FACULTAD DE CIENCIAS HUMANAS Y SOCIALES

DESARROLLO DE LA INTELIGENCIA ARTIFICIAL: LIMITACIONES Y POSIBILIDADES

Autora: Álvaro Arrieta Martínez

Director: Juan Pedro Núñez Partido

Madrid

Abril 2018

ÍNDICE

1. Introducción

La inteligencia artificial es probablemente uno de los temas más recurrentes dentro de la ciencia ficción. La idea de que una máquina llegue a exhibir las mismas capacidades que los seres humanos, o incluso llegar a superarlas, es una idea que nos aterroriza y al mismo tiempo, nos apasiona.

Las definiciones de la inteligencia artificial son muy extensas. En este trabajo resultan de especial interés las ofrecidas por Barr y Feigenbaum (1981):

> *La inteligencia artificial es la parte de la ciencia que se ocupa del diseño de sistemas de computación inteligentes, es decir, sistemas que exhiben las características que asociamos a la inteligencia en el comportamiento humano que se refiere a la comprensión del lenguaje, el aprendizaje, el razonamiento, la resolución de problemas, etc.* (Castillo et al, 2011)

Y la definición ofrecida por Rich y Knight (1991):

> *El estudio de cómo hacer que os ordenadores hagan cosas que, en el momento actual, se les da mejor a las personas.* (Russell y Norvig, 1995)

El ser humano ha destacado notablemente entre las demás especies, eso hay poca gente que lo ponga en duda. Sin embargo, el hecho de que hayamos sido capaces de colocarnos en la cima evolutiva no ha tenido nada que ver con nuestra fuerza, velocidad o agilidad, sino por nuestra capacidad de adaptación flexible al medio y habilidad para resolver problemas, entre otras características. Como señalan Arsuaga y Martínez (1999); esto, no nos convierte en una suerte de "especie elegida", solo nos hace una especie única entre millones de especies únicas, aunque eso sí, maravillosamente inteligente.

Esta idea se empezó a ver cuestionada hace algunos años. El concepto de inteligencia artificial fue creado por Alan Turing en 1950, solo 10 años después de la invención del primer ordenador computacional moderno, el *Heath Robinson*, creado por el mismo Turing. (Russell y Norvig, 1995). En 1997, el superordenador creado por IBM, *Deep Blue*, derrota en una partida de ajedrez a Gary Kaspárov, campeón mundial de ajedrez en aquel momento. El ajedrez es un juego de estrategia que requiere de respuestas flexibles a las jugadas del oponente y una gran capacidad de planificación a largo plazo; aun así, en comparación con el go es bastante sencillo. El go, al igual que el ajedrez, es un juego de tablero estratégico para dos jugadores. Debido a que el elevado número de posibles

movimientos difíciles de evaluar en su valor estratégico a largo plazo que requiere este juego son superiores en complejidad y versatilidad, crear una máquina que pudiera competir con los jugadores expertos era todo un desafío. Pero en marzo del 2016, el programa informático desarrollado por Google, *AlphaGo,* derrota a Lee Sedol, uno de los mejores jugadores de go del mundo.

Por suerte para nosotros, hay muchas más cosas que definen la condición humana que el ajedrez y el go. En términos de lenguaje, las máquinas son capaces de realizar una amplia variedad de tareas. En 2014 una máquina de aprendizaje algorítmico escribió un poema habiendo sido entrenada con las obras de Shakespeare (Lomas, 2014). El siguiente es un fragmento del mismo:

When I in dreams behold thy fairest shade

Whose shade in dreams doth wake the sleeping morn

The daytime shadow of my love betray'd

Lends hideous night to dreaming's faded form

Quizás los poemas no son una representación precisa de lo que implica la capacidad humana para transmitir información, pero las noticias escritas se acercan un poco más. En 2014 la primera noticia de un terremoto sucedido en Los Ángeles fue redactada por una IA (Oremus, 2014). Fragmento de la noticia:

> *A shallow magnitude 4.7 earthquake was reported Monday morning five miles from Westwood, California, according to the U.S. Geological Survey. The temblor occurred at 6:25 a.m. Pacific time at a depth of 5.0 miles.*

Del mismo modo que *Deep Blue* y *AlphaGo* derrotaron a los campeones mundiales de ajedrez y go respectivamente, en 2011 fuimos testigos de un evento similar en el ámbito del lenguaje. En ese año, el superordenador *Watson*, desarrollado por IBM, derrota en el programa televisivo estadounidense *Jeopardy!* a los dos vencedores históricos Brad Rutter y Ken Jennigs. El programa consiste en responder a una serie de preguntas de una amplia variedad de temas. Lo peculiar que tiene el modo de juego es que la pista que te ofrecen al jugador no es una pregunta, si no una respuesta; el jugador entonces tiene que averiguar cuál es la pregunta. Al mismo tiempo, el lenguaje natural empleado es muy rico, con muchas expresiones, frecuentemente ambiguas, sobre temas muy diversos y con

pistas muy complejas (Ferrucci et al., 2013). *Watson* funciona en base al software *DeepQA* **(***Question Answering***)**, ese sistema es muy efectivo en tareas relacionadas con el procesamiento de lenguajes naturales, la recuperación de la información, la representación del conocimiento y el razonamiento y aprendizaje automático. Las aplicaciones de esta tecnología están siendo desarrolladas con el objetivo de cumplir una serie de funciones laborales, tradicionalmente destinadas a especialistas: los sistemas expertos.

La creación de sistemas expertos es otra importante línea de avance en la I.A. Stevens (1984) define los sistemas expertos como máquinas que piensan y razonan como un experto lo haría en una cierta especialidad o campo, manejando grandes cantidades de datos y estrategias de razonamiento, además de comunicar el resultado de sus evaluaciones y lo que considera la decisión más acertada de forma que el resultado sea inteligible. Como nos cuentan Castillo et al. (2011), estás máquinas son creadas a partir de la colaboración de tres grupos: los expertos humanos especialistas, los ingenieros del conocimiento y el colectivo de usuarios al que está dirigido el programa. El proceso de comunicación entre estos tres grupos es de vital importancia. Debido a que la capacidad del sistema para transmitir información inteligible es definitoria en su valor, la calidad del lenguaje en que estos operan es fundamental. No solo porque los usuarios deban ser capaces de entender y usar correctamente aquello que les está trasmitiendo el sistema, sino porque estos programas deben ser capaces de explicar y justificar la decisión tomada, en el caso de que les sea requerido. Actualmente, estos sistemas están siendo utilizados de manera principal en campos como la economía, la industria y el diagnóstico médico. Al tratarse de ordenadores que reúnen el conocimiento de varios expertos humanos, son capaces de resolver problemas de una complejidad inabarcable para los profesionales, en un periodo de tiempo muy corto y con grado de fiabilidad muy alta; siendo su coste mucho menor a largo plazo que lo que cuesta tener contratado a un experto humano; experto humano cuyo proceso de aprendizaje pasa por años de adquirir conocimiento y experiencia. El aprendizaje de un ordenador simplemente pasa por el acceso al conocimiento y la experiencia de otros ordenadores.

Esto viene a anunciar un proceso de cambio muy significativo en la actividad laboral humana. Si la mayor parte de las tareas hasta ahora realizadas por trabajadores de cuello blanco y profesionales son cada vez más cubiertas y superadas por la IA, cada vez menos humanos serán necesarios en esos trabajos. Grey (2014) afirma que, si la creación de

músculos mecánicos apartó a los animales del proceso de producción, la creación de mentes mecánicas definitivamente apartará a los humanos del mismo.

¿Qué se puede esperar de una época en la que presenciamos de manera progresiva la aparición de entidades que desafían la excepcionalidad humana que ha sido exclusiva desde hace más de 2 millones de años? El ser humano se ha deleitado con esta idea de excepcionalidad durante gran parte de su historia. Si somos lo más alto de la creación, significa que la creación misma está a nuestro servicio. Los demás animales no son más que objetos para nuestro uso y disfrute, los bosques son nuestros para talarlos y explotar sus recursos y los océanos hacen las veces de vertederos para nuestros desechos. La aparición de algo que llegue a cuestionar esta sensación de superioridad absoluta se recibe con una cierta ambivalencia; pavimentamos el camino hacia nuestra irrelevancia al mismo tiempo que miles de científicos en ingenieros alrededor del mundo intentan conseguir lo que la naturaleza ha hecho con nosotros.

Ante esta situación de incertidumbre surgen muchas preguntas. Si las máquinas son excepcionalmente mejores que nosotros en cualquier tarea, ¿qué propósito nos queda?; si funcionan de manera más eficiente, rápida y eficaz, ¿cómo nos ganaremos el pan?; y aún más preocupante: si son superiores a nosotros en inteligencia, fuerza y funcionalidad, ¿qué les impide llegar tratarnos cómo nosotros hemos tratado a las especies, poblaciones o culturas que llegamos a considerar inferiores?

La idea de la maleficencia y el despotismo tiránico como característica de una máquina es categóricamente ficticia. Resulta extraño, aunque no imposible, pensar que la inteligencia artificial llegue a desarrollar sentimientos de odio y pulsiones genocidas hacia sus creadores. No obstante, Wångstedt (2016) realiza un paralelismo que puede servir de ayuda para ilustrar esto: hormigas. La mayoría de nosotros no tenemos profundos sentimientos de aversión y rechazo hacia las hormigas, pero si en un momento dado se encuentran entre nosotros y algo que queremos conseguir, no podría importarnos menos deshacernos de ellas. E incluso aunque nos importara, ¿cómo podríamos explicarles que se tienen que ir porque necesitamos usar ese terreno para otra cosa? O intentar explicarles a los millones de animales que viven en una selva las razones geopolíticas y socioeconómicas por las que continuamos destruyendo su hábitat. Sería imposible, simplemente carecen de la inteligencia necesaria para entenderlo.

Eso precisamente es lo que a muchas personas les preocupa, que en algún momento nos falte la inteligencia para seguir el ritmo del progreso.

¿Será la inteligencia artificial capaz de superar de forma definitiva las capacidades cognitivas humanas? ¿Llegará el momento en el que una máquina sea indistinguible de un ser humano? Si bien no podemos demostrar con lo que sabemos hasta la fecha que esto sea imposible, sí que podemos afirmar que con los conocimientos actuales que tenemos, a día de hoy es poco probable que la inteligencia artificial vaya a resultar en una simulación del 100% de las características mentales humanas.

Las limitaciones tratadas en el presente trabajo son teóricas y corresponden al momento actual de desarrollo tecnológico. La ciencia de la computación es un área en constante evolución, no solo por los problemas a los que se están enfrentando, sino a las soluciones que se proponen para solucionarlos. Aquí no solo intervienen razones técnicas, sino socioeconómicas y culturales. Los propios intereses del mercado, y las actitudes del consumidor son claves en qué aspectos se trabajan para su desarrollo y cuáles no

Como se ha mencionado anteriormente, nuestro punto de referencia de estudio son las capacidades mentales humanas, por lo que los límites se tienen en cuenta en relación dichas características.

En este trabajo las limitaciones a revisar han sido clasificadas en dos niveles: nivel cuantitativo y cualitativo. El nivel cuantitativo hace referencia a la cantidad de datos que maneja el sistema y al tiempo que tarda en procesarlos. Por otro lado, el nivel cualitativo trata aspectos más relacionados con la capacidad del sistema de dar respuestas flexibles y adaptadas a información ambigua, incompleta y circunstancial.

No obstante, conviene revisar la distinción entre los dos tipos principales de modos de trabajo de los sistemas artificiales, puesto que sus limitaciones y posibilidades están vinculadas a cómo procesan la información.

2. I.A. Simbólica e I.A. de Redes Neuronales Artificiales (PDP)

Existe una confusión generalizada sobre cómo funcionan este tipo de sistemas artificiales. Sabemos que las máquinas no comprenden, lo que hacen es seguir rutinas muy complejas. Por ello, nos cuesta entender como un ordenador es capaz de aprender y ajustar sus rutinas según sus experiencias si, técnicamente, su rutina está programada de principio a fin. Pues bien, resulta que esto último no es del todo cierto.

Tal y como nos cuenta Núñez (2012) se suele diferenciar entre dos tipos de IA, la Simbólica o Clásica y la de Redes Neuronales Artificiales o PDP (Procesamiento Distribuido en Paralelo). Sus componentes, su ciencia de programación y su funcionamiento son radicalmente diferentes, aunque ambas poseen una capacidad de computación muy alta.

La IA Simbólica procesa la información y se comporta en base a una serie de instrucciones que se organiza con una jerarquía previamente establecida. El formato de cada una de estas instrucciones se establece mediante criterios lógicos, probabilísticos o reglas y distribuciones de probabilidad que representan el conocimiento de un experto en un área concreta, tal y como se ha descrito en el párrafo sobre sistemas expertos (Castillo et al, 2011). Todas las rutinas que este sistema efectúa son cuidadosamente programadas por los ingenieros que desarrollan estas máquinas a través de algoritmos. Básicamente, se trata de conocimiento humano trasladado a la máquina. Por estas razones son muy efectivos realizando tareas específicas, como por ejemplo el ajedrez (*Deep Blue* contra Kaspárov).

Sin embargo, llega un punto en que se quieren resolver problemas que requerirían de instrucciones demasiado complejas y enrevesadas como para ser implementadas una a una por los ingenieros de programación. Para resolver este problema, se desarrolló un tipo de máquina que para resolver un problema determinado no necesita reglas ni instrucciones previamente pautadas, sino que las aprenden por ellas mismas. (Grey, 2017)

La I.A. de redes neuronales artificiales están formadas por una serie de nodos o "neuronas" interconectadas entre sí, puesto que su estructura está basada en el funcionamiento cerebral y las conexiones entre las neuronas. A este tipo de sistemas, no se les dan instrucciones o reglas que modulen el procesamiento de los datos; los ingenieros y programadores únicamente construyen la estructura de clasificación de

datos, el objetivo final de la computación, la forma de establecer las conexiones y los criterios de éxito o fracaso. A partir de aquí, se entrena a la red con una cantidad muy alta de datos. Estos consisten en problemas asociados a lo que se consideraría un buen resultado (criterio de éxito/fracaso). Por ejemplo, si quisiéramos enseñar a un sistema de redes neuronales a resolver raíces cuadradas, bastaría con entrenarla con ejemplos de raíces cuadradas ya resueltas. Estas máquinas han demostrado ser altamente eficaces resolviendo problemas de reconocimiento de patrones (p.e. reconocimiento visual de imágenes), al estar basado su aprendizaje en la detección de regularidades bien establecidas. Una dificultad que plantea este tipo de computadoras es que, para cuando el sistema ha aprendido las reglas que debe seguir para obtener un buen resultado, los ingenieros y programadores que la han construido no tienen forma de acceder a como la máquina ha realizado dichas computaciones (Núñez, 2012). Esto marca una diferencia notable a lo observado con el funcionamiento de la I.A. en forma de sistemas expertos, al tener que ofrecer al usuario una justificación de la decisión tomada. La máquina que derrotó a Lee Sedol, *AlphaGo*, usaba redes neuronales artificiales.

No obstante, la red neuronal artificial siempre tiene que tener una parte de simbólica. Los objetivos de computación y los criterios de éxito y fracaso se tienen que programar desde fuera, no pueden ser aprendidos. Por lo que no existen sistemas que se basen exclusivamente en redes neuronales artificiales. De hecho, la mayoría de los sistemas de alto nivel integran aspectos de ambos tipos en forma de sistemas híbridos neuro-simbólicos (Fernández-Riverola y Corchado, 2000).

3. Limitaciones a nivel cuantitativo

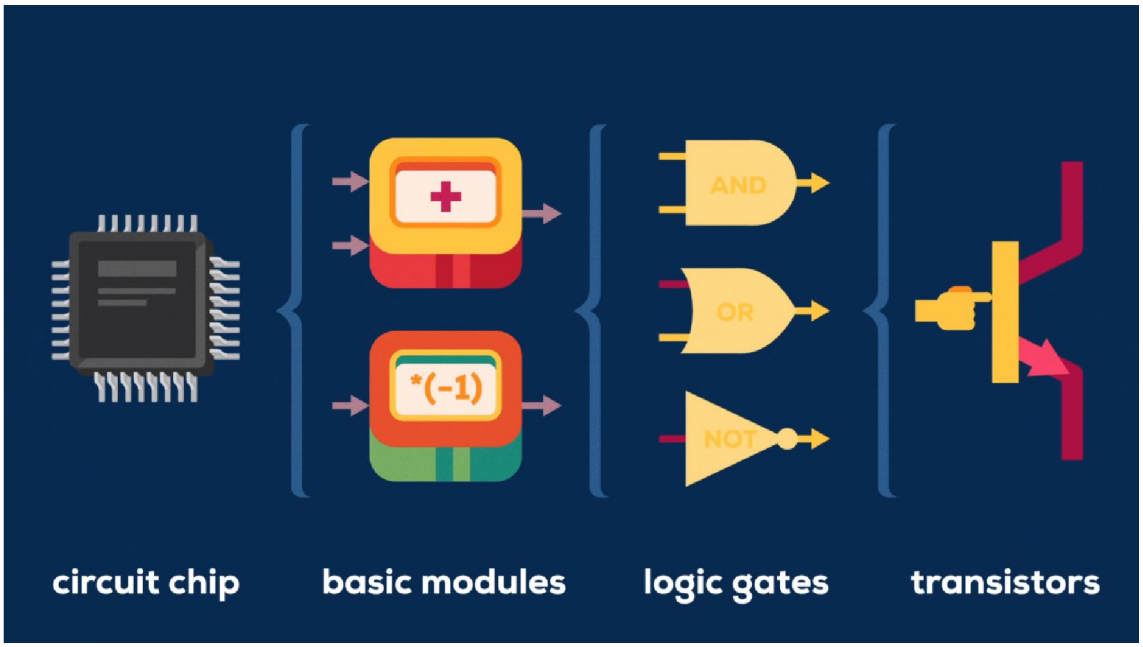
Desde 1960 el desarrollo de la computación ha sido exponencial. Cada vez nos enfrentamos a la existencia de hardwares más pequeños y a softwares cada vez más potentes (Montoya, 2004).

La cuestión de si las máquinas son capaces de superar al cerebro humano ya no es ningún misterio. Ya lo han superado, al menos en lo que se refiere a complejidad, tiempo y cantidad de datos procesados. Por poner un ejemplo, el sistema informático ganador del concurso televisivo *Jeopardy!*, *Watson*, es capaz de procesar 500 giga bytes por segundo, el cual es el equivalente a un millón de libros (Ferrucci et al, 2013). Esto nos da a entender que las capacidades cuantitativas del procesamiento computacional de la información supera ampliamente las capacidades humanas. Lo que tardaría una persona en leer un millón de libros varía según el individuo, pero probablemente estemos hablando de décadas.

Como afirma Núñez (2012) la clave de esto es el formato en que se procesan los datos: información de tipo digital. Tanto la actividad neurológica inconsciente como los ordenadores hoy en día utilizan este tipo de datos para procesar la información. Sin embargo, también existe una diferencia significativa en la base física utilizada por ambos sistemas, cerebro y ordenador. Los elementos biestables de un ordenador convencional tardan en reaccionar la millonésima parte que una neurona. Aun así, el trabajo de esta va más allá de una mera activación/desactivación, también hay que tener en cuenta las diferentes tasas de impulso nervioso; mientras que los elementos básicos de un sistema informático únicamente pueden adoptar un valor de 1 o un valor de 0. Los niveles de actividad de las redes neuronales artificiales sí que pueden adoptar todos los valores que hay entre el cero y el uno, pero ahí estamos hablando de la unidad de procesamiento, que es la red en sí, no de los elementos básicos que la componen.

Sabiendo que el nivel de complejidad de las computaciones, la cantidad de datos manejados y el tiempo que los sistemas artificiales tardan en procesarlos superan ampliamente a la capacidad humana; la pregunta no es si son mejores, sino si lo serán cada vez más. Hoy en día, es imposible saber si este crecimiento exponencial en la capacidad de computación está cerca de su fin, o nos avocamos a una revolución tecnológica sin precedentes.

Lo que sí sabemos es que actualmente este crecimiento está en un proceso de desaceleración. Tal y como ha sido mencionado anteriormente, el desarrollo de softwares cada vez más potentes está ligado a la construcción de hardwares cada vez más pequeños. Según Montoya (2004) hemos llegado a una situación previamente preconizada por la ley de Moore; es decir, hemos alcanzado un límite que pone freno a la miniaturización de los chips. En un ordenador que utiliza criptografía convencional, dichos chips están compuestos por módulos, que a su vez están compuestos por puertas lógicas, que a su vez están compuestos por transistores.



Kurzgesagt – In a Nutshell (2015) Limits of Human Technology

Los transistores son la forma más simple de procesamiento de datos en un ordenador. Básicamente funcionan igual que un interruptor de electricidad: si no está bloqueado, la electricidad pasa; y si está bloqueado, la electricidad no pasa. Estos estados se traducen en BITS, unidades de información. Si hay un bloqueo, es un 0; si no hay un bloqueo es un 1. La combinación de estos BITS es utilizada para representar información más compleja en código binario:

01010100 01101000 01100101 01110011 01100101 00100000 01110110 01101001 01101111 01101100 01100101 01101110 01110100 00100000 01100100 01100101 01101100 01101001 01100111 01101000 01110100 01110011 00100000 01101000 01100001 01110110 01100101 00100000 01110110 01101001 01101111 01101100 01100101 01101110 01110100 00100000 01100101 01101110 01100100 01110011 00101110

La escala típica de un transistor ronda los 14 nanómetros, lo cual es 500 veces más pequeño que un glóbulo rojo. Es decir, con la evolución tecnológica, los transistores se están aproximando al tamaño de unos cuántos átomos. Cuando se opera con elementos tan pequeños, el sistema empieza a manifestar problemas de funcionamiento por el simple hecho de interactuar con el mundo subatómico, donde las leyes de la física convencional no funcionan de una manera tan clara y predecible como a escalas más grandes. A escalas tan pequeñas, el transistor puede ser víctima de un efecto llamado "tunelación cuántica" o "efecto túnel", que consiste en el paso de una partícula subatómica (en este caso, los electrones que componen la corriente eléctrica) de un punto A en el espacio a un punto B, sin pasar por en medio. Esto puede llevar a que el transistor no haga bien su trabajo de bloquear el paso de la corriente eléctrica, y se produzcan fallos en el sistema.

No obstante, se encontró una manera de usar estas inusuales propiedades cuánticas para el propio beneficio del desarrollo de la computación. La solución a este problema fueron los ordenadores cuánticos. Estos ordenadores, por una serie de características físicas del hardware disponen de mucha mayor potencia de procesamiento y son capaces de realizar determinadas tareas con un grado de efectividad mucho mayor y en menos tiempo que la criptografía convencional.

Entre algunas de estas tareas se encuentran las búsquedas en bases de datos, donde usando criptografía cuántica se emplea la raíz cuadrada del tiempo que tarda un ordenador de criptografía clásica en comprobar todas y cada una de las entradas de la base de datos; lo cual es de una relevancia significativa cuando nos enfrentamos a bases de datos muy grandes. También destacan en atravesar la seguridad IT, donde gracias a su procesamiento acelerado exponencial, no necesita años de ensayo y error para averiguar la clave de desencriptación, como necesitaría un ordenador convencional.

Esto no quiere decir que los ordenadores personales con los que trabajamos cada día vayan a verse extinguidos. Actualmente no hay manera de saber si los ordenadores cuánticos van a resultar una herramienta especializada o una gran revolución para la tecnología humana. Tampoco podemos estar seguros de los límites de esta. No sabemos cuándo nos encontraremos con la próxima barrera, lo que sí hemos visto es que el progreso tecnológico está logrando sortear la ley de Moore de una manera considerablemente resolutiva. De momento no parece que este vaya a ser el fin del crecimiento computacional de las máquinas a nivel cuantitativo.

4. Limitaciones a nivel cualitativo

Cuando observamos el desarrollo de capacidades cualitativas en la inteligencia artificial, las capacidades cognitivas humanas tienen mucho más peso de comparación. Ya no se trata de crear potentes sistemas que operan con millones de datos en segundos, sino de que estos sistemas manifiesten un grado de flexibilidad y adaptabilidad que esperamos de una persona. Núñez (2012) afirma que, cualitativamente hablando, un teléfono es lo mismo que un superordenador, puesto que ambos realizan rutinas de menor y mayor complejidad respectivamente, pero ninguno tiene la capacidad de comprender y dar un significado a dichas rutinas.

Aunque sí que ha existido un crecimiento exponencial de la complejidad computacional desde 1960, en lo que respecta a las capacidades cualitativas, no ha sido tan sencillo. La implementación de sistemas que permitan a la máquina un grado humano de flexibilidad en sus rutinas de funcionamiento ha supuesto históricamente un gran desafío para los ingenieros de la computación.

Actualmente, los sistemas de alto nivel exhiben unas capacidades de flexibilidad de respuesta y manejo de la ambigüedad bastante razonables, aunque siendo ampliamente superadas por las capacidades cognitivas humanas (Kurzweil, 2012). El problema que se va a tratar en el presente trabajo no es hasta donde llegan, sino cómo lo hacen en comparación a los humanos.

Cognitivamente, los humanos resolvemos estos problemas gracias al papel que juega nuestra consciencia en los procesos mentales. Núñez (2012) define la consciencia como el "sistema compuesto por el conjunto de contenidos, actividades y procesos cognitivos de los que el organismo tiene una vivencia propia, que le permite dar cuenta de ellos en un momento dado". En los seres conscientes, como somos los humanos, hay parte de la actividad mental a la que el sujeto tiene acceso y parte a la que no; a esta última la denominamos actividad mental inconsciente. Como hemos visto en el apartado anterior, tanto los ordenadores como la actividad mental inconsciente siempre operan en un formato de datos al que conocemos por información digital o "de efecto". La característica principal de la consciencia es que este formato de datos son las propias experiencias subjetivas, a las que nos referiremos como "qualia". Los qualia son experiencias que generan un nivel de actividad excepcionalmente poderoso respecto a su

versatilidad, condensación de información, capacidad prospectiva e impacto afectivo sobre nuestra motivación.

El *leitmotiv* de este apartado es la flexibilidad, pero esta se compone de una amplia multitud de facetas y también trae consigo una serie de consecuencias fenoménicas. Por eso las revisaremos una a una, pese a que todas guarden relación entre ellas.

Los ordenadores hace mucho que se convirtieron en mucho más que grandes y estúpidas máquinas haciendo tareas repetitivas. Si bien es cierto que hoy en día siguen existiendo máquinas muy simples que cubren necesidades simples, a más pautas y criterios de decisión introducimos en un sistema, más plasticidad y flexibilidad va a desarrollar la máquina. Aunque cualitativamente, tenga las mismas limitaciones (Núñez, 2012).

4.1. Respuesta a la ambigüedad

Una faceta importante a tener en cuenta cuando hablamos de flexibilidad es la respuesta a una realidad ambigua e inestable. Los seres humanos, y demás organismos conscientes, nos enfrentamos día a día a la ambigüedad. La realidad en la que vivimos está llena de datos nuevos, incompletos y cambiantes. Por eso mismo, el tipo de procesamiento mental en humanos que predomina en estas situaciones es el pensamiento heurístico. Según Pérez y Bautista (2014) el pensamiento heurístico surge de las limitaciones estructurales en el procesamiento de la información. Se trata de mecanismos que sirven para hacer que la incertidumbre producida por nuestras propias limitaciones cognitivas se disminuya a la hora de enfrentarnos a una serie de estímulos ambientales de un grado alto de complejidad, transformándolos a una dimensión manejable por nuestro sistema. Con este objetivo reducen tareas complejas a juicios sencillos, prescindiendo de la necesidad de un análisis exhaustivo de toda la información disponible. Núñez (2012) explica que los criterios del filtro selectivo de información que pasa a consciencia, es decir que se manifiestan en formato qualia, es la relevancia de dicha información. La relevancia puede estar marcada por intensidad del estímulo, su duración, su importancia afectiva y su relación con lo que está activado en conciencia en ese momento. Por eso mismo, la información en la que un sujeto humano se fijará en una situación en particular tiene que ver con esos criterios del filtro selectivo que hacen posible el razonamiento heurístico.

Los heurísticos son también conocidos en el campo de la psicología cognitiva como sesgos o prejuicios cognitivos. Como es de esperar, los sesgos y prejuicios son una fuente de innumerables errores en nuestro pensamiento. Uno de los más comunes es el heurístico

de representatividad, que permite valorar en qué medida una muestra es representativa de un modelo. De esta manera, los elementos más prototípicos en una categoría serán mejor recordados (Pérez y Bautista, 2014). Por ejemplo, a la hora de determinar si un alumno en un campus universitario está estudiando ciencias políticas o derecho, si viste de manera muy informal, probablemente el sujeto determine que es un estudiante de ciencias políticas; incluso aunque el sujeto sea conocedor de que hay el doble de estudiantes de derecho respecto a los de ciencias políticas en dicho campus. El sujeto, de esta manera, estaría violando la regla de Bayes, en términos probabilísticos.

A pesar de los errores que se puedan cometer usando este mecanismo, el hecho de tener la capacidad de usarlo proporciona un sinfín de ventajas evolutivas. La actividad consciente, a través de heurísticos, posee la habilidad de resolver problemas en ausencia de reglas mentales específicas, donde más vale tomar una decisión y llegar a un resultado en el menor tiempo posible, que analizar el total de la información disponible.

En lo que respecta a las máquinas, o bien están compuestas por sistemas que necesitan reglas programadas y establecidas (I.A. Simbólica) o bien estos sistemas las aprenden por ellas mismas (I.A. de redes neuronales artificiales). Sin embargo, incluso aprendiéndolas por sí mismas, la cantidad de datos que necesitan para entrenar al sistema y resolver un problema es infinitamente superior al que necesitaría un humano enfrentándose a una situación desconocida.

Los contenidos en conciencia interactúan de una manera cuyo resultado es imposible de determinar. Podemos señalar otros heurísticos que también son comunes, como el heurístico de disponibilidad o el de anclaje, pero eso no quiere decir que dichos sesgos se vayan a producir siempre y del modo que esperamos. La manera que tienen los seres humanos de enfrentarse a la ambigüedad supone una habilidad que transgrede cualquier principio matemático y, por lo tanto, computable. Esta habilidad es la función propia de la actividad consciente de operar con unidades de diferentes categorías. En otras palabras, restar peras a manzanas (Núñez, 2012). Esto ofrece la posibilidad de tomar decisiones teniendo en cuenta al mismo tiempo múltiples facetas de la realidad con la que interactuamos. Como caso ilustrativo: El trabajador con síndrome de burnout que quiere dejar su trabajo tendrá que valorar muchas dimensiones para tomar la decisión: el valor de su salud mental, cuánto está dispuesto a ponerse a él y a su familia en riesgo económico, cuanto valora el viaje con su mujer que lleva meses planeando, etc. Para ello,

su mente tiene que mezclar en la misma operación el miedo a la enfermedad, el miedo a no poder pagar la universidad de sus hijos y la ilusión prospectiva del viaje.

Debido a que las máquinas, y la actividad mental inconsciente, operan utilizando información digital, son incapaces de realizar este tipo de operaciones. Y, por lo tanto, los parámetros de decisión siempre tienen que estar bien definidos. La información integrada de vivencia interna o qualia, no tiene esa necesidad. Hoy en día, ignoramos cómo es esto posible, debido a que desconocemos la naturaleza última de la consciencia y su relación con el tejido neurológico. Aquí entra el debate histórico que enfrenta al dualismo, monismo materialista, monismo espiritualista y funcionalismo. Nosotros apostamos por un tipo de monismo materialista conocido como emergentismo, que establece que la consciencia es el resultado del fenómeno producido por la interacción de sus componentes, y sus propiedades no son reducibles a las de sus partes constituyentes.

Sin embargo, hemos mencionado previamente que la inteligencia artificial exhibe unas capacidades de flexibilidad de respuesta y manejo de la ambigüedad bastante razonables. Si esto es así, ¿cómo es posible que resuelvan estos problemas sin las características habilidades del procesamiento consciente de la información?

La actividad consciente siempre tiene una base en la actividad inconsciente. Es una realidad que, a mayor complejidad del sistema inconsciente, mayor complejidad consciente tendrá la actividad mental. No hay que considerar al consciente y al inconsciente como dos módulos separados, sino como dos formatos de procesamiento de la información dentro de la misma actividad mental, que operan con datos cualitativamente diferentes (Núñez, 2012). El inconsciente, al operar con información digital, es reproducible por los sistemas artificiales, a los que es posible implementar determinados trucos en el conjunto de sus operaciones para lograr resultados que se aproximan a un funcionamiento flexible de respuesta a la ambigüedad (Kurzweil, 2012). Debido a que el desarrollo de la IA es un área extremadamente comercial, los métodos técnicos que utilizan las corporaciones que desarrollan estas máquinas de alto nivel son de difícil acceso al público por el secretismo que supone el mantenimiento de la patente y el copyright. Lo que también significa que las técnicas utilizadas por diferentes empresas son muy diferentes entre ellas. La manera de resolver un problema visoespacial un superordenador desarrollado por Google no tiene nada que ver con cómo lo resuelve

uno desarrollado por IBM; y a su vez un superordenador desarrollado por Apple no tiene nada que ver con ninguno de los otros dos.

4.2. Objetivos y criterios de éxito/fracaso

Cuando hablamos sobre el propósito último del sistema y cómo va a determinar qué es un buen resultado y qué es un mal resultado, es una referencia directa al debate sobre la libertad y el libre albedrío. Como se ha mencionado en el apartado sobre tipos de inteligencia artificial, los ingenieros que construyen los ordenadores que integran redes neuronales artificiales establecen el objetivo final de la computación y los criterios de éxito o fracaso. También se mencionó que estos elementos, no pueden formar parte de la red neuronal en sí, sino que se tienen que establecer a través de algoritmos, es decir, inteligencia artificial simbólica. Esto nos viene a decir que los criterios que guían los objetivos de acción han sido prestados por una consciencia humana. Núñez (2012) explica que, para que el propio sistema pudiera deshacerse de dichos criterios y obtener unos nuevos, necesitaría disponer de consciencia, o al menos un mecanismo similar. Bastaría con un dispositivo autorreferencial que le permitiera experimentar subjetivamente y operar con información integrada de vivencia interna, en lugar de información digital. Esto le proporcionaría la capacidad de detectar qué es bueno y qué es malo tanto en situaciones conocidas como en desconocidas, excluyendo la necesidad de detectar las regularidades de las consecuencias de sus acciones y de la interacción con los estímulos. Razón que está detrás de que la mayor parte de nuestro pensamiento esté basado en heurísticos.

En humanos, los objetivos de acción están ligados a la experiencia emocional. Sin embargo, no están tan definidos como en el caso de las máquinas. Esto no implica la prueba definitiva de la libertad en humanos, pero un argumento frecuente es que la relación entre los qualia y sus combinaciones posibles son imposibles de determinar (Núñez, 2012). Obviamente, la indeterminación no equivale a libertad, sin embargo, sí que es un requisito indispensable para que esta pueda tener lugar.

No obstante, la relevancia funcional de la consciencia está muy discutida actualmente. Autores como Searle, Chalmers y Dennet defienden que la actividad mental consciente no tiene capacidad causal de influir en la conducta del individuo.

Para ilustrar las ventajas evolutivas de adquirir unos criterios de éxito/fracaso basados en la experiencia subjetiva, Johnson-Laird propone un experimento hipotético conocido como "*La Ratita Sedienta*" (Núñez, 2012):

Supongamos dos ratas que están perdidas en el desierto, ambas altamente deshidratadas, Una de ellas tiene la capacidad de procesar en formato consciente y la otra sin esta capacidad, solo actividad inconsciente. Llegan a un punto en el que se encuentran con los restos de un campamento humano, donde encuentran restos de latas de gasolina, refrescos y cerveza; pero no agua. Para la rata no consciente, el problema al que se tiene que enfrentar es la deshidratación, y si la única regularidad que ha establecido es que la hidratación se soluciona con agua, solo buscará agua; ignorando los demás líquidos. Al no encontrar agua, abandonará el campamento sin beber nada y morirá por deshidratación, pero sin haber tenido sensación de sed y sufrimiento. Sin embargo, para la rata consciente el problema no es la deshidratación, sino la sed. Por lo tanto, buscará cualquier cosa que le haga calmar la sensación subjetiva de sed, incluso sin tener experiencia previa con dicha cosa. Puede que beba la gasolina y muera por su toxicidad; pero también puede que beba la cerveza y, emborrachándose, sobreviva. O que beba el refresco y sobreviva sin efectos adversos.

Esto está estrechamente relacionado con la respuesta a una realidad ambigua y desconocida. Definitivamente, la capacidad para calmar la sed bebiendo gasolina supone un peligro para el organismo, de la misma manera que el pensamiento heurístico puede llevarnos a un error en nuestro juicio. Pero poseer la capacidad de hacerlo, nos coloca en una posición privilegiada en lo que respecta a nuestra adaptación al medio, respecto a los demás organismos no conscientes.

Ahora mismo, nuestros superordenadores, sistemas expertos y demás sistemas informáticos son la rata sin consciencia. Pueden salvar determinadas limitaciones gracias a la implementación de trucos, pero llegar a superar esta brecha supondrá un cambio en el paradigma de procesamiento artificial de la información hasta ahora utilizado.

4.3. Lenguaje

Como hemos visto en la introducción, en términos de lenguaje la inteligencia artificial es capaz de realizar una amplia variedad de tareas, entre ellas escribir poemas y redactar noticias de actualidad. No obstante, dominar el lenguaje supone más que establecer un proceso de comunicación unidireccional.

Para entender por qué es importante el lenguaje dentro del análisis de la flexibilidad en las máquinas, tenemos que fijarnos en dos aspectos: manejo del lenguaje y comprensión del lenguaje. Lo que viene a ser el aspecto sintáctico y el aspecto semántico, donde ambos están estrechamente relacionados.

Como vimos con *Watson* y el programa *Jeopardy!*, dominar el lenguaje natural supone enfrentarse a expresiones coloquiales, dobles sentidos y ambigüedades mientras, en el común de las conversaciones, se tratan temas muy diversos uno detrás de otro. Esto último destaca un área importante en el desarrollo de la inteligencia artificial. El hecho de que una máquina, cuando es especialmente buena en una tarea, solo sea buena en esa tarea; mientras que un humano puede ser razonablemente bueno en un gran número de actividades. Esto claramente supone una limitación en la evolución tecnológica, pero no necesariamente supone un problema. Dado que el desarrollo de las sociedades humanas siempre ha estado ligado a la división y especialización del trabajo, disponer de máquinas hiperespecializadas exclusivamente en una determinada tarea, mientras son inútiles tratando de resolver cualquier otra, no convierte a esa máquina en inempleable para el sistema socioeconómico, sino todo lo contrario.

Watson fue la demostración de que un sistema informático podía dominar tareas concretas usando el lenguaje natural. La aplicación de su tecnología es utilizada actualmente para la construcción de sistemas expertos. Como vimos, el aspecto comunicacional y lingüístico es fundamental para estas máquinas. El lenguaje con el que programan estos softwares viene a ser una integración entre las aportaciones de los expertos humanos especialistas, los ingenieros del conocimiento y los usuarios a los que está dirigido el programa; es decir, terminología profesional, lenguaje de programación y lenguaje coloquial, respectivamente. Para sistemas expertos en el campo del diagnóstico médico, el sistema recibiría en el input expresiones más o menos coloquiales sobre el problema o la dolencia en cuestión. Entonces, tendría que traducirlo al lenguaje terminológico

registrado en la base de datos y ofrecer de output una serie de preguntas para recabar información sobre el problema (Castillo et al., 2011).

Este modo de funcionamiento responde a un modelo de interfaz usuario-sistema experto que en estos sectores se ha denominado *chatbot*, agente conversacional o sistema de diálogo. Como explican Shawar y Atwell (2007) la función de los chatbots está extendida en áreas muy dispares del proceso comercial en la informática y la inteligencia artificial. Pueden servir como herramienta de atención al cliente, de búsqueda y recuperación de la información o como un mecanismo para el mero entretenimiento del usuario.

El primer chatbot del que tenemos conocimiento, ELIZA, fue creado en 1960 con este propósito: entretener. Para desempeñar su función, ELIZA emularía a un psicoterapeuta rogeriano en un contexto clínico. Lo siguiente es una muestra de una conversación con este bot:

```
Human: Men are all alike.
ELIZA: In what way?
Human: They're always bugging us about something.
ELIZA: Can you think of a special example?
Human: Well, my boyfriend made me come here.
ELIZA: Your boyfriend made you come here?
Human: He says I'm depressed much of the time.
ELIZA: I'm sorry to hear that you are depressed much of
the time.
Human: It's true. I'm unhappy.
ELIZA: Do you think coming here will help you not to be
unhappy?
Human: I need some help, that much seems certain.
```

También hacía cosas como que, si en el input se encontraba la palabra "madre", ELIZA respondía al usuario "cuéntame más sobre tu familia". Este funcionamiento estaba inspirado en las teorías sobre la madre y la familia como centrales a los conflictos psíquicos en la persona (Shawar y Atwell, 2007).

No hace falta señalar que ELIZA no hacía un trabajo perfecto en el campo de la psicoterapia. Sin embargo, sentó las bases del trabajo de los chatbots como herramienta dentro de sistemas expertos. Una interfaz que pudiera conectar al usuario con el sistema a través de un lenguaje comprensible y que fuera la propia máquina la que hace la traducción de la terminología profesional, y que sea capaz de justificar la decisión tomada es definitivo en la construcción de sistemas expertos.

ELIZA fue construida en 1960, año en el cual empezó el crecimiento exponencial en el desarrollo de la inteligencia artificial. Siendo esto así, podríamos esperar que las habilidades de los chatbots hubieran mejorado de forma inimaginable y que actualmente fueran indistinguibles de un ser humano. Pero lo cierto es que no es así.

Como hemos visto, el desarrollo tecnológico en las capacidades cualitativas de la I.A. no se está dando al mismo ritmo que el desarrollo de las capacidades cuantitativas. Y el lenguaje tiene mucho de cualitativo. Requiere complejidad computacional, sí; pero no es lo único. Como cualquier máquina de alto nivel, son eficaces en tareas muy específicas. El lenguaje está cargado de ambigüedad, dobles sentidos y en una conversación hay que seguir el hilo de muchos temas que se dan al mismo tiempo, incluso de contenido que no se ha hecho explícito. Por eso mismo, los chatbots actualmente dejan mucho que desear. Al menos, a lo que tenemos acceso. Es posible que existan chatbots que funcionen de una manera mucho más similar a como lo haría un humano, pero no estén disponibles al público por motivo de las aplicaciones específicas de esta tecnología en las grandes empresas. De manera que no hay modo de saberlo con seguridad. Como se ha dicho antes, el desarrollo de la I.A. es muy comercial y la parte científica y de investigación está estrictamente ligada a lo que nos permite la primera.

Sin embargo, merece la pena mencionar que, incluso con el desarrollo de los primeros chatbots, había personas que creían firmemente que la entidad con la que se estaban comunicando, era una persona (Shawar y Atwell, 2007). Aquí ya nos empezamos a aproximar al segundo aspecto del lenguaje en las máquinas: la comprensión.

Alan Turing, después de crear el *Heath Robinson* y, 10 años después, el concepto mismo de inteligencia artificial; reflexionó mucho sobre si las máquinas eran inteligentes y si podrían llegar a pensar como los humanos. Para tener la posibilidad de acercarse a este problema, transformó la pregunta por "¿pueden las máquinas hacer lo que nosotros hacemos?". Con este objetivo creó el Test de Turing, una prueba que consistía en que un

sujeto detectara, a través de conversaciones escritas, si la entidad con la que estaba hablando era una máquina o una persona (Russell y Norvig, 1995).

Existe una fuerte discusión sobre las reglas que debe de cumplir esta prueba, ya que la implementación de las mismas es definitoria a la hora de decidir si actualmente hay máquinas que la pasan o no. Como hemos visto, muchos usuarios pensaban que el programa ELIZA era una persona humana al otro lado del monitor. Si definimos el Test de Turing como la capacidad de la máquina para hacerse pasar por un ser humano, la máquina lo pasa fácilmente. De hecho, hasta *Cleverbot*, un chatbot gratuito (online) bastante mediocre, podría pasar este formato de prueba, dadas las condiciones adecuadas (Stevens, 2017).

Sin embargo, si definimos el Test de Turing como la capacidad de la máquina para realizar las mismas funciones lingüísticas que realizaría un humano, entonces no hay necesidad de enmascarar la existencia de la máquina en la prueba. Mediante este formato, pasar el test de Turing supone un reto mucho más exigente, tanto que ninguna máquina a día de hoy lo ha pasado (Velasco, 2017).

No obstante, Turing no solo planteó esta prueba con el objetivo de evaluar el manejo del lenguaje, sino como una medida de la propia inteligencia de la máquina. Si para Turing la prueba de que una máquina pudiera pensar como una persona radicaba en el hecho de que pasara esta prueba, entonces el test de Turing sirve para saber si la máquina comprende el lenguaje tal y como lo comprenden los humanos. Pero no es ninguna sorpresa que esto no vaya a resultar tan sencillo. El concepto de comprensión computacional del lenguaje ha estado y sigue estando muy discutido entre multitud de expertos e intelectuales.

Gelepithis (1984) defiende que, dado un robot ideal, R, sea un robot a) equipado con sensores funcionalmente equivalentes a los de los humanos; b) es capaz de manejar información lingüística; y c) es capaz de conectar de cualquier manera a) con b). Asumiendo que R sea construido, entonces R es, en principio, incapaz de entender el lenguaje humano. Este autor cree necesario redefinir los conceptos de significado y comprensión para este problema.

En la búsqueda de redefinir estos conceptos, Searle en 1980 (Velasco, 2017) postuló que el experimento del test de Turing vale tanto para medir la comprensión del lenguaje como el siguiente planteamiento:

Supongamos un hombre encerrado en una habitación con un libro de instrucciones en inglés de como interactuar en chino. Un segundo hombre, que habla chino, se acerca a la puerta de la habitación donde está encerrado el primero, y se empieza a comunicar con él mediante mensajes escritos. En segundo hombre escribe los mensajes en chino, y el primero utiliza en libro para saber que lo tiene que responderle. En este proceso el hombre de fuera tendrá la sensación de que el hombre de dentro sabe chino. La cuestión es, ¿sabe el hombre encerrado en la habitación realmente chino?

Por supuesto, Searle defiende que el hombre encerrado no sabe chino, y probablemente todos estemos de acuerdo con esa afirmación. El autor utiliza esta metáfora del test de Turing para señalar que los ordenadores utilizan elementos sintácticos de la lengua, no semánticos.

Por otro lado, Chalmers ofrece el contraargumento de que el sistema no está compuesto únicamente por la persona encerrada en la habitación, sino por todo el conjunto de objetos con lo que se ha interaccionado, incluyendo el libro de instrucciones (Núñez, 2012). De esta manera, Chalmers defiende que, aunque el individuo no sabe chino, el sistema compuesto por todos los elementos de la habitación sí sabe el idioma.

En lo que respecta a nuestra postura, nuestra posición es más cercana a la de Searle que a la de Chalmers. Aunque ambos defienden la irrelevancia funcional de la consciencia, de manera que no termina de encajar del todo con ninguno. Velasco (2017) afirma que el propio test de Turing cae en el error de confundir epistemología y ontología; es decir, el cómo con el qué, o la sintáctica con la semántica. Parece que está bastante claro que el debate sobre el lenguaje en la inteligencia artificial es una lucha del manejo del lenguaje contra la comprensión del mismo. Nosotros consideramos estos dos aspectos dos caras de la misma moneda, ya que la capacidad de tener una referencia directa del significado es clave para conseguir un manejo adecuado, flexible y adaptado del lenguaje.

4.4. Creatividad

Cuando hablamos de creatividad, nos suele venir a la mente un cúmulo de ideas romantizadas como “sacar algo de la nada” o “generar algo nuevo”. A veces podemos llegar a considerar que es un proceso místico o divino. Si partimos de conceptos mitificados cuando pensamos en la creatividad en humanos, aún más desconcierto y rechazo nos va a provocar la creatividad artificial.

De la misma manera que en el lenguaje hay que redefinir los conceptos de significado y comprensión, en la creatividad debemos hacer lo propio. Como señala Cabañes (2017) a la hora de establecer un concepto claro de creatividad debemos tener en cuenta que nada puede surgir del vacío, y toda idea creativa requiere una serie de esquemas histórico-culturales previos y de experiencias vividas por el individuo. Así pues, definimos la creatividad como el producto de establecer relaciones nuevas entre datos que ya se poseen. De lo que podemos sacar que a más datos, experiencias y conocimientos tenga el individuo, más altas son las posibilidades de encontrar una relación creativa. Claro que, también será necesaria una capacidad especial de relacionar dichos datos. Especial, porque a mayor disparidad de reinos conceptuales entre los que se establezcan las relaciones, más creativo va a ser el resultado. De manera que tenemos dos factores para que se produzca el proceso creativo: acumulación de datos y capacidad especial para relacionarlos.

En humanos, esta capacidad de relación especial tiene que ver con la naturaleza de la experiencia consciente. Cabañes (2017) recalca como ejemplo de esto el fenómeno sinestésico. La sinestesia va más allá de los episodios extremos provocados por el consumo de sustancias estupefacientes u otros trastornos mentales. Se trata de un fenómeno existente en el funcionamiento mental de todos nosotros: la asociación sensorial. Asociación que no es aleatoria, sino que parece seguir un patrón. Por ejemplo, las notas musicales más graves tienden a ser asociadas con colores más oscuros. Los artistas relacionan sus experiencias con su material de trabajo: colores, sonidos, formas…

Incluso el origen del habla y las lenguas pudo haberse dado por la influencia de procesos sinestésicos. Según Ramachandran (Cabañes, 2017) los movimientos de los labios y de la lengua al hablar mantienen una relación con los gestos que hacemos con las manos al hablar, como queriendo representar espacialmente el objeto al que nos referimos.

Precisamente, esto es posible por la capacidad de la consciencia de operar con datos de diferentes categorías. Relacionar color con temperatura, sonido con forma, emociones con música, es algo que solo puede ser posible cuando esta información es procesada en modo qualia. De la misma manera que los humanos podemos responder a la ambigüedad, podemos relacionar datos o representaciones mentales de una manera original y, como se ha dicho antes, impredecible.

Puesto que las máquinas operan exclusivamente con información digital, sabemos que no hay forma de que las máquinas puedan tener un proceso creativo del mismo modo que nosotros. Pero el caso es que sí hay máquinas que diseñan obras artísticas, lo que por defecto las convierte en creativas.

Csikszentmihslyi (Cabañes, 2017) afirma que la creatividad es el producto de la interacción entre tres elementos:

1. Una cultura con reglas simbólicas
2. Un individuo que aporta novedad al campo simbólico
3. Un ámbito de expertos que reconocen la innovación

Csikszentmihslyi acierta en que la creatividad, como muchos otros fenómenos en el desarrollo de la persona, surge de la interacción social. Si queremos que las máquinas sigan estas "reglas de juego", tendremos que dotarlas de los recursos para acceder a estos elementos, exceptuando el punto 2 que sería la máquina en sí.

Según Cabañes (2017) hacer que el programa sigua las reglas simbólicas de la cultura, se puede llevar a cabo introduciéndole directamente las reglas y definiciones simbólicas, entrenando a la máquina con ejemplos de obras creativas o entrenándola en un contexto cultural. En la introducción se mencionó una máquina que, habiendo sido entrenada con las obras de Shakespeare, escribió su propio poema. Esta máquina es un ejemplo de la introducción de las reglas simbólicas mediante ejemplos. Como vimos en las redes neuronales artificiales, estas son entrenadas usando ejemplos ya resueltos sobre qué se considera un buen resultado qué se considera un mal resultado (criterios de éxito/fracaso). La ventaja que tiene esto es que con una definición simbólica solo das información explícita, mientras que con ejemplos almacena todos los datos que podemos pasar por alto.

Dicho esto, queda claro que la máquina necesitaría del algún nexo o interfaz con la sociedad a la que pretende introducirse artísticamente. Y hoy en día esto es más fácil que nunca. Nos encontramos en un momento de la historia en el que estamos recogiendo datos sobre infinidad de aspectos de nuestras vidas: patrones de clima, registros médicos, sistemas de comunicación, datos de viajes, tareas laborales y, por supuesto, comportamiento humano. Creando así grandes bibliotecas que las máquinas pueden usar aprender sobre las cosas que nos gustan y como crear otras que nos gusten más.

Al final el mercado el que decide qué de todo lo que crea la máquina le gusta más. Y por lo tanto el ámbito de "expertos" que reconoce la innovación como creativa es el conjunto de los consumidores de la obra artística en cuestión.

La industria de la música es un buen ejemplo de esto. Puede que las máquinas no tengan mucha idea sobre los gustos de los coleccionistas de arte costumbrista, pero definitivamente cada vez tiene más información sobre los gustos musicales de la gente. Qué escuchamos, cuándo lo escuchamos, con quién lo escuchamos, cómo se trasmite una canción que se vuelve viral son solo algunas de las preguntas a las que los Big Data les resulta cada vez más fácil responder.

Las máquinas artistas no son ciencia ficción, ya están entre nosotros. Hay máquinas que pintan y hay máquinas que componen música. Y a diferencia de los chatbots, estos sí que pasan su propio test de Turing. *Emily Howell* es un bot que puede estar componiendo de manera constante piezas de música. La gente que la escucha es incapaz de diferenciar su música de la de compositores humanos en pruebas con los mismos criterios que el test de Turing en su versión exigente (Grey, 2014). Es cierto que componer una pieza de música no requiere de las vastas habilidades cualitativas necesarias para mantener una conversación sobre cualquier tema por tiempo indefinido. Sin embargo, es interesante como algo que considerábamos tan característico de lo humano, como es la creatividad artística, este siendo cubierto de una manera tan efectiva por la inteligencia artificial.

Obviamente, todas estas tareas se realizan sin intervención alguna de la consciencia. Las máquinas cuando crean música no están integrando emociones y notas musicales, ni están convirtiendo experiencias pasadas en melodías. Pero están haciendo que nosotros si identifiquemos esas melodías con emociones, que nos muevan y nos causen las mismas sensaciones que, en el fondo, buscamos en los compositores humanos.

4.5. Implementar la consciencia

Después de haber repasado y analizado todos estos puntos, la siguiente cuestión obligatoria es si las máquinas dispondrán, en algún momento de su desarrollo, de consciencia; haciendo que sus limitaciones cualitativas desaparezcan y puedan desarrollarse junto a sus capacidades cuantitativas; donde empezarán a evolucionar exponencialmente hasta esferas inconcebibles por el intelecto humano.

La respuesta a eso: no está claro. Actualmente y con el conocimiento que a día de hoy tenemos resulta poco probable que las máquinas sean dotadas de un dispositivo similar a nuestra consciencia, aunque eso no quiere decir que sea imposible (Núñez, 2012). Claro que, para establecer que esta cuestión tenga la importancia que le damos en el presente trabajo, tenemos que asumir de partida que la consciencia tiene relevancia funcional y capacidad causal en nuestra conducta. Cosa que ya hemos visto que no todos los autores están de acuerdo con eso.

Tratar de definir qué sistema es consciente y cual no es una tarea más complicada de lo que puede parecer a priori. Sabemos que algunos animales no son conscientes, sabemos que otros si lo son, pero también hay muchos en los que no nos queda claro (Núñez, 2012). Entre humanos, utilizamos el lenguaje para establecer que otra persona es consciente. En animales, se utiliza la respuesta análoga al dolor, entre otras técnicas. Pero no tenemos ninguna manera de comprobar si un sistema artificial se ha vuelto consciente.

Por los propios intereses del mercado, a veces interesa que el sistema parezca que tiene consciencia. Este es el caso de los chatbots y los robots humanoides, donde lo que muchas veces interesa al consumidor es que el sistema parezca que tiene sentimientos, parezca que se preocupa por el usuario o parezca que está excitado sexualmente. Por otro lado, al mercado tampoco le interesa en principio que sus creaciones puedan llegar a comportarse consecuentemente con un procesamiento computacional consciente. A nadie le interesa que una herramienta tenga libre albedrío, capacidad para cansarse, enfadarse o revelarse (Núñez, 2018).

Hay que tener claro que crear una simulación artificial perfecta de un humano, no significaría que el sistema sea consciente. Una simulación, por precisa y exacta que sea, es tan real como una obra de teatro, donde puedes percibir ira, deseo, conflicto, alegría, etc., pero sabes que son actores interpretando un papel. Núñez (2012) denomina a esto el "como sí"; como sí sintiera, como sí le doliera, como sí procesara conscientemente; pero

este sistema seguiría tan carente de experiencias subjetivas como el termostato más simple.

Velasco (2017) nos cuenta que cuando la humanidad quiso construir una máquina que volara, solía fijarse en las aves con el objetivo de imitar su técnica de vuelo, y traspasarla así a la máquina; unas alas que se agitaran y una cola que marcara la dirección. Parece que una imitación exacta de un pájaro no era la mejor idea, pero lo cierto es que hoy los aviones tienen alas y cola, aunque su funcionamiento difiera significativamente del de un ave.

Por supuesto, el principal problema de pretender implementar voluntariamente la consciencia a un sistema artificial es que desconocemos la naturaleza de esta y su relación con el cerebro. No podemos tratar de crear algo que ni siquiera comprendemos cómo es posible su existencia en primer lugar.

Claro que, puede darse el caso de que la aparición de la actividad consciente en los sistemas artificiales no sea una decisión de los ingenieros que estén construyendo la máquina, sino que surja por sí misma debido a la propia y creciente complejidad del sistema. Esta hipótesis se denomina el paradigma de la complejidad (Núñez, 2012).

El problema del paradigma de la complejidad es que, de primeras, entiende la consciencia como una cuestión cuantitativa, no como un salto cualitativo, tal y como lo entendemos nosotros. Sin embargo, es comprensible que se le de este enfoque teniendo en cuenta que, en los sistemas biológicos, la consciencia correlaciona con complejidad. Los animales de los que tenemos conocimiento sobre su actividad consciente son en su mayoría vertebrados (Núñez, 2012). Lo que quiere decir que, por lo menos, la complejidad es un requisito para que surja la consciencia, pero no la garantiza.

Otro argumento a favor del paradigma de la complejidad es el proceso por el cual se entrena a las redes neuronales artificiales. Sabemos que, biológicamente, la consciencia emergió cuando determinada especie llegó a un punto “x” de complejidad neurológica después de millones de años de evolución y selección natural. Y, ¿qué es la selección natural sino ensayo y error? Como nos cuenta Grey (2017) en proceso por el cual se entrena a las redes neuronales no es entrenando a una sola red, sino entrenando a muchas. En los primeros ensayos los sistemas dan respuestas prácticamente aleatorias, después se evalúan (con resultados pésimos en los primeros ensayos), se elige al mejor sistema y se

descartan los demás; se hacen copias del mejor, se introducen pequeñas variaciones aleatorias y se vuelven a evaluar; ya sí sucesivamente. La cuestión es, ¿qué tan diferente es este sistema del ensayo y error que plantea la evolución de las especies y la selección natural por genética?

Este argumento es altamente reflexivo y especulativo, pero resulta interesante porque nos confronta con el desconocimiento y la falta de control que tenemos sobre este aspecto. Puede que ni siquiera sea necesaria la implementación o el surgimiento de la consciencia para que la inteligencia artificial supere sus limitaciones cualitativas. Aunque hoy día no vemos otra solución, en frecuentes ocasiones el progreso tecnológico sigue avanzando sin que seamos capaces de concebir cómo puede ser posible.

6. Transhumanismo

A lo largo del presente trabajo, hemos observado que las máquinas, aunque se están desarrollando de manera efectiva, todavía tienen que salvar determinadas limitaciones. Sin embargo, llama la atención la insistencia mediática que estamos viviendo en nuestra época sobre qué cosas pueden hacer y hasta dónde pueden llegar. Es natural que se viva la llegada de la inteligencia artificial como una amenaza. A fin de cuentas, las grandes obras de ciencia ficción que tratan este tema son en su mayoría distópicas, y presentan un futuro terrorífico, inseguro o carente de libertad. Pero ¿hasta qué punto será la inteligencia artificial una entidad con la que tendremos que competir y no el futuro de nuestra propia identidad?

Como explica Koval (2011) el concepto de singularidad tecnológica hace referencia al punto en el tiempo donde las máquinas se vuelven más inteligentes que los humanos en todos los aspectos; y donde sean ellas mismas las que diseñen y apliquen sus mejoras y actualizaciones. El término se desarrolló originariamente en el campo de la astrofísica para el modelo del agujero negro. Y fue tomado prestado por el área tecnológica y la inteligencia artificial para referirse al mismo concepto: un cambio cualitativo y trascendental donde los modelos científicos ya no sirven para realizar predicciones y proponer explicaciones, y consecuentemente deben de ser remplazados por un paradigma nuevo de conocimiento.

De esta afirmación se deduce que este va a ser un apartado categóricamente especulativo. Los sucesos que acontezcan pasada la singularidad son extremadamente complicados de predecir o imaginar, incluso contando con la mejor información disponible. Lo que si podemos afirmar es que el camino hacia la singularidad pasa necesariamente por superar las barreras cualitativas y cuantitativas.

Cuando hablamos de la integración hombre-máquina, nos referimos a la pérdida de las fronteras taxonómicas entre un artefacto mecánico y un ser humano (Koval, 2011). El hecho de que exista esta posibilidad plantea una oportunidad de superación de las limitaciones de ambos tipos, a través herramientas como las interfaces cerebro-ordenador.

Según da Silva et al. (2011) las interfaces cerebro-ordenador son sistemas que permiten la interacción entre el cerebro humano y un ordenador, donde el análisis de señales del electroencefalograma genera comandos de control. Esto significa la posibilidad de usar sistemas artificiales que potencian y mejoran nuestro rendimiento neurológico.

De momento, estos aparatos se están usando para suplir carencias provocadas por la enfermedad. El binomio inteligencia artificial-medicina ya es una realidad. Gracias a esto, hay personas ciegas que han recuperado la vista, personas sordas que han vuelto a oír; pacientes con depresión que se ha curado a través de dispositivos neurales que regulan los desequilibrios neuroquímicos; y personas que han vuelto a sentir y controlar un miembro perdido gracias a prótesis motoras (Núñez, 2018).

No obstante, sus aplicaciones van más allá de cubrir capacidades perdidas en el campo médico. También se pretende crear una tecnología que mejore las capacidades que tenemos por defecto. Este concepto de denomina *enhancement* y consiste en la mejora de artificial de nuestras capacidades físicas y cognitivas. Claro que, existe un debate sobre que se considera natural y que se considera artificial. La medicina positiva y la psicología de la salud existen desde hace unos cuantos años. ¿Es artificial consultar con un médico para desarrollar hábitos de vida más saludables y aumentar el rendimiento físico? ¿O visitar a un terapeuta para trabajar las habilidades sociales? La verdad es que no está claro. Aquí nos centramos en aplicaciones más obvias, como es el hecho de introducir un dispositivo en el cerebro que te permita acceder a la Wikipedia mientras vas tomando notas a las que posteriormente puedes acceder con tu cuenta de One Drive, por poner un ejemplo.

Esto es lo que implica el concepto de transhumanismo. Crear, por medio de la tecnología, humanos cada vez más capaces una múltiple disparidad de aspectos. Sin embargo, el transhumanismo no es exactamente una solución a las limitaciones cualitativas de la inteligencia artificial, sino el fenómeno de integrar las capacidades humanas a nivel cualitativo con las capacidades artificiales a nivel cuantitativo. Y en ese sentido, para que la integración hombre-máquina pueda llegar a darse, debe de existir una relación recíproca entre el humano y la máquina. Mientras solo haya una transmisión unidireccional, no se está dando una integración real.

El problema de esto es que el abuso de una interfaz ordenador-cerebro puede ser peligrosa. Como explica Wångstedt (2016) mientras que en el caso de que en una trasmisión del cerebro al ordenador hay algún problema, la consecuencia es un cortocircuito en el ordenador; si el fallo se produce en una transmisión del ordenador al cerebro, puede producirse daño cerebral o un coma; lo que lo hace mucho más complejo de implementar.

Entre algunas aplicaciones de esta tecnología, quizás no se tenga por objetivo la mejora artificial de la raza humana, pero definitivamente se salen de las aplicaciones médicas antes mencionadas. Si se posee la capacidad para crear artificialmente la experiencia consciente de una imagen visual, ¿qué nos impide desarrollarla para el entretenimiento del usuario? La realidad simulada es una opción cada vez más atractiva para la industria del videojuego. Y si en lugar de llevar un casco con unas gafas de realidad virtual tienes acceso neurológicamente a un mundo virtual tan realista que puedes moverte, tocar y sentir olvidándote de que no es real, ¿dónde acaban las posibilidades de la creación de aventuras, viajes y experiencias a la hora de diseñar un videojuego?

Hay determinados problemas que pueden ser solucionados con la tecnología adecuada, eso es una obviedad. Sin embargo, la reacción pública a estas soluciones muchas veces es negativa por el desconocimiento de las consecuencias. A los coches autónomos, sin ir más lejos, se les exige en un nivel social y ético una perfección que de ninguna manera exigiríamos si fuera una persona la que conduce el vehículo. Y no es porque sean más peligrosos, sino porque les tenemos miedo. Una cosa que debemos entender es que para que los coches autónomos salgan al mercado no tiene que ser perfectos, solo tienen que ser mejores que los humanos, y a los humanos se nos da bastante mal, si lo miramos en términos de accidentes y número de muertes por año (Grey, 2014). Hay quien optaría por

no dejar que sea una máquina quién decida contra qué o contra quién estrella el coche en caso de accidente. Sin embargo, solo permitir que sea un humano en que pueda tomar decisiones en estos casos no es una solución, es ignorar el problema.

Núñez (2018) afirma que, en el fenómeno de la integración hombre-máquina, somos los humanos los que podemos salir mal parados. A fin de cuentas, los que tenemos el peligro de atrofiar funciones que dejaríamos de usar debido a la constante asistencia de los dispositivos neurales y de volvemos adictos a estos, somos nosotros. Las máquinas no se van a volver adictas a los que les aportemos.

¿Hasta qué punto somos esclavos de nuestras propias limitaciones biológicas o de la tecnología que nosotros mismos creamos?

6. Conclusiones

El futuro es incierto, eso parece que está más claro que nunca. No obstante, por lo que sabemos a día de hoy, el reto que más trabajo y creatividad requerirá será que la inteligencia artificial supere el salto cualitativo que le impide llevar a cabo las tareas relacionadas con determinadas facetas de la flexibilidad. Es decir, superar la carencia de un dispositivo autorreferencial que le permita experimentar subjetivamente el mundo y decidir por sí misma que resultado está bien y que resultado está mal.

Al final, por mucho que los datos que tenemos nos indiquen que la consciencia es clave para superar estas limitaciones, puede que no se trate de algo tan importante como creemos ahora mismo. Puede que el problema en sí no sea tan importante y estemos tratando de responder a algo que en unos años ni siquiera se considere un inconveniente. Quizás la importancia que hoy le damos se disipe con la llegada de nuevos datos y nuevos problemas. Como es frecuente con el progreso, cada vez que se encuentra con una limitación, encuentra una forma de seguir adelante. Si se agotan todas las posibilidades, se redirecciona hacia otras alternativas de avance, del mismo modo que, cuando el mundo subatómico puso freno a la miniaturización de los chips, surgieron los ordenadores cuánticos. En ese sentido, el progreso es como una marea a la que, si pones una barrera, el agua irá por otro lado; si pones muchas, provocas una inundación.

Desde este punto de vista, las maquinas que piensan, al igual que las máquinas que vuelan, no se parecerán en nada a lo que en un principio tenemos en mente. Como señala Núñez (2012) los ingenieros creadores de inteligencia artificial dejaron de consultar a los psicólogos hace décadas. Las máquinas no tienen intención de parecerse a los humanos ni se desarrollan para hacer de nosotros mejores personas.

Lo único que sí está claro es que nuestro futuro como especie está ligado a la inteligencia artificial, con todas sus consecuencias. Quién estará al mando en esa nueva sociedad es otro asunto para el qué solo existe una manera de averiguarlo.

7. Bibliografía

Cabañes, E. (2008). Creadores Artificiales: ¿la creatividad más allá de lo humano?. *Actas del I Congreso de Jóvenes Investigadores en Filosofía*.

Cabañes, E. Creatividad Artificial: Cuestionando los límites humano/artificial. *Actas del XLV Congreso de Filósofos Jóvenes. Intervenciones filosóficas: Filosofía en acción*. Recuperado el 26 de octubre de 2017 en http://euridicecabanes.es.tl/Creatividad-Artificial.htm

Castillo, E., Gutiérrez, J. M., & Hadi, A. S. (2011). Sistemas expertos y modelos de redes probabilísticas. B-Enrique Castillo (Editor).

Fernández-Riverola, F., & Corchado, J. M. (2000). Sistemas híbridos neuro-simbólicos: una revisión. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, *4*(11).

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., & Mueller, E. T. (2013). Watson: beyond Jeopardy!. *Artificial Intelligence*, *199*, 93-105.

Gelepithis, P. A. (1984). On the Foundation of Artificial Intelligence and Human Cognition (Tesis doctoral). Recuperado de Google Scholar.

Grey, C. P. G. (2014). Humans Need Not Apply. *C. P. G. Grey*. Recuperado el 12 de octubre de 2017 en http://www.cgpgrey.com/blog/humans-need-not-apply

Grey, C. P. G. (2017). How Machines Learn. *C. P. G. Grey*. Recuperado el 22 de febrero de 2017 en http://www.cgpgrey.com/blog/how-do-machines-learn

Koval, S. (2011). Convergencias tecnológicas en la era de la integración hombre-máquina. *Razón y Palabra, 16* (75).

Kurzweil, Ray (2012), *How to Create a Mind: The Secret of Human Thought Revealed*. New York: Viking Books.

Lomas, N. (2014). Read the sonnet co-authored by Shakespeare, an MIT PhD student & a machine-learning algorithm. *TechCrunch.* Recuperado en el 20 de febrero de 2018 en https://techcrunch.com/2014/01/26/swift-speare/

Montoya, F. (2004). La Criptografía Cuántica, ¿Realidad O Ficción?. *Instituto de Física Aplicada, Departamento del Tratamiento de la Información y Codificación, Consejo Superior de Investigaciones Científicas*.

Núñez, J. P. (2012). Conocimiento único de naturaleza desconocida. *La mente: la última frontera*, (1ª ed), (pp. 243-307). Madrid: Universidad Pontifica de Comillas (Publicaciones).

Núñez, J. P. (2012). Humanoides. *La mente: la última frontera*, (1ª ed), (pp. 309-349). Madrid: Universidad Pontifica de Comillas (Publicaciones).

Núñez, J. P. (2018) Hombres y máquinas: futuro y límites del transhumanismo. *FronterasCTR*. Recuperado el 28 de febrero de 2018 en https://blogs.comillas.edu/FronterasCTR/2018/03/21/hombres-maquinas-transhumanismo/

Oremus, W. (2014). The first news report on the L.A. earthquake was written by a robot. *Future Tense, the citizen's guide to the future.* Recuperado en el 20 de febrero de 2018 en http://www.slate.com/blogs/future_tense/2014/03/17/quakebot_los_angeles_times_robot_journalist_writes_article_on_la_earthquake.html

Pérez, M. P., & Bautista, A. (2014) Pensamiento probabilístico. En Carretero, M., & Asensio, M. (Eds.), *Psicología del pensamiento*, (pp. 23-35). Madrid: Alianza.

Russell, S. & Norvig, P. (1995). A modern approach. *Artificial Intelligence. Prentice-Hall, Englewood Cliffs*, *25*, 27.

Shawar, B. A., & Atwell, E. (2007). Chatbots: are they really useful?. *In Ldv forum*, 22 (1), (pp. 29-49).

da Silva Sauer, L., Valero Aguayo, L., Velasco Álvarez, F., & Ron Angevin, R. (2011). Variables Psicológicas En El Control De Interfaces Cerebro-Computadora. *Psicothema, 23* (4), (pp. 745-751).

Stevens, M. (2017). Artificial Intelligence. *Mind Field*. Recuperado el 2 de abril de 2018 en https://www.youtube.com/watch?v=qZXpgf8N6hs

Velasco, J. A. M. Inteligencia Artificial y Conciencia. *Departamento de Matemáticas de la UAH.* Recuperado en el 26 de octubre de 2017 en http://www2.uah.es/benito_fraile/ponencias/inteligencia-artificial.pdf

Wångstedt, D. (2016). Artificial Intelligence. *LEMMiNO.* Recuperado el 12 de octubre de 2017 en https://www.lemmi.no/post/artificial-intelligence

Wångstedt, D. (2016). Simulated Reality. *LEMMiNO.* Recuperado el 3 de abril de 2018 en https://www.lemmi.no/post/simulated-reality?rq=simulated%20rea