La noción de inteligencia después de *Computing Machinery and Intelligence*

UNA PERSPECTIVA HISTÓRICA

Pedro Montoto García (USC)
Enrique Alonso González (UAM)

20 de septiembre de 2015

"*Intelligence is* what is measured by intelligence tests."

E. Boring, circa 1920

Resumen

Este trabajo pretende estudiar la evolución del concepto de inteligencia en relación a la Inteligencia Artificial a partir de la publicación de *Computing Machinery and Intelligence* (Turing 1950). Incluímos un comentario crítico de éste artículo y una compilación y resumen de los problemas que genera la pregunta *¿Puede pensar una máquina?*, ejemplificados por las críticas academícas a dicho artículo. Se hace un compendio también de los tipos de soluciones que se dan, técnicos y matemáticos, y de los nuevos problemas y conclusiones filosóficas a los que nos lleva ésta. Como guía organizativa se ha usado *The Quest for Artificial Intelligence: A history of ideas and achievements* (Nilsson 2009)[1]. Como introducción hemos aprovechado para hacer un repaso de los hitos históricos que conducen a la fundación de la Inteligencia Artificial. Además, se ligan los desarrollos tecnológicos de la revolución de los computadores durante la segunda mitad del siglo XX a la disciplina de la IA.

Palabras clave: Historia de la Inteligencia Artificial, Filosofía de la Inteligencia Artificial, Alan Turing, Ciencias Cognitivas, Cibernética

Abstract

This work intends to study the evolution of the concept of intelligence in relation to Artificial Intelligence after *Computing Machinery and Intelligence* (Turing 1950). We include a critical commentary of Turing's article and a summarized compilation of the problems that the *Can a machine think?* question generates exemplified by academical critique to that same article. A compendium of types of technical and mathematical solutions and new problems and philosophical conlclusions that this question brings is also included. *The Quest for Artificial Intelligence: A history of ideas and achievements* (Nilsson 2009) is used as an organizative guide. As an introduction we also take a look to the historical landmarks that brought Artificial Intelligence as an academical discipline into being. The tecnological development during the computer revolution in the second half of the twentieth century is linked with this discipline as well.

Keywords: History of Artificial Intelligence, Philosophy of Artificial Intelligence, Alan Turing, Cognitive Sciences, Cybernetics

[1]Nils Nilsson es investigador en el *Department of Computer Science* de la Universidad de Stanford, inventor de varios algoritmos de gran importancia en redes neuronales y co-creador del algoritmo A^*, que permite descubrir el camino más corto entre dos nodos de un grafo mejorando la búsqueda exhaustiva ponderada (conocida como *algoritmo de Dijkstra*) mediante el uso de métricas heurísticas.

Índice general

Capítulo **1**

Introducción Histórica a lo Inteligente y lo Artificial

> "*Intelligence is* a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience."
>
> Common statement with 52 expert signatories, 1997

Todo lo que llamamos IA comienza con una pregunta de apariencia simple, *¿Es posible construír algo que pueda pensar?*, que en realidad presenta muchas cuestiones asociadas. Podemos citar otras ideas análogas en cierta medida, como crear vida o crear una máquina que resuelva todos los problemas matemáticos. Esta idea no es en absoluto reciente ni mucho menos. La idea de inteligencia artificial ha existido en diversas formas durante la historia del pensamiento occidental, al menos desde la grecia clásica, en mitos, leyendas, historias, especulación y autómatas mecánicos y que, a favor o en contra, intentan dar una respuesta final a esta especie de sueño colectivo.

Las primeras leyendas griegas relevantes de las que tenemos noticia, alrededor del siglo V a.C., de las estatuas de Pigmalión traídas a la vida por Afrodita, diosa de la vida y del amor, y de Hefesto, dios de la forja, construyendo ayudantes dorados para los dioses. De todo esto nos llega noticia a través de la *Política* de Aristóteles, que crea probablemente por accidente uno de los primeros ejemplos de ciencia ficción político-social, planteando la cuestión de qué ocurriría si tuviésemos *máquinas/autómatas/seres artificiales* inteligentes:

> Pues si cada uno de los instrumentos pudiera cumplir por sí mismo su cometido obedeciendo órdenes o anticipándose a ellas, si, como cuentan de las estatuas de Dédalo o de los trípodes de Hefesto, de los que dice el poeta que entraban por sí solos en la asamblea de los dioses las lanzaderas tejieran solas y los plectros tocaran la cítara, los constructores no necesitarían ayudantes ni los amos esclavos. (Aristóteles 1988, Aristot. Pol. 1.1253b)

Más ejemplos se dan posteriormente. El Talmud, compilado entre los siglos I y VI, habla de *golems* creados de tierra que hombres santos y doctos pueden infundir de vida.

En el siglo XII el catalán Ramon Llull inventa lo que el llama *Ars Magna*, un conjunto de discos de papel que convenientemente combinados y rotados permitirían dirimir cualquier discusión teológica mediante el uso de la lógica, con la intención de convertir a las personas de fe musulmana al cristianismo mediante la misma. En algún momento entre finales del siglo XV ó principios del XVI Leonardo da Vinci crea unos esquemas para un robot-caballero que sería supuestamente capaz de sentarse, levantarse y mover los brazos manipulado, eso sí, por un humano[1].

En el siglo XVII se desarrolla, gracias al auge del racionalismo y las ideas humanistas, la idea de que todo en el universo físico puede ser mecanizable, incluídos los seres vivos. Se puede decir incluso que hasta ésta época, desde la Grecia clásica, vida e inteligencia sólo podía ser algo otorgado por dioses u otros seres omnipotentes más allá del universo físico, al darse que cualquier creación humana no puede superar una imitación de la vida que la divinidad otorga. Hobbes, en este mismo siglo en su *Leviathan*, contempla la posibilidad de crear un ingenio mecánico que se comporte como un animal, pues todos los órganos para él tienen paralelismos con la mecánica: "el corazón no es más que un muelle, los nervios son cuerdas" y conceptos similares aparecen a lo largo de su obra. Descartes, por el contrario, creía que las máquinas serían incapaces de pensamiento real pues sólo están formadas por materia y sería imposible dotarlas de mente. Algunos pensadores creen ver en Leibniz un avance de la Inteligencia Artificial ya que, al igual que Hobbes, concebía que todo lo que hace la mente son computaciones, en *De arte combinatoria*.

En el siglo XVIII Jacques de Vaucanson presenta un autómata que es capaz de simular un pato vivo en ciertos de sus aspectos, construído mediante un armazón metálico adornado con plumas de pato y componentes de relojería y mecánica, que era capaz de comer grano, beber y "digerir". En realidad el producto de la digestión ya estaba en el interior del pato para la simulación. Jacques de Vaucanson también es el inventor de las primeras tarjetas perforadas entendidas como contenedoras de programas, secuencias de acciones, para entes mecánicos. Éstas tarjetas se usarán en la "programación" de telares mecánicos en el siglo XIX y a principios del siglo XX podremos ver ya máquinas calculadoras que permitían hacer operaciones aritméticas usando estas tarjetas como entrada y salida de información.

También en el siglo XVIII se creó el "Turco mecánico", un autómata que podía jugar al ajedrez como un maestro y que como exhibición del mismo fue enviado de gira por las cortes de Europa de la época retando y ganando a soberanos y estrategas. Más tarde se supo que éste "autómata" debía su genialidad a un maestro de ajedrez humano que

[1] Éste robot se presentó supuestamente funcionando en una fiesta de la época en la corte de Venecia en 1495 organizada por Ludovico Sforza, y más recientemente un empresario llamado Mark Elling Rosheim reconstruyó los diseños de Leonardo, probando que los mismos eran sensatos y funcionales.

se ponía en su interior en cada partida. La compañía Amazon crea en 2013 un servicio que distribuye y automatiza tareas simples realizables por humanos a voluntarios que cobran por tarea realizada. El nombre del mismo, en homenaje, es Amazon Mechanical Turk y nos permite resaltar que el uso de la inteligencia humana en tareas simples aún no automatizables es, por tanto, una solución bien conocida para las lagunas que la Inteligencia Artificial aún tiene.

A partir del siglo XIX comenzamos a ver por doquier obras de teatro, narraciones y películas que hablan del *qué ocurriría* si las máquinas pudiesen pensar o actuar como humanos[2]. Es evidente, por tanto. que muchos de los problemas de este *posible* que plantea el nacimiento de la IA no son nuevos y sin embargo, los mayores avances que se han dado en este campo han sido sin duda alguna durante el siglo XX en los campos de la maquinaria computacional, la psicología, la neurología y la vertiente artística en la ciencia-ficción.

En última instancia los objetivos de la Inteligencia Artificial son múltiples y variados. Damos una lista de las diversas técnicas que se han creado y describiremos después, compilando y resumiendo desde McCarthy 2007, ya que sus categorías se adecúan a los experimentos concretos de cada campo pero no generalizan a través de ellos:

búsqueda en grafos Englobamos aquí todos los sistemas cuyo espacio de búsqueda es recursivamente enumerable o generable mediante una función recursiva, aunque también tiene aplicación en subconjuntos recursivamente enumerables de conjuntos no-enumerables. Entran por tanto en esta categoría la resolución de juegos finitos como el ajedrez o la elaboración de planes de acuerdo a restricciones. Razonamiento no monotónico, razonamiento difuso, bots de chat suelen ser representados como grafos difusos (redes de Markov etc.). Podemos incluír también la generación de espacios de soluciones con selección difusa como los algoritmos genéticos, ya que generan un grafo dirigido de orden en el espacio de búsqueda.

representación del conocimiento Todo problema que plantee resolución lógica determinista o no-determinista requiere la representación adecuada del universo del discurso. Hablamos de ontologías y epistemología computacional, e interpretación de lenguajes naturales. Las heurísticas, que son conocimientos a posteriori adquiridos de la experiencia que pueden ser dados al programa o generados por el programa también son representación del conocimiento.

[2]Como veremos, del pensar al actuar como humanos hay diferencias que además, se han ido acentuando con el progreso en IA.

pattern matching Entran en esta categoría las técnicas que permiten resolver problemas de asociación de elementos de un conjunto generalmente grande (imágenes, tomas de muestras de ADN,...) a clases difusas, normalmente mediante redes de neuronas o lógica difusa. La identificación de objetos en imágenes, la clasificación de conjuntos de datos (cuando los puntos de datos son independientes) o segmentos de datos (cuando existe una dependencia entre puntos de datos, como las series temporales) en clases también encaja en esta clase.

Las principales preguntas que derivamos de *¿Es posible construír algo que pueda pensar?* son dos: *¿Qué es pensar?* y *¿Qué es construír?*. Trataremos de ofrecer una visión histórica y panorámica de ambas cuestiones.

Capítulo **2**

¿Qué entendemos por construír? La base física de la IA

> "*Intelligence is* the ability to use optimally limited resources – including time – to achieve goals."
>
> Ray Kurzweil, 2000

Para la construcción de inteligencia precisamos, por tanto, definir qué entendemos por construcción. Esto presenta un número de problemas, como veremos, que no entran por completo en el alcance de este trabajo, pero creemos que conviene reflejar el cambio de concepto que se ha dado en el último siglo, aunque los grupos de tecnólogos y científicos (ingenieros, matemáticos, biólogos, psicólogos y científicos sociales) que trabajan en Inteligencia Artificial no suelan tenerlo en cuenta.

Normalmente definimos constructo o artefacto[1] como una entidad en la que algunas o todas sus propiedades preexisten en la intencionalidad de un autor. No definiremos entidad porque es un problema, el del ser, que no corresponde tratar en este trabajo. Es un artefacto si éstas propiedades existen en la descripción de la intencionalidad del autor y/o son aceptadas como válidas en la descripción de lo construído por el autor, i.e. el autor determina unas propiedades y valida que el constructo para el que éste ha guiado o accionado el proceso son las que se espera del mismo. En éste sentido la IA es una disciplina científica cuya *tecnología* asociada genera comportamientos inteligentes como artefacto, en palabras de Miguel Ángel Quintanilla citado en (Echeverría 1998). Por tanto, entendemos también que *tecnología* es la especificación de un proceso con bases científicas para la obtención de ciertos resultados, lo que implica intencionalidad.

Ya desde los tiempos de Aristóteles, en su *Física* se intuye esta definición en la diferencia entre los "productos naturales" que se generan por sus propios impulsos internos mientras que los "productos artificiales" precisan de una intencionalidad humana. Avicena criticaba en la edad media que la alquimia jamás podría conseguir "sustancias genuínas" como las presentes en la naturaleza precisamente por ser un constructo con intencionalidad humana. En este caso, para mantener el debate cerrado

[1] Para una referencia básica en la filosofía de la tecnología es recomendable consultar Franssen, Lokhorst y Poel 2013 y Hilpinen 2011.

en torno a lo que nos proponemos definir, i.e. lo que es un constructo, asumiremos que no existe autor en los entes naturales y que son autogenerados en cierta medida por sí mismos o sus predecesores en el tiempo.

Podemos criticar que la "vida natural" no se diferenciaría en absoluto de una supuesta "vida artificial" pues la intencionalidad del autor al crear dicha vida natural, esto es, las propiedades de dicha forma de vida, debería poseer características en común con la vida "artificial", habiendo sólamente cambios en el proceso y los materiales fuente si es una imitación perfecta.

Los avances más recientes en biología, química e Inteligencia Artificial y las promesas de avances futuros nos llevan en última instancia que lo "artificial" podría diferir sólamente de lo "natural" en una cuestión de proceso, lo que podría convertir la dicotomía "natural"-"artificial" en vacua: Si no podemos distinguir natural de artificial sin conocer el proceso que ha generado la entidad en cuestión y las entidades "naturales" o "artificiales" son indistinguibles, conocer el proceso después de conocer la entidad impide distinguir qué proceso es "natural" o "artificial" a menos que los etiquetemos arbitrariamente. La razón es que las cualidades del objeto que lo hacen "natural" o "artificial" no son inherentes al objeto una vez la técnica de imitación ha avanzado lo suficiente, como creemos que es la dirección que se está tomando actualmente. Este hecho será de especial relevancia cuando comentemos el juego de la imitación de Turing.

En el contexto de Inteligencia Artificial esto nos lleva, más adelante, al si existe distinción ontológica entre la inteligencia creada artificialmente y la inteligencia natural. Es decir, si logramos crear inteligencia, sea ésta lo que sea, ¿existe aún alguna diferencia entre lo que esta inteligencia es y la inteligencia que nosotros pretendemos? Por ello entramos en una distinción similar a la que propone John Searle (como veremos posteriormente) entre "IA débil", significando "la IA sólo imita la acción de una mente real", e "IA fuerte", significando "la IA *es* una mente real" (Searle 1980). En el argumento que comentamos más arriba las consecuencias son similares, aunque la distinción se haga irrelevante. No debe entenderse este argumento como crítica a Searle pues en el argumento de Searle, que habla sobre el proceso inteligente como constructo versus el proceso inteligente como ente natural, ambos constructos son distinguibles por sus propiedades y mateiales internos, aunque no los observables externamente, por lo que sí existe una posible distinción razonable.

Conviene resaltar que todas las referencias que se han podido compilar para este trabajo no adoptan una postura respecto a las bases filosóficas de la distinción natural-artificial o de la teleología de los componentes de la inteligencia. La mayoría de la literatura se limita a presentar diferentes modelos matemáticos que suelen tomar como base la máquina de Turing, las redes neuronales artificiales, aunque otros llegan a arguír

que es necesaria una imitación matemática de los procesos biológicos involucrados en las emociones, normalmente se trata de una lista de tipos de procesos matemáticos, es decir, herramientas para la imitación de la inteligencia según la definición de ésta que se dé.

La convención moderna del constructo computacional, que se instancia físicamente en un soporte tecnológico, y sirve para el soporte físico de la Inteligencia Artificial es la Turing-completitud. No nos detendremos aquí a dar una descripción completa de lo que es la máquina de Turing (o el equivalente cálculo lambda de Church, u otros), sirva decir que es el constructo conceptual en el que se basa la operación de la maquinaria de computación general[2]. En su contexto histórico, 1928, tres años antes de la primera publicación de Gödel sobre el segundo problema de Hilbert y, como continuación del programa formalista, Hilbert elabora una nueva lista de problemas para proponer a la comunidad matemática. Entscheidungsproblem, como se conoce, por su nombre alemán y que en castellano sería "problema de la decisión", consiste en dar un procedimiento finito para cualquier fórmula de lógica de primer orden, y por extensión de la aritmética, que nos diga si ésta es verdadera o falsa. Esta idea es heredera directa del décimo problema de Hilbert postulado en 1900, que consistía en similarmente dar un método de pasos finitos para resolver cualquier ecuación diofántica y en una perspectiva histórica de la filosofía leibniziana.

En 1936, cinco años después de On formally undecidable propositions se da un nuevo golpe a este formato de modernismo llamado formalismo o positivismo. Alan Turing publica *On computable numbers, with an application to the Entscheidungsproblem*; en el que formaliza una idea intuitiva de máquina capaz de realizar cálculos [3], demostrando que el Entscheidungsproblem es irresoluble mediante estas máquinas; y Alonzo Church publica A note on the Entscheidungsproblem; en el que demuestra que el Entscheidungsproblem es también irresoluble mediante las funciones lambda-definibles (o cálculo lambda) que había estado desarrollando con su estudiante Stephen Kleene. Se considera hoy, dado que Kleene demostró que estas clases de funciones son equivalentes; i.e. toda función lambda-definible puede transformarse mediante un procedimiento finito en una función Turing computable equivalente y viceversa; y que no se ha podido describir ningún algoritmo que no sea describible en estos términos, que esta es la noción formal de computabilidad mediante un procedimiento finito más aproximada a la idea intuitiva de computabilidad.

La máquina de Turing intuitiva, desprovista del aparataje formal matemático, consis-

[2]Sobre la teoría matemática que soporta esta máquina conceptual puede leerse en Hopcroft, Motwani y Ullman 2006 ó bien en el original Turing 1937.

[3]Que Turing usará en sus implementaciones de las *Bombes*, unas de las primeras máquinas de cálculo eléctricas que fueron usadas para acelerar el descifrado de mensajes alemanes en la Segunda Guerra Mundial

te en una cinta en la que se pueden escribir y sobreescribir símbolos, teniendo capacidad infinita numerable para símbolos y siendo el símbolo la unidad de lectura y escritura, y un dispositivo de máquina de estados que define una función parcial de aplicación continua hasta que se alcanza un estado de parada. Esta función nos dice que si la máquina se encuentra en un estado x_t y lee en la cinta un símbolo s_i: a) escribirá en esa misma posición un símbolo s_o (que puede ser igual a s_i), b) se moverá una posición adelante o atrás, y c) se moverá al estado x_{t+1}, o se quedará en x_t según el "programa" descrito por la función. La definición de función recursiva toma pues, la forma de esta máquina de estados con memoria sobreescribible.

Pasamos ahora a dar una descripción de los soportes tecnológicos que se han tenido en cuenta para la Inteligencia Artificial, lo que en la jerga del ingeniero en computación sería la infraestructura, y que se compone normalmente de cuatro elementos que establecen abstracciones en el elemento anterior dando lugar a la posibilidad de construír elementos más complejos:

- En el núcleo del artefacto tenemos el soporte físico de cómputo: hablamos generalmente de una máquina de cómputo de propósito general con procesador aritmético-lógico y memoria, que contendrá datos y programa de acuerdo a la arquitectura von Neumann, o más recientemente, de varias de éstas máquinas conectadas mediante una red de datos (llamado procesamiento distribuído) o funcionando sobre una memoria compartida (llamado procesamiento paralelo).

- Éste elemento se suele cargar con un sistema operativo que establece una capa de abstracción sobre los elementos físicos para que los programadores puedan usar todos los componentes de forma simple y eficiente, ya que se encarga también de otorgar el control de recursos físicos a programas.

- Sobre los sistemas operativos se opera mediante lenguajes de programación, que consisten en abstracciones sobre los sistemas del mismo que alcanzan al sistema físico de cómputo y que al mismo tiempo ofrecen una manera de expresar algoritmos. Cada lenguaje ofrece diferentes estructuras computacionales que los hacen más adecuados para expresar un tipo de algoritmo u otro, aunque todos suelen tener la misma capacidad teórica, esto es, la Turing-completitud.

- En última instancia tenemos los algoritmos y estructuras de datos que nos permiten resolver el problema que estemos tratando. Ésta es la parte esencial del sistema que identificamos como Inteligencia Artificial, ya que todos los sistemas actuales poseen pequeñas variaciones de los sistemas precedentes siendo los cambios en algoritmia los más relevantes hoy en día.

Cada una de estas partes puede considerarse como un problema de ingeniería en sí mismo, lo que ayuda a la simplificación de su resolución en conjunto, y a la división en tareas de los grupos de investigación, puesto que en las últimas décadas se ha pasado de la investigación general en computación a la especialización cada vez más granular en cada uno de estos campos. Por supuesto para llegar a estos elementos desde la concepción simple de las bombes, que eran computadores con un propósito definido y limitado, hubo una evolución de las máquinas de cómputo, que relatamos en el apéndice A.

Hoy en día se están empezando a estudiar tecnologías prometedoras como la expansión de la tecnología electrónica mediante elementos ópticos, bajo la promesa de más velocidad de cómputo al ser las interacciones entre fotones más rápidas que las interacciones entre electrones, en la computación cuántica, que utiliza los fenómenos a nivel de partícula atómica para implementar puertas lógicas en una variante de la lógica multivalorada denominada lógica cuántica que permitiría la paralelización masiva de las operaciones atómicas, la computación basada en ADN y biología molecular, que basa su implementación de puertas lógicas en interacciones químicas de proteínas y ácidos o intercambios químicos entre células, como en las sinapsis neuronales. Todo ello nos conduce a pensar que estas tecnologías tienen una relación inherente con la base física o biológica de los seres vivos, y aunque no resulten productivas para los problemas de computación actuales [4] podrían darnos esa capacidad de imitación a todos los niveles que mencionamos al principio del capítulo. En todo caso, rechazamos la hipótesis de Penrose en la que la imitación tenga que llegar hasta los fenómenos cuánticos que se dan en las neuronas, y asumimos que las neuronas pueden ser ajustadas mediante funciones continuas adaptadas a universos discretos (i.e. computables).

[4]Por ejemplo, la computación basada en elementos biológicos resulta más lenta que su equivalente electrónico pero permite el procesado paralelo en una escala mucho mayor, lo que puede ser también una propiedad de la computación cuántica según la implementación que siga

Capítulo **3**

La definición de inteligencia

> "[*Intelligence is*] Achieving complex goals in complex environments"
>
> B. Goertzel, 2006

La definición de lo que es inteligencia es naturalmente un apartado importante, aunque de difícil solución en el campo de la Filosofía de la Inteligencia Artificial. Se puede ver ver un pequeño compendio de definiciones históricas de inteligencia en diferentes ramas de las ciencias cognitivas en Legg y Hutter 2007. En primer lugar, trataremos de diferenciar la "inteligencia" en sí que nos parece de un platonismo peligroso al declarar que ésta existe como entidad al no ser designable lingüísticamente en el universo físico, y reemplazarla por "comportamientos inteligentes" o "tareas para las que es necesaria inteligencia". "Inteligencia", entendemos pues, no es predicable directamente de agentes, sino de actuaciones que califican a dicho agente como inteligente. Se predica normalmente que ser inteligente es un atributo humano, aunque si tomamos ciertas definiciones de inteligencia como válidas podemos atribuír inteligencia a todos los seres vivos o incluso a seres carentes de vida y capacidad de acción como piedras[1], que, intuitivamente no parecen tener ninguno de los atributos necesarios para la inteligencia. Es por tanto un tema que merece ser tratado con cuidado.

Entendemos por tanto que hemos de dar una definición de mente o consciencia desde la que trabajaremos, el punto de vista de éste trabajo. Decimos que la mente al igual que la consciencia es una propiedad emergente de los sistemas físicos que se manifiesta en los comportamientos inteligentes, pero que carece de sentido separada de éstos. Es decir, los comportamientos inteligentes son la base de lo que podemos llamar inteligencia, mente o consciencia, pero no existen comportamientos inteligentes no físicos, de la misma forma que un algoritmo anotado en un papel o cargado en una tarjeta de memoria electrónica no puede generar comportamientos inteligentes a no ser que haya una entidad actuando sobre el mismo.

[1] En el caso que nos ocupa podemos decir que si aceptamos "*Inteligencia es* la capacidad de mantener una existencia continuada en el tiempo", ésta definición no especifica que para tener inteligencia ha de tenerse capacidad de actuación. Test como el de Gunderson, que veremos más adelante, entran en este caso también.

No podemos decir que un cerebro de un animal muerto tenga mente, sin embargo, el cerebro que genera comportamiento en el animal sí podría tenerla. En última instancia, consideramos que mente o consciencia es una abreviación para referirnos a éstos comportamientos inteligentes. Lo que llamamos intencionalidad de los comportamientos proviene de la relación del sistema con el entorno, es decir, la inteligencia sin entorno no existe. Aceptamos la crítica de Lakoff a la congnición incorpórea y aceptamos que ésta es imposible, pero no creemos que haya suficientes diferencias relevantes entre un programa dotado de mecanismos de evaluación de su entorno y una persona cuyos mecanismos de percepción están incorporeizados, afectando a sus procesos de cognición y viceversa (Lakoff 1987).

Entendemos que para que un comportamiento pueda considerarse inteligente debe existir posibilidad de error, entendido como una desviación entre la intencionalidad a priori y el resultado a posteriori, y el agente que realiza dicho comportamiento debe ser capaz de detectar esta divergencia en algún momento y corregirla. Para ello el agente debe ser capaz de continuarse autónomamente durante un tiempo suficiente o crear sucesores de sí mismo.

Comencemos por el que creemos soporte físico de los comportamientos inteligentes en animales y humanos y que da partida probablemente a nuestra pregunta inicial, el cerebro. Se sabe que las señales del sistema nervioso son de naturaleza eléctrica desde los experimentos de Luigi Galvani en el siglo XVIII. La disciplina que estudia la "maquinaria" del cerebro, la neurociencia, fue inaugurada en a principios del siglo XX por Camilo Golgi y Santiago Ramón y Cajal. El primero había inventado un método de contraste que permitía distinguir la estructura del tejido cerebral y sus ramificaciones, mientras que el segundo, continuando el uso de éste método creó la teoría de que el tejido cerebral no se componía de una malla de hilos neuronales, sino que dicho tejido estaba formado por células contiguas pero separadas entre sí.

Las experiencias clínicas de Carl Wernicke y Paul Broca con pacientes con partes del cerebro dañadas, pero que eran capaces de realizar la mayoría de sus funciones vitales, a finales del siglo XIX habían concluído en la teoría de que existían diferentes zonas en el cerebro con diferentes responsabilidades en los procesos de pensamiento y fisiológicos y, por tanto, al ser el daño cerebral localizado sólo estarían afectadas las funciones correspondientes a las zonas dañadas. En 1952 se teoriza el primer modelo de propagación de señales entre neuronas, y se concluye que las neuronas reciben un número de señales de otras neuronas y se activan totalmente o inhiben totalmente las señales recibidas de acuerdo a un umbral interno (1 y 0, de nuevo).

Tenemos por tanto una arquitectura general, los elementos mínimos del sistema estudiados en detalle y además una relación clara con la disciplina de la computación y

la comunicación[2], por lo que dada la interpretación mecanicista habitual en la época bastaba estudiar sus relaciones y establecer un modelo matemático para obtener un sistema que nos permitiese conocer y predecir su comportamiento o simulase el objeto de estudio.

Tal y como se dice en Ashby 1956: "La cibernética es a la máquina real (electrónica, mecánica, neural o económica) lo que la geometría es a los objetos materiales de nuestro espacio terrestre", en el sentido de que es una descripción simplificada, y certifica más adelante con "La cibernética es entonces indiferente al reproche de que algunas de las máquinas que estudia no están incluídas en las que encontramos entre nosotros". La palabra *Cibernética* se refiere aquí una disciplina generalista creada en la década de 1940 y que dio origen, entre otras cosas, a lo que hoy referimos como Inteligencia Artificial en la "Conferencia Dartmouth" de 1956, instituyéndose como una de las ciencias cognitivas como veremos. Descendiente éste término de la misma palabra que dio origen a "gobierno" comprendía una descripción general de todos los sistemas complejos, incluídas la vida, las máquinas, la mente humana, la economía y varias disciplinas matemáticas aplicadas intentando encontrar un lenguaje que permitiese expresar las interacciones en todos ellos, es decir "la ciencia del control" de todos éstos sistemas. Hoy en día "cibernética" se usa en lengauje académico como término para agrupar diferentes campos de éstas ciencias además de en la política, la sociología y los estudios organizativos y de empresa. Parafraseando a Pylyshyn 1970 en su comentario sobre algoritmia, la cibernética es más una cosmovisión que un campo de la ciencia aislado.

El campo de trabajo de Ashby era, en términos generales, la aplicación de los principios de la ciencia de la computación a la biología y especialmente a la neurociencia. Dejando a un lado que Ashby asume que el comportamiento inteligente es claramente mecánico, entendemos que la Inteligencia Artificial no escapa a la creación de comportamientos inteligentes fuera del ámbito humano pues, en primer lugar, tal vez no interesa la reproducción exacta y completa de todos los comportamientos a los que atribuiríamos inteligencia. Se entiende todo sistema complejo, desde la maquinaria formada por múltiples máquinas simples hasta las redes sociales pasando por los sistemas neuronales como una máquina compleja cuyas leyes son: finitas y cogentes; en el sentido de que unas no contradicen a otras en el mismo sistema, es decir, que el universo descrito no entra en contradicción interna; imitables por otros sistemas de la misma complejidad y sólo dependientes de lo observable externamente.

Es decir, en el caso de Ashby aún no se había dado el giro cognitivista respecto a los

[2]La publicación en 1949 de *A mathematical theory of Communication* por Warren Weaver, fundador de la cibernética, y Claude Shannon, que inaugura la cuantificación de la información contenida en un mensaje, es clave para entender esta relación.

fenómenos mentales, y Ashby trata la mente como un constructo mecánico cuyas piezas son irrelevantes y de las cuales el funcionamiento observable es lo único relevante. Un funcionalismo que será, 50 años más tarde, destituído en las ciencias cognitivas por dos elementos: en la inteligencia artificial por las técnicas aplicadas a problemas reales, ya que la investigación en la explicación de la mente y el conocimiento será otorgada a los psicólogos, y en la psicología por el cognitivisimo, una corriente que identifica como clave en el comportamiento inteligente a los procesos mentales no observables directamente.

Veremos a continuación un pequeño resumen de la segunda mitad del siglo XX que explica la convergencia de diversos grupos científicos de lingüistas, psicólogos, neurocientíficos, economistas, filósofos y computólogos en lo que hoy se conoce como *ciencias cognitivas*.

3.1. El giro cognitivista

Definimos *ciencia cognitiva* como toda ciencia que se ocupa de uno o más aspectos de los fenómenos de la cognición. Como tales, las ciencias cognitivas forman un conjunto de ciencias y al mismo tiempo un campo de investigación interdisciplinar cuyo tema central es el estudio de la cognición humana, animal y mecánica (Martínez-Freire 2005, p.20), que es a la vez la fuente, se cree, de los comportamientos inteligentes. Se distinguen dos conceptos de cognición: cognición A, aquella que se refiere a la acción de tomar en cuenta una realidad o, dicho de un modo más apropiado a la terminología científica, como recepción de información y cognición B, como el uso y manejo de dicha información. Tanto en el primer y segundo caso se hacen asunciones sobre la capacidad del ser humano de recibir o hacer uso de cualquier información, así como la existencia en cierto grado de dicha información. Existen múltiples concepciones de la información que sirven de base a conocimientos tan diversos como la medición cuantitativa del contenido de un texto fuente (p.e. la concepción informativa de Shannon como desorden, que ya hemos visto en el capítulo anterior) o la fundación de la capacidad deductiva de la lógica clásica (p.e. la concepción objetivista informativa de Corcoran).

Tal y como lo expresan Allen Newell y Herbert Simon en *Human Problem Solving* (1972) el ser humano y su capacidad intelectual, y al mismo tiempo las máquinas formales de cómputo (i.e. las máquinas Turing-equivalentes) pertenecen al género, por analogía biológica, 'sistema de procesamiento de información'. La constitución de las ciencias cognitivas como grupo de interés en las funciones cognitivas de animales, humanos y máquinas, comienza en la llamada "conferencia Dartmouth", una conferencia de matemáticos y lógicos que pone de manifiesto el auge de una disciplina informática

llamada inteligencia artificial, cuyo objetivo es la obtención de comportamiento que denominaríamos inteligente en sistemas artificiales: a saber, máquinas sentientes, capaces de actuación o razonamiento autónomo. Más allá de la cognición en animales y máquinas, se comienza a sospechar que los reinos de las plantas, los hongos e incluso los virus tienen algún tipo de cognición. Por ejemplo, se ha demostrado que las plantas segregan neurotransmisores cuando se identifica una situación de estrés (i.e. aumento en la acidez del suelo o carencia de agua) y mediante esos neurotransmisores se regula su propio crecimiento [3].

Otro de los momentos importantes para este grupo de interés es el nacimiento de la psicología cognitiva, manifestado en tres hechos: la fundación en 1960 del Harvard Center for Cognitive Studies, la publicación en ese mismo año de *Plans and the Structure of Behaviour* de George A. Miller y la publicación en 1967 del primer libro de texto de esta escuela de pensamiento psicológico. Con la psicología cognitiva se recupera el mentalismo, esto es, la existencia de una vida mental interna que es al menos parcialmente ajena a la materia y se deja de lado el conductismo, y por tanto el funcionalismo, como mencionamos antes. Además, el postulado de los procesos mentales que es inherente a la psicología cognitiva, asume un cierto compromiso con el computacionalismo, la existencia de procesos que podrían ser simulados o reproducidos mediante máquinas, y acerca esta escuela a los propósitos de la inteligencia artificial.

Aunque inteligencia artificial y psicología cognitiva son nucleares a las ciencias cognitivas no son, evidentemente, las únicas disciplinas que se centran en el estudio de los mecanismos de consciencia e inteligencia que se dan en animales y máquinas. Ciertas disciplinas, como la sociología cognitiva, la pedagogía y la filosofía de la mente, entre otras, trabajan también con este rumbo aportando diferentes perspectivas y técnicas de trabajo. Se pone el énfasis en que las ciencias cognitivas asumen una naturalización materialista de sus postulados, i.e. no existen entidades fuera del mundo físico, por tanto la mente, el espíritu y cualquier entidad a la que pueda ser asignada la función creadora de inteligencia debe tener una manifestación física y ningún tipo de atributo no-físico. Por lo tanto, todas las disciplinas que toman parte en las construcciones de las ciencias cognitivas si no asumen esta doctrina metafísica, por así llamarla, deben compatibilizarla con ella[4].

[3] `http://www.sciencedaily.com/releases/2015/07/150729085922.htm`

[4] Es éste el caso de la psicología, cuyo estudio de los procesos mentales obvia en muchos casos los procesos fisiológicos que los desencadenan.

3.2. Computing Machinery and Intelligence

El *paper* fundamental que da título a este trabajo e inaugura formalmente la investigación hacia comportamientos inteligentes y vida artificial por parte de los computólogos e ingenieros en computación es *Computing Machinery and Intelligence* de Alan Turing, publicado en 1950. Es éste el primer momento en que puede verse una definición, aunque no formal, sí observable y reproducible de lo que es un comportamiento inteligente. A medida que la Inteligencia Artificial ha ido enfocándose en aplicaciones concretas como la creación de algoritmos para diferentes análisis de datos, el *paper* de Turing ha perdido relevancia en éste campo, mientras que en disciplinas en las que el sujeto de la discusión es propiamente la mente humana como la psicología y la filosofía, éste sigue estando en el punto de mira. Si bien el paper ha perdido relevancia en las ciencias computacionales debido a su enfoque en la resolución de problemas concretos no deja de ser sorprendente lo precisas que son las intuiciones de arquitectura y aumento en capacidad computacional que predice Turing.

En resumidas cuentas, Turing describe la máquina formal de una manera breve e intuitiva y enuncia las reglas del juego que permitiría descubrir si una máquina de éste tipo cargada con cierto programa está pensando o no. Éste juego debe prepararse de tal manera de que una mujer, un hombre y un interrogador que puede hacer preguntas a ambos. Es conveniente que la sala de testeo separe a los sujetos par que no puedan verse entre ellos pero haya canales de comunicación claros, preferentemente mediante texto, para que sea imposible inferir la cualidad humana o no del estilo comunicativo o la voz de cada sujeto.

Turing plantea la ejecución de tal forma que el interrogador debe adivinar cuál de los dos sujetos es el hombre y cual la mujer, siendo la tarea del hombre convencer al interrogador de que es la mujer y la de la mujer ayudarle a hacer la identificación correcta[5]. Seguidamente pasa a considerar qué ocurriría si sustituyésemos al hombre por una máquina capaz de realizar conversaciones imitando a humanos. El valor del test en sí, por tanto, no es la definición de inteligencia que usa, sino la identificación de inteligencia con capacidad de realizar actos inteligentes a nivel humano.

Hay un punto que suele escapar a las discusiones sobre este paper sobre el que nos gustaría llamar la atención y es que Turing explícitamente dice “Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, *Can machines think?*” y es que el juego no trata de identificar siempre a la máquina como

[5]Este hecho ha sido analizado por diversos autores, como David Leavitt en *El hombre que sabía demasiado*, como muestra de sexismo o desconfianza para con las mujeres por parte de Turing. Si bien es un tema interesante, no consideramos que sea relevante en el legado del paper de Turing, por lo que no nos detendremos en consideraciones al respecto.

humana sino que los errores de identificación de un humano como mujer u hombre sean estadísticamente parecidos a los que se cometan en la identificación como hombre o máquina. Es decir, Turing acepta el error en la identificación y lo hace parte de la prueba ya que es consciente de que una característica fundamental del pensamiento humano es que éste no es exacto y, además, es propenso a errores en múltiples niveles[6].

Turing pasa a describir las máquinas que podrían ser programadas para participar en el test, que se corresponden con el nivel máquina de la arquitectura de computador digital que describimos en el capítulo anterior, es decir, los niveles de abstracción, como ya hemos explicado, facilitan la implementación del algoritmo proveyendo abstracciones sobre el "bare metal", la máquina básica de cómputo, pero no pueden aumentar las capacidades teóricas de la máquina. De la misma manera que al realizar deducciones matemáticas no partimos simplemente del concepto de número y nuestra demostración debe partir de ahí siempre, se toman generalizaciones dadas, por ejemplo, al inferir hechos sobre triángulos se recurre en primera instancia a la trigonometría y no se deduce, de nuevo, la trigonometría de los axiomas de la geometría. Turing indica claramente que lo relevante en éste constructo no es la máquina en sí, sino el algoritmo que produce los comportamientos inteligentes.

Turing comenta una serie de 9 objeciones que impedirían la construcción de dicha máquina, que aquí comentamos desde una perspectiva generalizadora, actualizando las críticas a dichas objeciones:

1, 2, 4, 5, 6, 9: La máquina no puede ser humana Agrupamos 1) Las máquinas carecen de alma, 2) Las consecuencias de la máquina inteligente son temibles ya que el hombre no debe ser sobrepasado, 4) Las máquinas no pueden tener consciencia ni sensibilidad, 5) Las máquinas no pueden X, siendo X una cualidad humana como "amar, aprender, ser el sujeto de su propio pensamiento, tener errores, etc.", 6) Ada Lovelace, en sus notas sobre el motor diferencial de Babbage, dice que la máquina "no puede hacer otra cosa que aquella que sepamos ordenarle cómo hacer, mas no tiene pretensión de generar nada por sí misma", es decir, carecen de creatividad y 9) Las máquinas no pueden tener Percepción Extra-Sensorial, como equivalentes al basarse todas ellas en la idea de que algo construído "no puede tener características humanas". Consideramos 1, 2, 4 y 9 no afectan al test al ser éste funcionalista[7] y las críticas estar fundamentadas en supuestas propiedades de la esencia humana cuya existencia no ha sido probada, o de la estructura interna de la mente. 5 y 6 son dudables ya que también carecen de justificación, basándose

[6]La pregunta de si un humano no pasa el test deja o no de ser inteligente queda implícita, por ello se hace ésta distinción estatística.

[7]Una de las pocas alternativas a 4, la máquina no puede ser consciente, es caer en el solipsismo, de hecho.

en los prejuicios del criticante, siendo "al estar programadas para el test, deberían ser capaces de simular X" una contra-crítica equivalente, que es la que de hecho Turing parece esgrimir.

3, 7: Objeciones matemáticas Al ser la base de la máquina inteligente propuesta por Turing un constructo capaz de realizar matemática discreta, surgen las siguientes objecciones: 3) La máquina no puede tratar problemas indecidibles de la lógica al estar basada en ella (i.e. preguntada por el problema de la parada la máquina daría una respuesta errónea o entraría en un bucle infinito), 7) La máquina es un constructo discreto mientras que el cerebro humano es continuo. Turing considera que no existe demostración de que las limitaciones de 3 no ocurran también en la mente humana y, que si una máquina da una respuesta errónea a estos problemas no es relevante puesto que la mayoría de humanos también la daría. Como respuestas a 7 se dan que los ajustes discretos a funciones continuas podrían dar resultados suficientemente similares para las preguntas formuladas en el juego, y además, como ya hemos visto las neuronas operan fundamentalmente en términos discretos.

8: El comportamiento humano es arbitrario Ésta objeción consiste en que no existe un procedimiento finito, algoritmo, para calcular o predecir el comportamiento de un sujeto humano. Se puede contraargumentar que no existe *aún*, como no existe una descripción completa de la física etc. o bien se puede argumentar que dado un programa que responde a un número de dieciséis cifras con otro arbitrariamente uno podría similarmente concluír que el programa es humano puesto que exhibe arbitrariedad y no hay reglas claras de su funcionamiento, aunque sabemos que éstas están en su código. Aún así consideramos que es una crítica relevante al funcionamiento del test, puesto que un humano puede identificar a la máquina como humana mientras que otro puede no hacerlo de forma arbitraria.

En la parte final del paper se describe el proceso de aprendizaje máquina, que ha dado un campo propiamente establecido en la inteligencia artificial, y que consiste en hipotetizar que la mente del niño es sustancialmente más simple que la del adulto, siendo la mente del adulto la mente del niño más experiencias de aprendizaje y, esto siendo la clave del constructo, la capacidad de generar ideas secundarias y terciarias a partir de experiencias o ideas inyectadas. Lo que Turing denomina, por analogía con los materiales atómicos, una *mente supercrítica*, en las que la inyección de material nuevo, i.e. ideas y experiencias, provoca una reacción que conduce a ideas nuevas: es decir, capacidad creativa y arbitrariedad, de la misma forma que los humanos generarían conceptos nuevos para ellos mismos o incluso para otros humanos. Por la misma

analogía, Turing describe las mentes animales como *mentes subcríticas* e incapaces de creatividad o generación de ideas a partir de ideas. Turing presupone que un método aleatoreizado puede ser mejor para la generación de este tipo de aprendizaje, ya que no cree ver reglas en la generación de éste tipo de conocimiento: todo ésto es también la base de otro campo del conocimiento, el razonamiento inductivo.

Existen reformulaciones posteriores del TT que admiten errores y la posibilidad de que la identificación de comportamiento inteligente se haga por un juicio grupal, como el usado en el premio Loebner[8], en el que la identificación humano o no-humano mayoritaria sea la aceptada por votos, o bien, en otros casos, mediante la aplicación del mismo test varias veces en diferentes momentos[9] dando a entender que la atribución de inteligencia debe estar sujeta a revisión. Schweizer en *The Truly Total Turing Test* (1998) arguye que el test debería ser aplicado no a individuos sino a máquinas capaces de reproducirse y tener descendencias en su misma "especie". El TTTT es un sucesor del Test de Turing Total propuesto por Harnad en *Other Bodies, Other Minds* en el que propone que las máquinas deberían ser capaces de realizar actos físicos para ser consideradas inteligentes, lo que se conoce como mente corporeizada[10], y que la capacidad lingüística es en realidad equivalente e indicativa de habilidad motora. Resulta evidente que una máquina que no pudiese pasar el TT pudiera pasar algunas de sus extensiones, por tanto debería ser uno de los primeros objetivos.

Como crítica, el funcionalismo del que habíamos hablado al citar a Ashby, se da también en la definición de Turing 1950, ya que identifica comportamiento inteligente como "comportamiento percibido como humano por un ser humano", además de ser un funcionalismo subjetivo en el que la validez de la observación es dependiente de un agente o varios, según la formulación del test, i.e. no es una métrica objetiva. Además, ésta definición es limitante, pues sólo identifica comportamientos inteligentes o pensamientos con actuaciones humanas, siendo ambas identificaciones comportamiento-pensamiento y humano-inteligente problemáticas: la identificación comportamiento-pensamiento impide a la IA dar una descripción de lo que es pensamiento más allá de "es comportamiento" y la identificación humano-inteligente impide la resolución de problemas que están más allá del proceso de pensamiento humano en formas diferentes que un simple aumento en la velocidad o precisión de los mismos.

Ésta crítica al funcionalismo también está implícita en el contra-argumento de la habitación china, por John Searle, uno de los más importantes. A continuación veremos ésta y otras críticas posteriores al paper de Turing.

[8] `http://www.aisb.org.uk/events/loebner-prize`

[9] Detallados en Moor, *An analysis of the Turing Test* (1976), lo que además indica que el término "Test de Turing" ya era usado en ese año.

[10] Harnad lo propone como crítica al TT, aunque Turing no dice que las habilidades lingüísticas deberían ser las únicas

3.3. Limitaciones del Test de Turing

Más allá de las objeciones que Turing anticipa en su propio paper, es interesante comprobar la polémica que éste suscita en sus contemporáneos y sucesores. Seguiremos para el estudio de la misma el comentario de (Pinar Saygin, Cicekli y Akman 2000) y (LaCurts 2011).

Al tratarse el test de Turing de una sustitución, de una pregunta demasiado amplia (*¿Puede pensar una máquina?*), por una pregunta restringida a un objetivo (*¿Puede convencer una máquina a un humano de que conversa como un humano?*) es natural que surjan maneras de cumplir éste objetivo que quizás entren en conflicto con nuestras preconcepciones sobre la primera pregunta. Esto es, en nuestro modelo de inteligencia existen falsos positivos aparentes; casos positivos en el método de testeo que no son, en apariencia, casos positivos en lo que intentamos modelar, la inteligencia; y falsos negativos, análogamente. Podemos arguír, como Purtill en *Beating the Imitation Game* (1971), que la máquina no es necesaria en ningún caso, y que realmente lo único que demuestra el Test de Turing es la inteligencia del constructor de la máquina. Creemos que ésta crítica es vacua al no existir diferencias no esenciales entre la creación de una máquina inteligente y la crianza de un niño inteligente, si la teoría sobre los indistinguibles que presentamos anteriormente se acepta.

Al tratarse de una sustitución de habilidades inteligentes por aquellas habilidades que pueden reflejarse únicamente mediante el lenguaje escrito, se entiende que el rango de inteligencia que puede medirse es menor. Gunderson, en *The imitation game* (1964), critica la validez de esta métrica de inteligencia usando un experimento mental. Imaginemos que tratamos de adivinar si el sujeto en la habitación es humano o no. Para ello disponemos de un agujero por el que meter el pie en la habitación del sujeto que se supone debe pisar para calificar como humano. Si se construye un sistema que deje caer una piedra cada vez que alguien introduzca el pie por el agujero, dice Gunderson, podríamos identificar la piedra como humano. Esta crítica una ejemplificación de un falso positivo, que se sostiene en que la capacidad de reconocer humanidad a partir de poder pisar un pie es limitada. Creemos que esta crítica es sensata en general, pero obvia que las capacidades derivadas del lenguaje tienen una relación más profunda con la inteligencia que pisar un pie con la cualidad de ser humano, a saber, con habilidades lingüísticas puede explicarse a otro ser que comprenda el lenguaje cómo realizar casi cualquier tarea. En todo caso, para que el TT sea una métrica completa de inteligencia debemos asegurar que esté libre de falsos positivos y falsos negativos.

Esto nos lleva a pensar en cómo atribuímos inteligencia a otros humanos. Normalmente ésto pasa por asumir que uno mismo es humano e inteligente, y mediante la interacción con otros humanos atribuír también a éstos inteligencia. La mayor parte

de éstas interacciones discurren, después de la identificación del sujeto en cuestión como humano[11], mediante la evaluación de los actos, no sólo de habla, de dicho sujeto. La parte lingüística y de comparación de ideas a través del lenguaje es ciertamente importante, aunque no es la única. Michie, en *Turing's Test and Conscious Thought* (1992) afirma que el TT no captura el pensamiento subconsciente y que las máquinas son incapaces de generar este tipo de pensamiento. Mitchie da como ejemplo que cualquier ser humano debería saber cuál es el plural de una serie de palabras inventadas en inglés. La crítica consiste en que es imposible que un programador considere todas las reglas del lenguaje, como las de pluralización, y las programe en una máquina. En realidad ésto puede ser innecesario al programar la máquina para que aprenda de usuarios del lenguaje reales y además, no evidencia la existencia de pensamiento subconsciente en el proceso lingüístico, sino que se trata de una regla lingüística aprendida y generalizada en la mayor parte de los humanos que conocen el idioma inglés. En cualquier caso, ante el hecho de que la máquina demuestre inteligencia que el TT no captura, existen competencias clave que el TT sí captura

Volviendo al tema de la atribución de inteligencia es interesante considerar qué ocurriría si sustituímos al evaluador del TT por una máquina. La máquina, de poseer inteligencia equivalente a la humana, debería ser capaz de identificar correcta e incorrectamente máquinas y humanos con una tasa de errores similar a la de un humano[12]. Podemos considerar si para la identificación correcta y la imitación de un humano correcta (ambos roles del TT) debe ser necesaria la existencia de una teoría de la mente en dicha máquina, es decir, la capacidad de atribuír estados mentales a entidades que interaccionen con la misma. La respuesta a esta pregunta determinaría si la atribución de estados mentales y la identificación de actos inteligentes es un comportamiento inteligente en sí mismo que el TT original no captura. Este test fue propuesto por Watt en *Naive Psychology and the Inverted Turing Test* (1996).

En relación a esta crítica está la habitación china de Searle, que postula una habitación que recibe mensajes escritos en chino teniendo al propio Searle dentro (que no entiende el chino) equipado con un manual que le dice qué simbolos escoger para elaborar sus respuestas a los mensajes entrantes. Searle critica que dicha habitación podría ser capaz, en teoría, de superar el TT. Sin embargo, ninguna de las partes componentes entiende el lenguaje y por tanto la habitación carece de intencionalidad, que para Searle es una parte esencial de la inteligencia. Podemos arguír que el único libro que puede hacer que Searle sea capaz de elaborar respuestas con sentido para que pueda pasar

[11] La atribución de inteligencia humana se simplifica normalmente como atribuíble por humanos al ser una cualidad esencial de éstos.

[12] Si la máquina tuviese menos errores podríamos concluír únicamente que tiene una inteligencia especializada en éste problema a menos que de alguna otra manera muestre que esta especialización surge de habilidades mayores que los humanos.

el TT es un manual que enseñe chino a Searle o un manual que traduzca entre frases del chino y el inglés perfectamente o una transformación entre frases en chino y un modelo de datos de la situación que se va desarrollando durante la conversación[13], ya que cualquier conversación en chino o en cualquier otro idioma no posee una estructura tan simple como una tabla pregunta-respuesta, como Searle parece entender. Evaluar el lenguaje no es trivial en todo caso, y en una conversación se establecen hechos y relaciones entre términos (véanse juegos de palabras por ejemplo) que una máquina que pase el TT debe ser capaz de identificar y desarrollar para que la comunicación sea humana[14]. Por lo tanto, o bien las instrucciones incorporan un mecanismo de comprensión de palabras en chino asociadas a la respuesta, lo que hace necesario un mecanismo recursivo sobre las comunicaciones recibidas y el modelo de la situación conversacional, o bien Searle puede entender el chino. El tema de la habitación china de Searle tiene el suficiente alcance para poder desarrollar varios artículos por sí mismo. Forma parte del movimiento anti-funcionalista o anti-conductualista, en el campo psicológico, de la década de los 80 y 90.

Más allá de la limitación del test respecto a la inteligencia, el identificar actos no inteligentes como inteligentes y viceversa, cabe preguntarse si los actos inteligentes que se identifiquen sean inteligencia humana o no. La crítica de French en *Subcognition and the Limits of the Turing Test* (1990) postula un test para la capacidad de volar en una isla del mar del Norte cuyos únicos ejemplos de objetos voladores son gaviotas. El paralelismo se traza en el momento en que una máquina capaz de volar cruza su modelo de "objeto volador", que está basado en las gaviotas y asume que la capacidad de volar de éstas debería ser indescirnible de cualquier entidad voladora. La conclusión de este experimento mental es la existencia de falsos negativos en el test, en el sentido de comportamientos inteligentes *no humanos* que pasen desapercibidos al test. De nuevo nos topamos con el problema de la definición de inteligencia, que en general toma la inteligencia asumida en humanos como base. Minsky, en concreto, en *Communication with Alien Intelligence* (1985) postula que la manera en la que los humanos piensan, dividiendo y abstrayendo la realidad en objetos y conceptos simples, es comparativamente eficiente y tal vez por eso sea una muestra de *inteligencia general* y el único modelo de inteligencia.

[13]Por ejemplo, si el mensaje incluye "Yo soy filósofo" en chino, recuerda que el adversario es filósofo y construye todas los hechos relativos a cualidades de los filósofos de acuerdo a x reglas. Otras críticas similares incluyen el "saber todas las conversaciones de longitud x previamente" de Block en *Troubles with Functionalism* y *Psychologism and Behaviorism* y se dedica a dar respuestas desde una lista que obvian que es imposible construír dicha lista al ser todas las conversaciones infinitas no numerables y dependientes del contexto y de sí mismas.

[14]Es posible incluso que la habilidad de la máquina para simular la cultura de su adversario haga que la identificación sea más o menos precisa. Existe material desarrollado en las ciencias sociales sobre el TT, listado en (Pinar Saygin, Cicekli y Akman 2000)

Otras críticas, también de French, involucran experiencias singularmente humanas como necesarias para la inteligencia, como el ser capaces de identificar hojas secas como un escondrijo útil o la evaluación de nombres de cereales como atractivos o no. Se trata de una limitación del TT clara, por falsos negativos potenciales: el hecho de que si una máquina no puede empatizar o pensar en una experiencia humana no será identificada como inteligente en el TT. Las soluciones propuestas desde la IA son o bien dar a la máquina entrenamiento en vivencias humanas o bien cambiar el test para que las preguntas relativas a vivencias personales no sean relevantes. Si la primera es una solución efectiva o la segunda una limitación a la evaluación completa de inteligencia son preguntas aún abiertas. Como extensión a esta crítica agregamos la super-expresividad (*superarticulacy* en inglés) que hace referencia a la capacidad de las máquinas inteligentes de, posiblemente, superar a los humanos en algunas tareas que éstos realizan de forma intuitiva, siendo rechazados en el TT por ésto mismo[15]: en todo caso, ésto podría ayudar a crear nuevas ideas más precisas sobre la inteligencia humana, el desarrollo de nuevos test más precisos y el aumento de las propias capacidades humanas relativas a dichas tareas.

[15]Crítica ya anticipada por Turing, que resuelve forzando a las máquinas a "hacerse estúpidas" en el momento que ésto ponga en peligro el test.

Capítulo 4

Conclusiones

En este trabajo hemos expuesto perspectivas históricas referentes a la creación de máquinas inteligentes en los campos de la maquinaria computacional y la algoritmia como referentes a la base de la inteligencia y la definición de inteligencia en la parte filosófica del problema. Se ha ligado el problema de la IA con el problema de los indescernibles y la filosofía de la tecnología además de señalar relaciones entre el TT y las disciplinas académicas que tratan el tema de la inteligencia y la maquinaria computacional (i.e. psicología, algoritmia, sociología, etc.), identificando puntos de partida para el desarrollo de investigaciones en dichas disciplinas.

Se ha hecho una diferencia entre la definición de inteligencia y los comportamientos inteligentes, permitiendo evadir el platonismo de asumir que inteligencia es una entidad, por lo que se caracteriza el Test de Turing como una forma productiva de comprobar si una entidad realiza, de facto, comportamientos inteligentes.

Se han identificado las limitaciones generales del TT:

1. Hace posible identificar máquinas carentes de inteligencia general como tales (falsos positivos).

2. Hace posible que máquinas capaces de actos inteligentes pero sin habilidades conversacionales no sean considerados como inteligentes (falsos negativos). Englobamos aquí la posible existencia de inteligencias no humanas que también pasarían desapercibidas al test.

3. En ningún caso el test avanza hacia una definición más precisa de inteligencia, siendo necesaria una meta-evaluación del test para llegar a ella.

Vemos que el TT aún así nos da una definición operativa, aunque no sea usada en los grupos que crean algoritmia inteligente orientada a tareas[1], es necesaria para la identificación de inteligencia general. Uno de los objetivos del *endgame* de la IA es la identificación y caracterización de los comportamientos inteligentes y el desarrollo

[1]La definición operativa suele basarse en métricas de *precision* y *recall* sobre conjuntos de datos estándar, para comprobar mejoras y diferencias en ajuste de funciones y clasificación.

de tests objetivos y reproducibles de la misma. La necesidad de colaboración con la psicología, la neurociencia y el resto de las ciencias cognitivas se hace aquí evidente.

Apéndice **A**

Historia del Hardware Computacional

Se incluye aquí un resumen de la historia del hardware de computación extraída de (Wikipedia 2015).

Históricamente el computador de propósito general fue diseñado por Charles Babbage en 1833, aunque no fue construído hasta el siglo XX por, primero, su hijo Prevost que construyó una parte mínima esencial del mismo que era capaz de ejecutar programas simples en 1910, y en última instancia fue imitado por el Museo de Ciencias de Londres con materiales de la época sin éxito a finales del siglo XX. Aún así se considera que el prototipo, de llegar a cumplir sus especificaciones de diseño, sería Turing-completo, sólo limitado por la precisión numérica y la memoria total de manera similar a la que cualquier computador actual lo sería. En la primera mitad del siglo XX encontramos gran cantidad de máquinas de cómputo, incluso cajeros automáticos de cobro mecánicos y electrónicos con sistemas de ayuda al cálculo de balances, pero ninguna cumple la propiedad de ser lo suficientemente general para ejecutar algoritmos recursivos.

Será en 1944 cuando se cree el Colossus Mark II, el primer computador electrónico digital Turing-completo que se mantuvo en secreto hasta la década de 1970 ya que fue planeado también para romper códigos, y usado durante la guerra fría. Éste constructo es otro ejemplo de la confluencia de ideas que se fue gestando desde el siglo XIX con la publicación de George Boole[1] *Las leyes del pensamiento*, que inaugura el cálculo lógico binario (o booleano), y desarrollado por Alfred Whitehead. Claude Shannon, fundador de la teoría de la información como entropía e ingeniero eléctrico, demuestra que las operaciones de la lógica booleana pueden construírse mediante la adaptación a circuítos eléctricos entendiendo el símbolo básico 0 como la ausencia de corriente eléctrica y el símbolo 1 como la presencia de la misma. La idea de usar la lógica simbólica como base de todas las matemáticas, y la idea de *Gödelización* como transformación de enunciados en números y su posterior procesado, confluyen también en la generación de la idea de la Máquina de Turing Universal ya presente en *On Computable Numbers* y consistente en la aplicación de una máquina de Turing que puede aceptar como entrada otras máquinas de Turing transformadas en números naturales y ejecutarlas. Se puede

[1] Otros ejemplos de ésta confluencia son los trabajos en lógica de Russell y Whitehead, el proyecto formalista de Hilbert o los teoremas de Gödel.

demostrar también que una máquina de Turing, para ser Turing-completa, únicamente precisa del uso de dos símbolos, 0 y 1.

Cabe notar que éste computador era programable en el sentido de que permitía la interconexión arbitraria de elementos que no es que tuviesen programas almacenados sino que eran módulos de cálculo especializados en una tarea concreta referida al cálculo booleano o de aplicación criptográfica, no porque permitiese especificar un programa a nivel de memoria como se hace actualmente, siendo gran parte de estas tareas gestionadas por el sistema operativo. Para la carga en memoria de programas se tendría que esperar a la arquitectura von Neumann y al desarrollo de los chips y memorias magnéticas.

En la década de los 50 todas estas ideas se amplían y se comercializan. La invención de la microprogramación, hoy conocida como *firmware* que permite la adición de nuevas instrucciones programadas sobre una máquina física, sin cambiar la máquina, simplifica el desarrollo de las máquinas de cómputo haciéndolas más flexibles. En 1955 los computadores sustituyen los circuítos basados en tubos de vacío por circuítos basados en transistores, reduciendo su tamaño en órdenes de magnitud y aumentando su velocidad en igual medida. Ésto se suele nombrar en la literatura como "segunda generación".

En la década de los 60 se produce el gran salto de la computación al consumo generalizado, no por parte de la inmensa mayoría de la población sino por instituciones y grandes empresas, lanzándose el primer computador que podríamos llamar "popular", el IBM 1401. Consistía en un computador configurable que se podía alquilar por unos veinte mil euros mensuales, al cambio actual, lo que permitió a un gran número de entidades entrar en el uso de las computadoras. Lo que hoy conocemos como mainframe requería que un ingeniero configurase el computador y cargase los programas ya escritos en el mismo y preparados para el lenguaje de la máquina que los ejecutaría, decidiendo cuánto tiempo debería gastar cada uno y en qué orden: la automatización de estas tareas dará lugar en las próximas décadas a los sistemas operativos, el uso de mainframes en tiempo compartido y en última instancia a los sistemas distribuídos y paralelos.

Tambien de esta época son los primeros lenguajes de programación por encima del nivel de lenguaje máquina, LISP y Fortran, que requerían una etapa de traducción intermedia y aún se usan hoy en día. LISP es uno de los lenguajes clave en el desarrollo de la IA en las últimas décadas, ya que, basado en el cálculo lambda de Church, es un lenguaje que permite expresar programas de procesado de listas mediante listas de instrucciones donde datos y programa están expresados en el mismo lenguaje. Los primeros sistemas operativos, como OS/360 cuya historia de desarrollo está narrada en *The mythical Man-Month*, también son de ésta época. En la era del mainframe cada

una de las máquinas solía ser enviada al cliente con un sistema operativo adaptado a sus necesidades específicas, los sistemas operativos de consumo general, configurables por sí mismos, no aparecerán hasta la generalización de componentes de consumo que permitan que el sistema operativo pueda ser manufacturado por separado de la máquina, e integrado en varias máquinas.

Los sistemas de tercera y cuarta generación, circuítos integrados conectados mediante cables y circuítos integrados conectados mediante circuítos integrados, serán desarrollados a mediados de la década de los 60 y de la década de los 70 respectivamente. Dado que las diferencias entre ambas generaciones sólo aparecen durante el proceso de manufactura de los mismos hemos decidido obviarlas y centrarnos en los efectos que produce sobre el comercio de computadores: mayor velocidad en menor tamaño y menor precio. Para finales de la década de los 70 el computador personal ya era un hecho y los Apple II, Commodore e IBM-PC podían verse en hogares de todo el mundo, incluída España, donde se considera que hubo una *edad de oro* del software durante los años 80.

Entre los años 80 y nuestra era los computadores han ido incrementando su capacidad computacional en velocidad y capacidad de memoria de acuerdo con la Ley de Moore, que postula que el avance de tecnología permite introducir el doble de transistores en una placa del mismo tamaño más o menos cada dos años. Ésta ley se ha topado recientemente con la limitación práctica de que el calor disipado y energía consumida por un computador altamente integrado también aumentan, precisando que dicho computador posea refrigeración y alimentación eléctrica no razonables. También, teóricamente, se postula un límite duro para dicha ley en el momento que los transistores no puedan funcionar debido a su relación de tamaño con la escala de los fenómenos cuánticos. Por estas razones hoy se tiende a aumentar la eficiencia de los sistemas no aumentando la cantidad de operaciones en serie, sino que se aumenta la cantidad de operaciones que pueden hacerse simultáneamente mediante paralelización o distribución de tareas.

Bibliografía

Aristóteles (1988). *Política*. Trad. por Manuela García Valdés. Gredos.

Ashby, William Ross. (1956). *An introduction to cybernetics*. Trad. por Jorge Santos. *Introducción a la cibernética, 1976, Ediciones Nueva Visión SAIC, Argentina*. Chapman & Hall Ltd., London.

Echeverría, Javier (1998). «Tecnologías, Espacios de Interacción y valores». En: *Teorema* XVII/3:11-25.

Franssen, Maarten, Gert-Jan Lokhorst e Ibo van de Poel (2013). «Philosophy of Technology». En: *The Stanford Encyclopedia of Philosophy*. Ed. por Edward N. Zalta. Winter 2013. URL: http://plato.stanford.edu/archives/win2013/entries/technology/.

Hilpinen, Risto (2011). «Artifact». En: *The Stanford Encyclopedia of Philosophy*. Ed. por Edward N. Zalta. Winter 2011. URL: http://plato.stanford.edu/archives/win2011/entries/artifact/.

Hopcroft, John E., Rajeev Motwani y Jeffrey D. Ullman (2006). *Introduction to Automata Theory, Languages, and Computation (3rd Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc. ISBN: 0321455363.

LaCurts, Katrina (2011). *Criticisms of the Turing Test. and Why You Should Ignore (Most of) Them*. MIT. URL: http://people.csail.mit.edu/katrina/papers/6893.pdf.

Lakoff, George (1987). *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press. ISBN: 978-0-226-46803-7. URL: http://emilkirkegaard.dk/en/wp-content/uploads/George-Lakoff-Women-Fire-and-Dangerous-Things.pdf.

Legg, Shane y Marcus Hutter (2007). «A Collection of Definitions of Intelligence». En: *CoRR* abs/0706.3639. URL: http://arxiv.org/abs/0706.3639.

Martínez-Freire, Pascual (2005). *La importancia del conocimiento: filosofía y ciencias cognitivas*. Segunda Edición. Thema, Universidad de Málaga.

McCarthy, John (2007). *What is Artificial Intelligence?* EN. Última revisión 2007, documento web vivo. URL: http://www-formal.stanford.edu/jmc/whatisai/whatisai.html.

Nilsson, Nils J. (2009). *The Quest for Artificial Intelligence: A history of ideas and achievements*. 1st. New York, NY, USA: Cambridge University Press. ISBN: 0521122937, 9780521122931. URL: http://ai.stanford.edu/~nilsson/QAI/qai.pdf.

Penrose, Roger (2006). *La Nueva Mente del Emperador*. Trad. por Javier García Sanz. Barcelona, España: Random House Mondadori.

Pinar Saygin, Ayse, Ilyas Cicekli y Varol Akman (2000). «Turing Test: 50 Years Later». English. En: *Minds and Machines* 10.4, págs. 463-518. ISSN: 0924-6495. DOI: 10.1023/A:1011288000451. URL: http://www.cs.bilkent.edu.tr/~ilyas/PDF/minds2000.pdf.

Pylyshyn, Z. W., ed. (1970). *Perspectives on the computer revolution*. Englewood Cliffs, NJ: Prentice-Hall.

Searle, John R. (1980). «Minds, brains, and programs». En: *Behavioral and Brain Sciences* 3 (03), págs. 417-424. ISSN: 1469-1825. DOI: 10.1017/S0140525X00005756. URL: http://journals.cambridge.org/article_S0140525X00005756.

Turing, Alan M. (1937). «On Computable Numbers, with an Application to the Entscheidungsproblem». En: *Proceedings of the London Mathematical Society*. Vol. 42. 2. URL: http://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf.

— (1950). «Computing Machinery and Intelligence». En: *Mind* 59, págs. 433-460. URL: http://cogprints.org/499/.

Wikipedia (2015). *History of Computing Hardware*. [Online; 05/08/2015]. URL: https://en.wikipedia.org/w/index.php?title=History_of_computing_hardware&oldid=673753505.

Grand Master Turing once dreamed that he was a machine. When he awoke he exclaimed:
"I don't know whether I am Turing dreaming that I am a machine, or a machine dreaming that I am Turing!"

From the Tao of Computer Programming