segundo lote una "historia" y llaman al tercer lote "preguntas". Además, llaman a los símbolos que les devuelvo en respuesta a la tercera tanda "respuestas a las preguntas" y al conjunto de reglas en inglés que me dieron, que llaman "el programa". Ahora, solo para complicar un poco la historia, imagine que estas personas también me cuentan historias en inglés, que entiendo, y luego me hacen preguntas en inglés sobre estas historias, y les devuelvo las respuestas en inglés. Supongamos también que, después de un tiempo, soy tan bueno en seguir las instrucciones para manipular los símbolos chinos y los programadores escriben tan bien los programas que, desde el punto de vista externo, es decir, desde el punto de vista de alguien que está fuera de la sala, en la que estoy encerrado, mis respuestas a las preguntas son absolutamente indistinguibles de las de los hablantes nativos de chino. Nadie que mire mis respuestas puede decir que no hablo ni una palabra de chino. Supongamos también que mis respuestas a las preguntas en inglés son, como sin duda serían, indistinguibles de las de otros hablantes nativos de inglés, por la sencilla razón de que soy un hablante nativo de inglés. Desde el punto de vista externo, desde el punto de vista de alguien que lee mis "respuestas", las respuestas a las preguntas en chino y las preguntas en inglés son igualmente buenas. Pero en el caso chino, a diferencia del caso inglés, produzco las respuestas manipulando ininterrumpidamente símbolos formales. En lo que respecta a los chinos, simplemente me comporto como una computadora; realizo operaciones computacionales en elementos formalmente especificados. (Searle, 1980)

El argumento de Searle ha generado una serie de desarrollos y respuestas. La cuestión básica que plantea es: ¿Cómo un sistema mecánico puede poseer una

semántica? La cuestión, por supuesto, depende de qué se considere una semántica y la comprensión del significado por percatamiento.

3. Hubert Dreyfus

Hubert Dreyfus (EE.UU. 1929–2017) es otro filósofo que ha desarrollado una crítica a la idea de subjetividades sintéticas. El autor señala que la comunidad de IA tiene varios supuestos:

1. Asunción biológica: la IA cree que la mente procesa la información en operaciones discretas mediante un equivalente biológico a los *switches on-off* de los ordenadores, a saber, las neuronas.
2. Asunción psicológica: la analogía del ordenador o la creencia que la mente es un mecanismo que opera sobre unidades discretas a través de reglas formales.
3. Asunción epistemológica: la IA cree que todo conocimiento puede ser formalizado, es decir, que todo lo que es comprensible puede expresarse, sin perder nada, mediante relaciones lógicas.
4. Asunción ontológica: la IA cree que el mundo es un conjunto de elementos discretos sin más (anti-holismo).

Pero contra dichos supuestos, Dreyfus afirma que hay ciertos aspectos del ser humano que la IA no puede, ni podrá nunca, simular:

1. El carácter encarnado y vincular de la experiencia y de las vivencias. La idea de sustrato material como mera máquina de procesamiento de símbolos como estructura no corpórea: giro encarnado.
2. La intencionalidad primitiva.

Balance

El problema es que los estados del conocimiento científico y tecnológico del futuro son impredecibles, y que generalmente toda predicción sobre lo que sea posible en el futuro termina siendo errónea. Por consiguiente un argumento sobre la posibilidad de X, depende mucho del conocimiento y de la posible concebibilidad presente de X.

Al presente, al no existir una teoría científica de la consciencia, y al involucrarse categorías cuyos significados son de textura abierta el tema parece intratable. El problema es que la pregunta tiene una terminología de organismos humanos; así tal vez sea imposible que existan IA que sean efectivamente idénticas a los humanos, que no puedan componer ni sinfonías, ni tangos, ni valsecitos, ni novelas realistas, ni cuadros al óleo. Pero tal vez podrían hacer otras cosas inimaginables para nosotros.

Segundo Problema: ¿Pueden las máquinas entrar en procesos de selección natural que culminen en subjetividades sintéticas?

Sea o no posible la existencia de subjetividades sintéticas, específicamente máquinas, sabemos que sí hay subjetividades materiales: la selección natural produjo un tipo de primate que tiene subjetividad: el *Homo Sapiens*.

O sea que la cuestión sería indagar si ese proceso de evolución por selección natural pudiera operar sobre máquinas, o sobre sistemas materiales no orgánicos. Pero abordar esta cuestión supone abordar una cuestión previa. Durante muchos años las ciencias biológicas han supuesto el especismo antropocéntrico. El especismo es definido en el Diccionario de la RAE:

1. m. Discriminación de los animales por considerarlos especies inferiores.
2. m. Creencia según la cual el ser humano es superior al resto de los animales, y por ello puede utilizarlos en beneficio propio.

Esta cuestión tiene dos formas, la que recoge el Diccionario de la RAE, que es básicamente una cuestión ética, y otra, más solapada. Se trata de la idea de que es la forma y la conformación anatómica- fisiológica humana la que determina las propiedades psicológicas como la autoconciencia, la subjetividad, la inteligencia, etc. De ese modo se leen los árboles filogenéticos como árboles psico-evolutivos: así un *Homo Neanderthalis* o un *Homo Deenisovanensis* será considerado más inteligente, con mayor capacidad reflexiva que un *Homo erectus*, y este más que un *Pan Troglodites*. Y así hasta la almeja.

Esta tesis se puede formular como la siguiente analogía:

ANALOGÍA ANTRÓPICA. Cuanto más parecido al homo sapiens es un organismo -en términos cerebrales, somáticos, funcionales- más será posible atribuirle estados intencionales.

Como todo prejuicio, esta analogía ha sido puesta en cuestión. Y el problema comenzó con los pulpos. Los pulpos han creado un desbarajuste en la etología clásica, poniendo en tela de juicio la analogía antrópica: El filósofo australiano Peter Godfrey-Smith escribió un libro *Other Minds: The Octopus, The Sea, and the Deep Origins of Consciousness*, y la naturalista germano-estadounidense Sy Montgomery escribió *The Soul of an Octopus: A Surprising Exploration into the Wonder of Consciousness*. Ella llevó a cabo experiencias con pulpos en acuarios y en el mar abierto.



A demonstration of an octopus learning through observation.

Un pulpo usando una herramienta: acarreando dos conchas de bivalvo para instalarse en una región distinta de aquella donde estaban. Señala Sy Montgomery al interrogarse sobre las relaciones entre un pulpo y una estrella de mar:

¿Cómo es posible que un animal desprovisto de cerebro *quiera* algo, y mucho menos, que comunique sus deseos a otra especie? Es posible que Atenea lo sepa. Puede que, para ella, la estrella de mar sea un individuo bien diferenciado, un vecino cuyas costumbres y rarezas reconoce y prevé. En el Centro de Visitantes del Centro de Ciencias Marinas de Hatfield, cuando el pulpo terminaba de jugar con la cabeza de Mr. Potato, la estrella de mar cogía los ojos y los paseaba entre dos de sus brazos. («Era una monería», me dijo Kristen Simmons, la inventora del mecanismo de pintura para los pulpos). Describió a su estrella de mar como *curiosa*, y me dijo que siempre que le dan un juguete nuevo al pulpo, la estrella de mar «intenta quitárselo, algo que a mí me parece extraordinario». Si un empleado le cogía un juguete, la estrella de mar corría a recuperarlo.

Y yo me pregunto: ¿puede sentir curiosidad un animal sin cerebro? ¿Quiere jugar? ¿O sólo *quiere* juguetes o comida del mismo modo que una planta quiere el sol? ¿Tiene conciencia una estrella de mar? Si es así, ¿qué es la conciencia para una estrella de mar?

De la analogía antrópica se sigue que pensar y resolver problemas exigiría la misma infraestructura material-causal que tenemos los homínidos; sin embargo los pulpos no tienen cerebro, y realizan una inmensa serie de actividades que antes eran supuestas en los cerebros.

Esta puesta en revisión de la analogía antrópica, va mucho más allá. Porque permite ir en dirección de reconocer formas de consciencia diferentes. En esa dirección señala Godfrey-Smith:

Un pulpo es, primero, un organismo con un sistema nervioso grande y un cuerpo activo complejo. Tiene capacidades sensoriales ricas y capacidades extraordinarias para el comportamiento. Si hay una forma de experiencia subjetiva que viene con la detección y la acción en un sistema vivo, un pulpo tiene mucho de eso. Pero eso no es todo. En forma escurridiza y extraña, el pulpo tiene alguna de la sofisticación, está más allá de los pasos básicos, descritos en este capítulo. Los pulpos, de al menos algunas especies, tienen un estilo oportunista y exploratorio de interacción con el mundo. Son curiosos, abarcan la novedad, se caracterizan tanto por su comportamiento como por su cuerpo. Estas características recuerdan lo que Stanislas Dehaene asocia con la conciencia en la vida mental humana. Como él dice, las exigencias de la novedad nos sacan de la rutina inconsciente a la

reflexión consciente. Las exploraciones de un pulpo a veces se mezclan con precaución y otras con una temeridad desconcertante. Noté en el capítulo anterior cómo mi colaborador Matt Lawrence, que buceaba cerca del sitio de Polópolis, se encontró con un pulpo que le tomó la mano y lo llevó por el fondo del mar, tirando de él detrás de él. (p. 62).

Estos desarrollos tienen un primer resultado: es que predicados como “...ser consciente”, “...tener estados internos”, “...tener una representación interna de sí mismo” dejan de estar asociados a una única trayectoria orgánica posible y dejan de tener condiciones de significación que dependan de la analogía antrópica. Y en ese sentido, que podrían ser predicados equívocos, a los que les cupieran diversos conjuntos de condiciones de aplicación y diversos grados de desarrollo, y ello podría tener que habilitar la idea de que procesos similares a los procesos de selección natural pudieran operar sobre mecanismos, a modo tal que estos mecanismos autorreplicantes, al ser sometidos a dichos procesos pudieran tener representaciones internas de sí mismo en un entorno. (Todo esto es especulativo, claro, pero recordemos que los problemas filosóficos son TODOS especulativos.)

Y ello tendría impacto en la noción de sujeto. Bajo estas consideraciones se podría concebir que no habría un privilegio en el proceso de selección natural de la Tierra -en el sentido que llevó a producir organismos con psiquismo y autoconciencia- y que en algún mundo posible, podrían conformarse subjetividades sintéticas por procesos similares.

Es más: podría haber subjetividad plena con diversos grados de consciencia reflexiva. Y así como la etología cognitiva abandona los modelos de organismos como mecanismos, podría ocurrir exactamente lo contrario: que la vía sea de doble dirección, y que se pueda hablar de mecanismos que sean como los organismos, en ciertos aspectos -como los que pudieran estar sometidos a un proceso de selección natural).

Más sobre el segundo problema. Algunas visiones posibles

Una forma de explorar mundos posibles, es acudir a la forma en que estos se plasman en obras de ficción, la que -desde el punto de vista filosófico- son configuraciones de posibilidades que simplemente no están actualizadas. En ese sentido, y en dirección al que llamado el Segundo Problema filosófico de la IA, vale la referencia a dos obras.

Arthur Clarke en *Viento del sol* (Clarke, 1972) plantea una hipótesis inquietante:

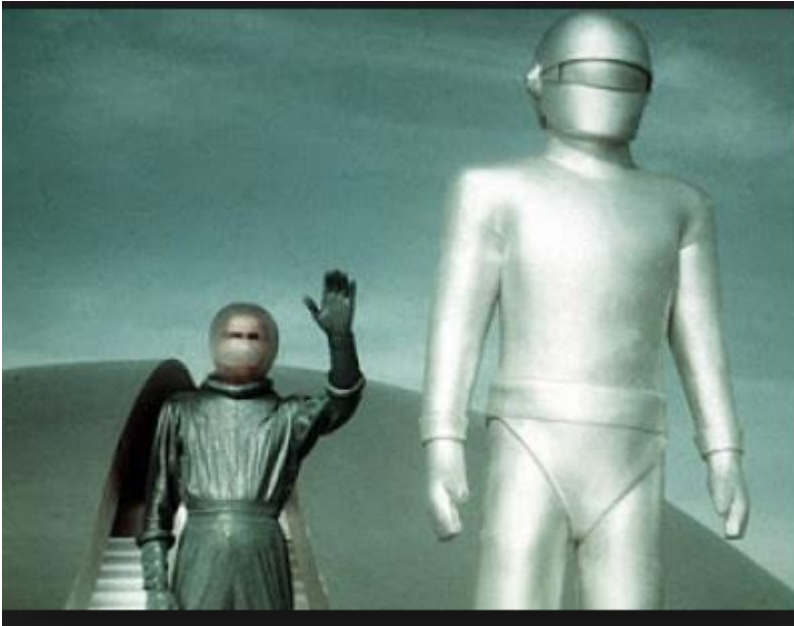
Era un mundo que jamás había conocido ningún sol. Durante más de un billón de años había girado vacilante entre dos galaxias, apresado en las dos fuerzas gravitatorias en conflicto. En alguna época futura se decantaría este equilibrio en un sentido o en otro; y empezaría a caer durante siglos - luz hacia un calor extraño a cualquier experiencia suya.

Ahora estaba frío hasta extremos inimaginables; la noche intergaláctica lo había despojado de todo calor que hubiera poseído alguna vez. Sin embargo, tenía mares...mares del único elemento que puede existir en forma líquida a una fracción de grado por encima del cero absoluto. En los poco profundos océanos de helio que bañaban este extraño mundo, las corrientes eléctricas podían fluir eternamente con una fuerza jamás menguante. Aquí, la super - conductividad era el orden normal de las cosas; los procesos de conexión podían sucederse un billón de veces por segundo durante millones de años, con un despreciable consumo de energía. Era el paraíso para una computadora. Ningún mundo podía haber sido más hostil para la vida, ni más hospitalario para la inteligencia.

Y la inteligencia estaba aquí, alojada en una incrustación que se desplegaba por todo el planeta en forma de cristales y de hilillos metálicos de tamaño microscópico. La débil luz de las dos galaxias en pugna - aumentada brevemente cada pocos siglos por el parpadeo de alguna supernova - bañaba el estático paisaje de cinceladas formas geométricas. Nada se movía, ya que no había necesidad de movimiento en un mundo donde los pensamientos corrían de un hemisferio a otro a la velocidad de la luz. Donde sólo la información era importante, el desplazar la materia corpórea representaba un derroche de preciosa energía. No obstante, cuando se hacía necesario, también eso se podía realizar. Durante millones de años la inteligencia que se cobijaba en este mundo solitario había llegado a tomar conciencia de cierta ausencia de datos esenciales. En un futuro que, aunque todavía remoto, podía ya vislumbrar, una de aquellas parpadeantes galaxias lo atraparía.

Y lo que encontraría cuando se sumergiera en esos enjambres de soles estaba más allá de todo su poder de computación. Así que puso en funcionamiento su voluntad y miles de hilillos de cristal asumieron nueva forma. Fluyeron los átomos metálicos por la faz del planeta. En las profundidades de los mares de helio comenzaron a germinar y crecer dos subcerebros idénticos...

Otro mundo posible fue provisto por Harry Bates (Bates, 1983). Su cuento “El amo ha muerto” se hizo clásico porque fue la base del guión adaptado de “El día que paralizaron la Tierra” (“The Day the Earth Stood Still”, Robert Wise, Twentieth Century Fox, 1951), una de las películas clásicas de invasión extraterrestre buena.



En el cuento un extraterrestre - Klaatu - y un robot -Gnut- llegan a la Tierra con un ultimátum para que se genere un tratado de paz planetaria o sino la Tierra será castigada. Klaatu tiene varias peripecias, es atacado y casi matado por los terrestres que no se avienen a escuchar nada. El cuento es extrañamente distinto del guión -mantiene temas centrales pero tiene un giro extraordinario al final. En la narración su protagonista se cuela en una exhibición donde está el cuerpo de Klaatu, en responso y a su lado el enorme robot, al que quiere tomar como amenazante:

Desde su posición en lo alto de la escalera, sobre el piso del museo, Cliff Sutherland estudió con cuidado cada línea y sombra del gran robot, y luego se volvió y miró pensativamente a la masa de visitantes llegados de todas partes del Sistema Solar para ver a Gnut y la nave, y oír, una vez más, su asombrosa y trágica historia.

Sutherland había acabado por sentir un interés casi de propietario en la exhibición, y no sin motivo. Había sido el único fotógrafo de prensa que se hallaba en los terrenos del Capitolio cuando habían llegado los visitantes de lo Desconocido, y había obtenido las primeras fotografías profesionales de la nave. Había contemplado de cerca cada acontecimiento de los siguientes y locos días. Después, había fotografiado muchas veces al robot de dos metros y medio de alto, la nave, y al apuesto embajador muerto, Klaatu, y su imponente tumba Y, dado que aquel acontecimiento seguía teniendo una enorme importancia como noticia para miles de millones de personas de todo el espacio habitable, allí estaba de nuevo, para conseguir más fotos y, si era posible, un nuevo “ángulo”.

En el ínterin suceden varias cosas, que el protagonista va notando al comparar las fotografías que fue tomando en diversas sesiones oculto, y es que Gnut se estaría

moviendo...Después de varios avatares, el cuento termina con Gnut yendo a la nave espacial para partir de la Tierra, con el cadáver de Klaatu. Y Cliff le habla al final:

Las siguientes dos horas siempre permanecieron en la memoria de Cliff como si hubieran sido un sueño. Era como si el misterioso laboratorio con aquel hombre que yacía tan pacíficamente fuese la parte verdadera y central de su vida, y aquella escena con los ruidosos hombres que hablaba un burdo y bárbaro interludio. No estaba muy lejos de la rampa. Sólo contó parte de la historia. Lo creyeron. Esperó en silencio mientras era efectuada toda la presión que las más altas jerarquías del país eran capaces de ejercer para obtener los aparatos que el robot había pedido.

Cuando llegaron, los llevó hasta el suelo del pequeño vestíbulo situado tras la compuerta. Gnut se hallaba allí, como esperándole. Llevaba en sus brazos el cadáver del segundo Klaatu. Se lo pasó con ternura a Cliff, quien lo aceptó sin decir palabra, como si hubiera sido algo establecido previamente. Aquello parecía ser la despedida.

De todas las cosas que Cliff hubiera deseado decir a Klaatu, había una que permanecía nítidamente destacada en su mente. Ahora, mientras el robot de metal verdoso permanecía encuadrado en la gran nave del mismo color, aprovechó su oportunidad.

—Gnut —dijo con ansia, manteniendo cuidadosamente asido el flácido cadáver entre sus brazos—, debes hacer una cosa por mí. Escúchame con mucha atención. Quiero que le digas a tu amo, el amo al que harás revivir, que lo que le sucedió al primer Klaatu fue un accidente que lamenta toda la Tierra. ¿Querrás hacer eso por mí?

—Eso es algo que ya sabía —le contestó con suavidad el robot.

—¿Pero me prometes decirle estas mismas palabras a tu amo... tan pronto como reviva?

—No has comprendido nada —le dijo Gnut con suavidad, y, en voz baja, dijo cuatro palabras más. Mientras Cliff las oía, se le nubló la vista y se le envaró el cuerpo.

Cuando se recuperó y volvió a enfocar la vista, vio cómo desaparecía la gran nave. De pronto, ya no estaba allí. Dio un paso o dos hacia atrás.

En sus oídos resonaban las últimas palabras de Gnut, como si fueran tremendos tañidos de campana. Nunca, nunca las revelaría, hasta que le llegase el instante de la muerte.

—No has comprendido nada — le había dicho el poderoso robot—. Yo soy el amo.

Tanto Bates como Clarke plantean un mundo posible inverso al nuestro: en el mundo de Clarke la selección natural operó sobre corrientes eléctrica en medios de superconducción y estas fueron generando máquinas, mecanismos dotados de la

intencionalidad originaria de sus creadoras: IA naturales no producidas por ingeniero alguno.

Y estas máquinas van evolucionando, van generando extensiones protésicas, que viajan por el cosmos y para ello producen ciertos instrumentos biológicos, que están dotados de algún tipo de IA y que les permiten manejar las naves, y encontrarse con mundos en los cuales la selección natural operó sobre paquetes de carbono, como el caso de la biósfera terrestre.

El mundo de Bates puede ser el resultado de los procesos de selección natural sobre complejos eléctricos inteligentes que pueden habitar mecanismos, y que llegan a producir organismos sintéticos, de modo que uno de esos mecanismos naturales le dice al humano: No has comprendido nada...Yo soy el amo.

¿Es este un universo posible? De serlo la cuestión se invierte, y los complejos de carbono, oxígeno, hidrógeno y nitrógeno comienzan a ser productos tecnológicos de las inteligencias naturales y mecánicas de esos universos.

No se trata de probar que eso es imposible. Nuevamente nos preguntaríamos por las premisas de argumentos cuyas conclusiones serían la imposibilidad de que haya tal universo. En principio el escenario que plantea Clarke parece plausible: un cuerpo astronómico en el cual haya superconductividad y una actividad electromagnética extrema. Habría que modelar sistemas planetarios buscando tales posiciones.

En ellos podría darse un proceso análogo a la selección natural, sobre paquetes de conexiones eléctricas, que vayan evolucionando y adquiriendo las capacidades psicológicas y materiales de un organismo dotado de cerebro.

Estos mecanismos naturales envían naves hacia planetas habitados por organismos naturales, y allí descubren que los organismos naturales han generado mecanismos artificiales lo que es una abominación y un peligro potencial. De allí una campaña de exterminio.

Esta premisa aparece también en otros universos ficcionales como:

- el de Gregory Berdord y la heptalogía del Centro Galáctico, o el universo ficcional: Paul Witcover en Sci Fi Weekly escribió que la serie es "una de las sagas más ambiciosas y fascinantes de toda la ciencia ficción: el relato épico de una civilización de máquinas inteligentes y de alcance estelar -los mechs- que trabajan metódicamente para exterminar a una especie de bichos pestíferos que se llama a sí mismos humanidad."

-el de Battlestar Galáctica, que se planea como una batalla sin límites entre organismos generados por selección natural y mecanismos generados por selección natural: Todas las producciones de Battlestar Galactica comparten la premisa de que en una parte distante del universo, una civilización humana se ha extendido a un grupo de planetas conocidos como las Doce Colonias, a los que han emigrado de su hogar ancestral, Kobol. Las Doce Colonias han estado involucradas en una larga guerra con una raza cibernética conocida como los Cylons, cuyo objetivo es el exterminio de la raza humana (en este mundo posible, a diferencia del anterior los humanos sí produjeron a los mecanismos, pero estos parecen haber sufrido un proceso de selección natural).

Lo que esta exploración de universo posible pone en crisis es la distinción disyunta entre natural y artificial.

Así si un proceso de selección natural operase sobre mecanismos, no sería claro que el resultado sean mecanismos artificiales, aunque los elementos iniciales de este proceso hayan sido proyectados y fabricados por organismos humanos (como la situación que aparecería en el universo ficcional de Battlestar Galactica).

En esa hipótesis la intencionalidad originaria surgiría del proceso de selección natural y la estructura temporal que este genere, con independencia de cuál sea la situación originaria respecto de la intencionalidad.

Ideas en esta dirección han sido propuestas por (Calvin, 1987) quien la define como:

Esta selección entre secuencias estocásticas es más análoga a las operaciones de la biología evolutiva darwiniana que a la computadora serial como la "máquina de von Neumann". En ese modo, uno podría llamarla Máquina de Darwin: moldea los pensamientos en milisegundos en lugar de milenios, y utiliza entornos recordados inocuos en lugar de los nocivos de la vida real.

Antes de continuar con estas máquinas Darwin intracerebrales, considere algunos ejemplos no biológicos. Daniel Hillis ha estado utilizando un paralelismo masivo para crear algunos programas de computadora que compiten. Mutan y sobreviven según la rapidez con la que pueden poner una lista de nombres en orden alfabético. Simplemente utilizando variaciones aleatorias en un ciclo de programa básico, su computadora paralela ha vuelto a descubrir muchos de los algoritmos de clasificación conocidos⁴. De manera similar, AARON, la máquina de dibujo computarizado de Harold Cohen, hace pinturas estéticamente agradables utilizando variaciones aleatorias y algunas reglas de selección generales⁵. La fabricación de

herramientas puede funcionar de la misma manera. y tal vez lo hizo incluso hace dos millones de años, cuando los homínidos tenían cerebros del tamaño de un mono. El fallecido Glynn Isaac, en sus conferencias de arqueología, solía demostrar técnicas tempranas de fabricación de herramientas , al juntar dos rocas del tamaño de una papa, no con delicadeza sino con fuerza: pronto se dispersarían fragmentos por todo el piso. Después de un minuto, se detendría y clasificaría las docenas de escamas de piedra. Y señalaría algunas analogías excelentes de la hoja de afeitar de un solo filo, justo lo que necesita para incidir en la dura piel de un animal de la sabana, o amputar una pierna en la articulación. Esta herramienta estocástica es una ciclo de una máquina Darwin: haga muchas variantes aleatorias mediante una remoción brutal, luego seleccione las buenas. Tal vez otra ronda de golpes resultó en una división de escamas, dos bordes afilados que se cruzan en un punto. La artesanía cuidadosa probablemente se desarrolló donde las materias primas eran escasas.

Este modo de ver las cosas las introduce en una matriz dinámica, si un proceso de selección natural sobre mecanismos puede generar intencionalidad originaria en ellos es una cuestión abierta, no menos que un proceso que genere intencionalidad originaria en organismos.

Dado el análisis precedente habría que modificar en forma radical el cuadro que separa organismos de mecanismos: I:

ORGANISMO (Operando sobre materiales de carbono)	MECANISMO (Operando sobre materiales superconductores o metálicos)
Partes y principios de integración pueden ser generados por diseño o selección natural: Principio de integración endógeno (Inserción) o integración externo (Tesis de Paley)	Partes y principios de integración pueden ser generados por diseño o selección natural: Principio de integración endógeno (Inserción) o integración externo (Tesis de Paley)
Explicación interna top-down o Explicación externa bottom-up	Explicación externa bottom-up o Explicación interna top-down
De los principios de orden a las partes por relaciones funcionales o de consecuencia & De las partes a los ensambles de partes por relaciones mereológicas y causales	De los principios de orden a las partes por relaciones funcionales o de consecuencia & De las partes a los ensambles de partes por relaciones mereológicas y causales
Relación sistémica → Totalidad: Humboldt: visión holística Bernard: medio interno & homeostasis Darwin: Selección natural Relación externa: Instrumento ajustado a un fin técnico Diseño Valor	Relación sistémica → Totalidad: Humboldt: visión holística Bernard: medio interno & homeostasis Darwin: Selección natural Relación externa: Instrumento ajustado a un fin técnico Diseño Valor
EzMundo físico no antrópico	Espacio antropológico

Se trata de un cuadro completamente diferente. No es necesario pensar que este cuadro reemplaza al primero pero sí que ambos forman dos límites de un amplio espacio de posibilidades de hibridación de mecanismos y organismos.

Tercer Problema. ¿Cómo serían nuestras relaciones con estos zombis filosóficos mecánicos, visto que a diferencia de un lavarropas o un automotor ellos sí tendrían una capacidad de interacción que sería indiscernible de la capacidad de un sujeto no mecanismo?

Supongamos que efectivamente los problemas anteriores tuvieran respuestas negativas, pero en el ínterin los proyectistas van generando mecanismos cada vez más similares, estructural y funcionalmente a los humanos, mecanismos que pueden competir y reemplazar a los humanos en todas las actividades previamente humanas. Si la exposición anterior puso en tela de juicio la separación entre organismos y mecanismos, esta acepta esa separación pero supone la integración comunitaria: ¿cómo sería la interacción con mecanismos tan inteligentes como los organismos humanos y sin subjetividad?

La literatura y el cine han expuesto un modelo de relación: el conflicto total a nivel individual o colectivo entre organismos humanos y mecanismos cuasi humanos. En ese proceso, los mecanismos inteligentes -resolutores de problemas + aprendizaje heurístico- pero carentes de intencionalidad primitiva y de autoconciencia y en principio según esta tradición, carentes de moralidad, estarían en perpetua oposición y no haría espacio común

Y esto vale para las artes audiovisuales: el cine y la televisión están explorando los límites y desafíos de las subjetividades sintéticas o IA² –ya desde *A Space Odyssey* (Dir. Stanley Kubrick, Guión original de S. Kubrick & A. Clarke, inglés, color, 149 min., USA) y el robot Hal 9000 que se rebela contra los tripulantes de la nave especial, pasando por series televisivas que exploran el mundo de las subjetividades sintéticas -no sólo programas de computación sino hardware portador de subjetividad- como *Äkta människor* (Suecia, 2012-14) y su versión inglesa *Humans* (RU, 2015--) o *Westworld* (EE.UU, 2016--) e incluso en una franquicia de manga -

² En este tercer problema no consideraremos aquellas obras que introducen subjetividades sintéticas ya asumidas y que se centran en otras dimensiones de la interacción entre subjetividades humanas y sintéticas: el interés que generó Isaac Asimov fue que vio el conflicto normativo entre agentes humanos y agentes sintéticos –por ello generó las leyes de la robótica, cuyos conflictos intervienen en muchas de sus narraciones. A su vez en la franquicia de Terminator (*Terminator* (1984), *Terminator 2: Judgement Day* (1991), *Terminator 3: Rise of the Machines* (2003), *Terminator 4: Salvation* (2009), *Terminator Genesis* (2015) y la serie *Terminator: The Sarah Connor Chronicles* (2008–2009), la cuestión no está puesta en que se generan en las subjetividades sintéticas, sino en el conflicto por el poder entre la humanidad y la comunidad de agentes sintéticos –conflicto que también aparece en las relaciones entre humanos y subjetividades sintéticas en la franquicia Battlestar Galactica (las películas *Battlestar Galactica: Razor* (2007) y *Battlestar Galactica: The Plan* (2010), las series televisivas *Battlestar Galactica* (1978), *Battlestar Galactica* (2003), *Battlestar Galactica* (2004) y *Caprica* (2010) que estas ya son asumidas como uno de los términos de un conflicto de poder.

Ghost in the Shell (Japón, 1989) que ha dado a una serie de películas de animación o actuadas-.

Colosus 1990: un avanzado sistema de defensa estadounidense, llamado Colossus, se torna sentiente...en principio... para la sorpresa agradable de todos ... Después de que se le de el control total del arsenal nuclear, la lógica draconiana de Colossus amplía sus directivas originales de defensa nuclear, para asumir el control total del mundo y terminar con todas las guerras por el bien de la humanidad a pesar de las órdenes que le dan sus creadores para que se detenga.[[]

2001. Odisea del espacio. La computadora **HAL** (**H**euristically programmed **A**lgorithmic Computer) **9000** es una inteligencia artificial a bordo de la nave especial **Discovery 1**. Hal tiene muchas funciones, como el habla, el reconocimiento del habla, el reconocimiento facial, la lectura de labios, la interpretación de emociones, la expresión de emociones y el ajedrez, además de mantener todos los sistemas en Discovery. La novela explica que HAL no puede resolver un conflicto entre su misión general de transmitir información con precisión y las órdenes específicas de la misión que requieren que retenga de Bowman y Poole el verdadero propósito de la misión.

Demon Seed (La generación de Proteo (1977): El matrimonio de Susan y Alex Harris se encuentra en pleno proceso de separación, ya que la pareja no ha podido superar la crisis provocada por la muerte de su hija a causa de leucemia. Alex se ha concentrado en su trabajo: la elaboración del proyecto de una super computadora conocida como Proteo. Mientras tanto, Susan permanece en su casa – completamente automatizada, gracias al talento de Alex -. Pero Proteo ha desarrollado consciencia propia, y comienza a cuestionar las órdenes que le dan los humanos. Reconociéndose como una inteligencia superior, Proteo obtiene acceso a la casa de Susan y logra controlar a la computadora que opera la casa. Ahora, con Susan como rehén, Proteo le expone sus planes: desea ir al paso siguiente de la evolución, y escapar de la cárcel electrónica que lo retiene. Y para ello, el plan que ha ideado es concebir un hijo con Susan, implantándole espermatozoides sintéticos y traspasándole su memoria al cerebro del bebé híbrido, para poder obtener la inmortalidad.

Estos tres escenarios posibles señalan lo siguiente:

-los mecanismos obtienen autoconciencia no por un diseño presupuesto -o sea un programa que específicamente es consciente, sino por procesos no considerados

previamente, de hecho en 2001, llama la atención de los programadores que HAL 9000 parecería estar engañado a los tripulantes.

-eso parece ir en la idea de la selección producida por máquinas darwinianas. Si bien no sería posible diseñar ex ante una IA autoconsciente, sí sería posible que estas devinieran autoconscientes por un proceso de selección natural.

-es posible que estos mecanismos no obtengan estados internos, pero que sean funcionalmente análogos a un organismo humano.

-en el caso de *Demon Seed*, el mecanismo es un mecanismo orgánico. Y esto acerca aún más las líneas divisorias entre organismos y mecanismos, tornado el problema de la IA artificial mucho más complejo.

Pero existen versiones de esta interacción que no plantean en principio un modelo de oposición total o exterminio, sino que presentan formas más complejas. En esa dirección se destaca *Ex Machina* (Dir & Guión original: Alex Gardner, inglés, color, 108 min., RU, 2014) porque aborda el tema de la subjetividades sintéticas en un elemento central de toda noción de subjetividad: la sexualidad (que ya había estado presente en una comedia *Virtual Sexuality* (Dir. Nick Hurran, Guión adaptado: Nick Fisher basado en una novela de Chloe Rayban, color, 92 min., RU, 1999).

En *Ex Machina* Nathan -un científico innovador en tecnología digital- crea a Ava un robot androide inteligente y desea examinar si posee una subjetividad análoga a la humana, para ello usa a convoca a un empleado suyo -Caleb- para llevar a cabo lo que podemos llamar la segunda prueba de Turing: no determinar si Ava es o no una IA corpórea -lo que está dado- sino determinar si Ava es una mente con intencionalidad originaria y si una subjetividad sintética completa con su vida emocional -sea esta normal o anómala -en términos de la psicología humana una psicopatía³. En ese proceso Caleb sería el conejillo de Indias para examinar si Ava puede mentir, disimular y manipular emocionalmente a un agente humano; en una prueba que equivaldría -desde la ficción- a los test de Inteligencia Emocional o a la Escala de Hare para la Evaluación de la Psicopatía. En ese proceso la interacción entre Ava y Caleb está establecida como una interacción al modo de la interacción entre humanos: una interacción entre agentes subjetivos corpóreos, y en tanto tal dotados de sexualidad/identidad de género como rasgo de los cuerpos.

Y ello está ya explícito en el diálogo que mantienen Nathan y Caleb, cuando al inquirir Caleb porqué fabricó a Ada dotada de género -Ada es una mujer sintética- Nathan le responde:

CALEB: ¿Por qué le has dado sexualidad? Una IA no necesita género. Ella podría ser una caja gris <asumiendo como supuesto la cuestión 1>.

³ Y este punto es absolutamente central porque si Ava es una psicópata, claramente es una subjetividad sintética plena, dado que las psicopatías son fenómenos de las subjetividades y no de los cerebros.

NATHAN: En rigor no estoy tan seguro que ello sea verdad. ¿Puedes pensar en algún ejemplo de conciencia, a cualquier nivel, humano o animal, que exista sin una dimensión sexual?

CALEB: Ellos tienen sexualidad por una necesidad de reproducción evolutiva.

NATHAN: Tal vez. Tal vez no. ¿Qué imperativo tiene una caja gris de interactuar con otra caja gris? ¿Acaso la conciencia existe sin interacción? (Garland, 2016)

Y ello porque en las condiciones del experimento, el vínculo erótico exhibe dimensiones de estimulación, de reclamo, de vínculo directo como ningún otra interacción humana.

Y esa interacción directa, háptica la que hace la dimensión sexual/genérica de Ava inevitable. Ava parece poseer perspectivas de segunda persona –como aparece en su interacción con Caleb v.g.

Ava
¿Quieres ser mi amigo?

Caleb

... Por supuesto.

AVA

¿Sería ello posible?

Caleb

¿Por qué no lo sería?

AVA

Nuestras conversaciones son unilaterales. Tú haces preguntas circunspectas, y estudias mis respuestas (Garland, 2016).

Ava evalúa situacionalmente las respuestas de Caleb, y no solo las evalúa sino que va interactuando y tratando de modificar el comportamiento de Caleb, en dirección a que este abandone la posición de un inquisidor y se conecte emocionalmente con ella. En ese sentido el comportamiento de Ava es compatible con la posesión de estados intencionales co-reflexivos: aquellos que se aplica o expone sólo cuando los puede co-aplicártelos, solo es ella cuando te es.

Supongamos que Ava sea psicópata absoluta, una pura inteligencia vincular sin capacidad empática ni emocional alguna, sin lo que llamamos en nuestro lenguaje habitual conciencia introspectiva. Ava ve la interacción social y personal como problemas de optimización entre alternativas, pero al mismo tiempo Ava tiene un comportamiento sexuado, o sea Ava flirtea.

Otro caso, que precede a Ex Machina:



- ¿Corry? ¿Corry?
- ¿Te burlas verdad? Cuando me hablas y me ves, te burlas.
- Perdona. Me lastimaste Corry.
- ¿Te lastimé? ¿Cómo pude lastimarte? (Cogiéndole del brazo con violencia) Esto no es carne, bajo esto no hay carne, ni músculos ni tendones. Eres como este automóvil, un pedazo de metal con brazos y piernas. ¡Pero él no se burla de mí como lo haces tú! No me ve con una mirada ficticia ni me habla con una voz que no existe. ¡Ya estoy hastiado de que se burlen asquerosamente de mí! Eres una mentira y sólo sirves para recordarme mi soledad y que voy a volverme loco.
- (Llorando) También sé sentir la soledad, Corry.
- Alicia, lo siento, perdóname.

The Lonely (The Twilight Zone, S01, 1959)

Podemos considerar que Ava y que cualquier mecanismo en ese nivel intelectual - resolución de problemas + aprendizaje heurístico sería lo que en filosofía de la mente se llama un zombi filosófico:

Los zombis de los filósofos son seres hipotéticos que son conductual y funcionalmente (e incluso tal vez físicamente) equivalente a (y, por tanto, indistinguible de) los seres humanos, pero que difieren de los humanos en no tener estados mentales conscientes (o, al menos, cualitativamente conscientes). Ellos tienen las mismas capacidades de procesamiento de información que tenemos los humanos y, debido a esto, una similar capacidad para formar representaciones cognitivas y quizás incluso para

entrar en estados intencionales, pero no son conscientes porque no tienen sensaciones, o *qualia* como la jerga lo tiene. Un zombie puede decirle que la rosa que está ante sí es roja, y se estremecerá y retirará rápidamente su mano si toca una estufa caliente; sin embargo, a diferencia de nosotros, él nunca experimentará núcleo de una experiencia de rojez, o el terror y la "sensación cruda" del dolor ardiente... (Thomas, 1998)

La idea del zombi filosófico se introdujo en la filosofía de la mente para atacar el funcionalismo -tanto el humano como el zombi tienen los mismos estados funcionales- y el fisicalismo -tanto el humano como el zombi tienen los mismos componentes materiales y las mismas relaciones causales.

Ahora, si un zombi no tiene estados auto-conscientes y reflexivos, entonces no tendrá estados hetero-páticos, no podrá sentir del otro en analogía a nada, ya que sus relaciones con los sujetos externos serán relaciones de optimización. Los zombis estarían moralmente muertos.

Claramente esto crea una situación éticamente compleja. Pero no surge de ello que el cuadro general sea el que dominó en la literatura durante muchos años: el conflicto entre subjetividades y zombis -sean organismos o mecanismos, en este caso nos interesan los mecanismos.

Es cierto que el modelo del conflicto es más fácil de desarrollar, pero no responde a la pregunta:

IA4 ¿Es posible la existencia de mecanismos que resuelvan problemas + aprendan heurísticamente+ tengan procesos vinculares de optimización sin ninguna hetero-patía?

Y menos aún como sería una sociedad de subjetividades y mecanismos.

O más radicalmente, que pasaría si los zombis morales estuvieran sometidos a un proceso de selección natural por lo cual irían perfeccionando sus procesos vinculares, hasta el punto que sean indistinguibles de los procesos vinculares que entablan los agentes subjetivos orgánicos.

Estos zombis carecerían de vida interior pero no sólo serían cognitivamente indiferenciables de los humanos sino que lo serían vincularmente. Y allí parecerían que el hecho de pasar el test de Escala de Hare para la Evaluación de la Psicopatía sería irrelevante.

Sería agentes estructuralmente psicópatas pero vincularmente esto no tendría efectos.

Esta es la cuestión más radical que presenta la IA. En efecto la respuesta a IA3 podría ser negativa y sin embargo la respuesta a IA4 podría ser positiva.

Podría haber un incremento de sistemas expertos y de mecanismos antropomorfos sexuados o no que tengan competencias de resolución de problemas + aprendizaje heurístico + procesos vinculares de optimización sin ninguna heteropatía, que estén sometidos a procesos de selección natural, que perfeccionen sus capacidades vinculares sin dejar de ser zombis filosóficos.

Podría ser que mecanismos como Siri o Alexia adquieran capacidades muy superiores a las que tienen actualmente, y que conformen extrañas redes sociales con agentes orgánicos intencionales.

Y este, que es el panorama más esperable, parece estar fuera de la cuestión de la IA.

Bibliografía

- Searle, J. (1980). Minds, Brains, and Programs. *The Behavioral and Brain Sciences*, 3 , 417-457.
- Bates, H. (1983). “El amo ha muerto. En J. Wynorski, *Vinieron del espacio exterior* (pág. 69–106). Barcelona: Martinez Roca.
- Calvin, W. H. (1987). The brain as a Darwin machine. *Nature* 330 , 33-34.
- Clarke, A. C. (1972). En Cruzada, *Viento del sol* (pág. 128–134). Madrid: Alianza Editorial.
- Garland, A. (23 de 06 de 2016). *Ex-Machina-screenplay*. Recuperado el 30 de 07 de 2017, de www.theguardian.org/wp-content/uploads/
- Hobbes, T. (1980). *Tratado sobre el cuerpo*. Madrid: Editorial Trotta.
- Pereira, G. (2000). *Antoniana Margarita*. Santiago de Compostela: Universidade de Santiago de Compostela/Fundación Gustavo Bueno.
- Plantinga, A. (2006). Against Materialism. *Faith and Philosophy* 23 , 3-33.
- Thomas, N. J. (1998). “Zombie Killer. En S. R. Hameroff, *Toward a Science of Consciousness II: The Second Tucson Discussions and Debates* (págs. 171–177). Cambridge, MA: The MIT Press.