eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Facultad de Informática

Grado de Ingeniería Informática

▪ Trabajo Fin de Grado ▪

Computación

Revisión Crítica del Test de Turing, la Inteligencia Artificial Fuerte y propuesta de nuevo Test

---

Alfredo Vallejo Martín

Junio 2022

Alfredo Vallejo Martín

Resumen

En el presente trabajo se pretende explorar en profundidad la idea de la Inteligencia Artificial y la posibilidad de la creación de una. Para llevarlo a cabo se va a recurrir a escuelas de la filosofía para definir una visión de conjunto desde la que conceptualizar el tema.

Para entender mejor la concepción que tenemos de la IA en la actualidad se ha llevado a cabo una investigación de la historia de los pensadores y movimientos filosóficos en que nació este campo.

De la misma forma se van a explorar los argumentos que se han dado contra la posibilidad de la creación de una entidad semejante, analizando de forma rigurosa su validez y solidez. Toda esta investigación servirá para poder llevar a cabo a continuación un análisis del test de Turing tal y como se formuló en origen. El objetivo de este análisis es identificar sus debilidades para posteriormente contrastarlo con el estado del arte actual.

Finalmente este trabajo desemboca en una propuesta alternativa para el test de Turing de carácter original que supere y evite los problemas identificados en apartados anteriores, y que esté alineada con las ideas e intuiciones desarrolladas en los primeros apartados. Así como un pequeño ejemplo de su aplicación a un caso de actualidad, LaMDA.

Índice

1

Introducción

El Test de Turing en su formulación original sustituye la pregunta “puede una máquina pensar” por “podría una máquina engañar a un humano”. En su momento esta reformulación fue un ejercicio novedoso de uno de los pioneros del campo de la computación. Sentó el objetivo a lograr para el incipiente campo de la inteligencia artificial, cuando no el campo mismo. Pero tras más de 70 años de avances en el campo y la posibilidad cercana de que pueda superarse, las deficiencias que tiene el test como proxy para la inteligencia son notables, así como los incentivos para aprovecharse de la formulación de Turing y lograr su objetivo sin conseguir una inteligencia real. Simulacros diseñados específicamente para pasarlo, pero sin ningún tipo de inteligencia real.

1.1 Objetivos

El presente trabajo se propone como una investigación rigurosa y profunda en la cuestión del Test de Turing y el campo de la Inteligencia Artificial en general. El objetivo principal del mismo es intentar aportar un enfoque novedoso que permita abordar el tema de la conciencia en la Inteligencia Artificial, huyendo de dogmatismos y dar una serie de directrices o indicadores para poder detectarla. Se plantea un objetivo tan ambicioso, debido a que sin una conciencia todo uso del término inteligencia ha de ser considerado metafórico, pero no real. Para ello el trabajo se compone de los siguientes momentos:

- Marco interdisciplinar: en este apartado se definirá de forma rigurosa el marco conceptual desde el que parto. Para ello me centraré en construirlo desde la concepción metafísica de la conciencia, sus características indispensables y un análisis crítico del concepto de la inteligencia, añadiendo en el camino algunas nociones de neurociencia y biología.

- Investigación de la IA: este apartado se centrará en un desarrollo histórico de algunas de las cuestiones centrales del campo de la IA. En esencia este apartado se podría entender como una vasta contextualización del campo mismo, poniéndolo en relación con el resto de disciplinas sobre las que se apoya o de las que recibe influencia. A su vez en él se explorarán algunas de las cuestiones centrales de la IA, siendo el debate sobre la

consciencia y las máquinas el que tendrá un papel preponderante. Necesario por otra parte si pretendemos tomarnos en serio a Turing y su propuesta.

- Crítica del Test de Turing: este apartado se centrará en elaborar una crítica al Test de Turing, fundamentada en una lectura atenta del Test tal y como lo propuso en su artículo original. Esta crítica se basará ostensiblemente en muchas de las intuiciones e ideas desarrolladas en el apartado anterior. Su objetivo es sacar a la luz todos los errores del planteamiento de Turing, como propedéutica para el apartado siguiente, gracias al cual poder evitar algunos de sus errores.

- Propuesta Alternativa: Este apartado, de los tres del trabajo, el de carácter más especulativo, se centrará en analizar el estado del arte de los tests que se proponen hasta la fecha como alternativas o mejoras al test de Turing, para a continuación realizar una síntesis de ellos y proponer un diseño, formalización u esquema para un nuevo test que venga a sustituir al de Turing. Para su elaboración se dependerá, al igual que en el apartado anterior, de todas las intuiciones y conceptos desarrollados en el primero.

1.2 Motivación y Justificación del contenido y la estructura del trabajo

El presente trabajo se presenta como una investigación del test de Turing, la inteligencia artificial y la posibilidad de una conciencia en las máquinas. Dado este objetivo tan ambicioso y extenso, he considerado necesario e indispensable tratar el tema desde una perspectiva holística en vez de limitar el análisis al enfoque y cuestiones que se dan en esta carrera. Esto ha supuesto que haya tenido que ampliar el enfoque de la investigación a otras áreas del conocimiento. El área que mayor influencia e importancia ha cobrado ha sido la filosofía, sobre todo la fenomenología y el postestructuralismo(términos que ya se explicarán más adelante), aunque no solo de ellas.

Tras investigar bastante sobre el tema, sobre todo las opiniones y desarrollos teóricos de profesionales adyacentes o provenientes de nuestro campo [2](solo por mencionar dos). He quedado francamente decepcionado por los planteamientos ahí desarrollados: llenos de asunciones ocultas que no se hacen explicitas en ningún momento, que hacen pasar razonamientos endebles y cargados de “peros” por certezas incontestables y autoevidentes. Asunciones sobre la conciencia, sobre la realidad, sobre la forma de procesarla, etc.; todas ellas de carácter marcadamente metafísico, pero en ningún momento discutidas ni reconocidas como tales. Igualmente usos de las palabras completamente licenciosos e imprecisos y que solo en virtud de estos usos dudosos, sus argumentos parecen sostenerse.

El caso paradigmático es el uso y equivalencia de la palabra memoria entre su uso para personas y para máquinas, que dado que se pueden decir que son funcionalmente isomorfas, por lo menos en cierto nivel de abstracción, les da la seguridad para hacer equivaler humano a máquina. Esta equivalencia pasa por alto que la memoria humana es fundamentalmente distinta a la máquina. Con esto no me refiero a que tengan sustratos distintos, ya que es una trivialidad tautológica que no tiene por qué comportar una imposibilidad insalvable. Me refiero al hecho de que a no ser que haya algún bitflip por un defecto mecánico, o un rayo cósmico impacte en el circuito de almacenamiento y altere el contenido de un bit, la memoria de una máquina se mantiene inalterada sin importar cuantas veces se revisite “un recuerdo” o recuerde un dato.

Por el contrario nuestra memoria es mucho más falible, y no me refiero a nuestra capacidad de olvido, sino que cada vez que revisitamos un recuerdo, lo recordamos distinto, incluso lo alteramos de cara a futuras revisiones. Nuestra memoria es narrativa, intenta integrar los hechos en una narración de nuestro yo presente, hasta el punto de ser capaz de retorcer los hechos y alterarlos. Este es el ejemplo más flagrante de usos impropios, pero no el único. Son ejemplos de palabras utilizadas metafóricamente o análogamente, pero de una forma que parece escapar al propio emisor, confundiendo la realidad enunciada por la realidad que se pretende enunciar.

Lo mismo puede decirse de la idea más generalizada en nuestro campo de entender la mente como un software y el resto del cuerpo como un hardware. Una idea que se crítica más adelante en un apartado posterior, por hacer retornar el fantasma, ya muerto y enterrado a nivel neurobiológico, del dualismo cartesiano.

Por este motivo he optado por un enfoque alternativo. He preferido realizar un excurso inicial por la filosofía para intentar definir un marco coherente que deje bien claro tanto mis asunciones de partida, como mi forma de entender la cuestión, para luego poder desarrollarla de la forma más satisfactoria posible.

A su vez, el enfoque y la filosofía traída para estudiar y entender la cuestión, es de un carácter marcadamente distinto y alejado de la filosofía habitual subyacente a las "ciencias duras" o "formales" como la nuestra (lo que más adelante se denomina como filosofía analítica). Si he elegido este punto de vista es porque me resultaba interesante descentrar la visión habitual que se tiene de estas cuestiones y comprobar si nuestra perspectiva y comprensión del problema podía verse favorecida o enriquecida por una visión alternativa.

Esta filosofía de la que hablo que ha permeado a las "ciencias duras" en general y a la nuestra en particular, se explora, justifica y desarrolla en el segundo apartado. He optado por añadir este apartado para mostrar como una inercia histórica ha limitado nuestra forma de interpretar esta cuestión y favorecido una serie de ideas mientras se alejaba de otras. Teniendo en cuenta esta visión se puede explicar la concepción de los pioneros de la informática y de la IA en particular y como determinadas visiones se convirtieron rápidamente en hegemónicas dentro del campo. A su vez, considero que solo mostrando, aunque sea de una forma sencilla y esquemática, como esa visión cristalizó y se solidificó en nuestro campo, mi motivación para desviarme tan notablemente de las concepciones habituales y traer conocimientos de otras áreas estaría justificado.

Al mismo tiempo, este desarrollo histórico y enfoque genealógico encaja especialmente bien con el trabajo, en la medida en que el apartado 3 se enfoca en analizar el test de Turing. Solo si tenemos en cuenta el contexto intelectual y filosófico en que vivió y se desarrolló Turing, podemos entender como su pensamiento evolucionó en esa dirección. De una forma similar, la segunda parte del apartado 3 se concentra en hacer un pequeño resumen de la historia de la IA después de la muerte de Turing, recapitulando sus primeros 30 años de vida. Considero de interés estos aspectos y desarrollos porque por un lado sirven de contexto en los que insertar los argumentos en contra de la IA fuerte que se describen en el apartado siguiente; mientras que también nos permiten ver el momento actual, las predicciones a futuro y las declaraciones

optimistas con un poco de perspectiva, reconociendo que en cierta medida estamos ante una historia que se repite.

El último apartado antes del contenido realmente original estará dedicado a llevar a cabo un análisis del test de Turing partiendo desde el artículo original en que lo presentó. Con esto pretendo identificar sus debilidades y traer a la luz la concepción que tenía Turing de la mente y su funcionamiento, para confrontarla con la desarrollada en el segundo apartado.

1.3 Conclusión

A la vista del desarrollo anterior, queda solventada la cuestión de la justificación de la inclusión del resto de ciencias y campos de estudio. Pero antes de dar paso al siguiente apartado me gustaría recalcar que el presente trabajo tampoco se va a convertir en un collage de doctrinas, reducido a una vorágine de citas bibliográficas dispersas; o que por ello la cuestión principal y objeto de estudio, la IA, vaya a perder su preponderancia y carácter primario.

A su vez, la presencia del resto de campos tampoco ha de entenderse como algo intimidatorio que pretende oscurecer y enfangar el estudio que estoy por desarrollar. Todas aquellas cuestiones que sean ajenas a nuestro campo de estudio serán debidamente contextualizadas y explicadas y si su presencia no se justifica en el desarrollo subsiguiente, serán recordados cuando jueguen un papel determinante en una argumentación o desarrollo.

Por último, dada la naturaleza heterogénea del primer apartado, en el que los temas serán tan dispares, es imposible articularlos en un una exposición en que cada apartado se refiera al anterior y deduzca el siguiente eslabón a comentar o al menos, es imposible que así sea por entero. Esto es algo inevitable dado que el campo de la IA se encuentra en la intersección de otros muchos, por ello mi intención es ir planteando tema por tema la constelación de disciplinas. Quien quiere hallar la intersección de múltiples conjuntos, necesita examinarlos uno por uno para al final reunirlos en su área de solpamiento. Esta labor de unificación se llevará a cabo al final del primer apartado en la sección denominada Marco de análisis.

2

Marco Interdisciplinar

Empezaremos la exposición de los temas que constituyen el marco por la parte más general de todos ellos, es decir aquella que refiere a la metafísica. Esto es necesario en la medida en que la metafísica aunque no se discuta de forma explícita, subyace como poso o resto, como principio dado por supuesto, a cualquier posición. Quien considere que su pensamiento está libre de metafísica está, al igual que quien cree que no tiene un ideología, precipitándose de cabeza a una posición metafísica de carácter acrítico o dogmático. Una vez dado este iremos procediendo poco a poco a otros temas de carácter más cercano a la realidad y la experiencia.

2.1 Marco metafísico

Consideramos necesario empezar la exposición así, debido a que el tema de la consciencia está íntimamente relacionado con él de la metafísica. Empecemos nuestro excurso por este campo clarificando cual va a ser el modelo de la realidad que vamos a utilizar, así como la concepción de la conciencia y sus propiedades.

Es necesario señalar que si la conciencia se la tratará en este apartado y no en otro es porque la conciencia es una de las entidades metafísicas por antonomasia. Como bien señala Kant en el apartado de los paralogismos de la razón pura [3], la conciencia es algo que se presupone en toda experiencia empírica, pero de la que nunca se tiene experiencia empírica.

2.1.1 La conciencia

La conciencia como concepto se haya enredada entre dos concepciones contrapuestas. Se la entiende tanto en un sentido fenoménico, como en un sentido lógico o referente a la capacidad de razonamiento.

En su dimensión fenoménica, la conciencia refiere a la experiencia subjetiva de la realidad por parte de la entidad que la posee. Dicho de otra forma, la conciencia sería la realidad mental inmediata que experimenta el sujeto, fruto de las percepciones sensoriales. A estas realidades mentales se las llama *qualias* y hacen referencia a las distintas propiedades o estados distinguibles que nos son dados en la experiencia. Un ejemplo, sencillo sería el color rojo, este, aunque es producto de la percepción de ondas electromagnéticas, de una determinada longitud, por parte

de células foto receptoras, en la conciencia se manifiesta como la propiedad cualitativa rojo, aplicada a un determinado objeto [4].

En su dimensión lógica la conciencia refiere a la capacidad de autoconciencia. En este aspecto, la conciencia se puede equiparar a "nuestra voz interior", en la que nos reconocemos como entidades corporeizadas y separadas de nuestro ambiente. A su vez, es gracias a ella que somos capaces de efectuar razonamientos y reflexionar, ya sea sobre el mundo exterior o sobre nosotros mismos [4].

Estas dos concepciones no son mutuamente excluyentes, sino que están integradas en una sola que es lo que *strictu sensu* llamaríamos la conciencia. La dimensión fenoménica supone el sustrato de experiencias y estados cualitativos que recibe y procesa la dimensión lógica, para llevar a cabo sus razonamientos y modificar su comportamiento. Cualquier intento de privilegiar una por encima de otra nos lleva a imposibles. Sin la autoconciencia para guiarnos la conciencia fenoménica solo sería un puro estar en el momento presente, mientras que sin fenómenos la autoconciencia no tendría capacidad de razonamiento. Idea resumida en la máxima kantiana: "Las intuiciones sin concepto están ciegas y los conceptos sin intuición, vacíos". [3]

En resumen, la concepción de conciencia de la que vamos a partir es aquella que integra ambos aspectos y es capaz de tener tanto estados cualitativos (referidos a experiencias), como cognitivos.

De la misma forma aunque no se deduzca a priori voy a huir de todo planteamiento que intente dividir ambos aspectos de la conciencia y plantear que puede existir uno sin el otro [4]**.** Encuentro que estos planteamientos son altamente dudosos y constituyen más un ejercicio de pensamiento mágico excesivamente optimista, que una intuición sólida a tener en cuenta.

Mi motivo para descartarlos tan apresuradamente, es que solo se basan en la premisa, errónea a mí parecer, de que dado que podemos discernir y discriminar un tipo del otro y concebir o aproximar la existencia de un sistema con solo un tipo de esas capacidades (generalmente un sistema exclusivamente cognitivo), entonces esto es una prueba de que son posibles. Pero al igual que con los argumentos ontológicos de la existencia de Dios, que podamos pensar en algo no significa que exista.

Implícitamente se está partiendo de la siguiente premisa: cómo podemos diferenciar estos dos aspectos en el sistema, son dos sistemas distintos que pueden existir de forma diferenciada y se acoplan dando mayor complejidad a las experiencias. Pero esta posibilidad no se demuestra ni discute abiertamente y los pocos ejemplos que he encontrado que la discuten tangencialmente, se basan en experimentos mentales y analogías con animales [4]**.** A mi juicio esta concepción es errónea y bebe excesivamente de una tradición que podríamos calificar de cartesiana, que concibe que la conciencia se puede dar sin un correlato objetivo en el mundo material y puede existir como pura idealidad mental.

2.1.2 El estatus ontológico de la conciencia

Esta cuestión lejos de ser trivial es de una importancia fundamental a la hora de proceder, ya que el marco elegido en esta cuestión determina por completo las conclusiones posteriores. Debido

a que la conciencia, como ente, es inmaterial, todo postulado sobre ella se cimenta sobre asunciones metafísicas. Por ello, será necesario presentar las principales candidatas y elegir una de ellas. Pero antes de nada será necesario clarificar cuales son las posturas de las que emergen.

- El monismo: como doctrina metafísica el monismo plantea que la realidad está constituida por una única sustancia. Esta sustancia o materia puede ir mutando y variando constituyéndose de distintas formas que permiten una interacción y variación que da pie a la riqueza plural de nuestra realidad. Un ejemplo sería el materialismo inmanente[1], en qué solo existe lo material o físico y todo lo demás es un subproducto del mismo originado por fuerzas interiores a él. [5] [6]

- El dualismo: el dualismo se erige por contraposición al monismo, postulando dos principios o sustancias irreconciliables o por lo menos bien diferenciadas y que es de su interacción de la que emerge la realidad. Un ejemplo sería la teología cristiana donde tenemos una instancia trascendente, Dios, frente a otra inmanente, el mundo. [5] [6]

Una vez presentados estos dos puntos, será más sencillo comprender las distintas concepciones de la conciencia pues se basan en uno de los dos modelos.

2.1.3 Estatus elegido: monismo emergente

El punto de partida a nivel metafísico que voy a mantener en todo este trabajo es el de un monismo de tipo emergentista [6], en clara sintonía con las intuiciones biológicas y cibernéticas desarrolladas en la teoría de la autopoiesis[2] de Humberto Maturana y Francisco Varela [7]

Esto implica que considero que la conciencia es un epifenómeno o proceso emergente de un sustrato físico. Para evitar el reduccionismo biológico de otros autores [8], es decir, entender que la conciencia solo puede darse debido a las circunstancias y características particulares de la anatomía del cerebro, he optado por una concepción más plena. Esta permite contemplar la conciencia como resultado de un proceso de integración en sistemas de altísima complejidad ya sean biológicos o artificiales.

Además las intuiciones sobre la Autopoiesis tal y como las formuló Maturana [9], conciben que del fenómeno de la autoorganización y la relación con el entorno emerge la cognición como un proceso encaminado a facilitar la Autopoiesis y la relación con el entorno. Grosso modo la conciencia emerge de este proceso, como un fenómeno que integra temporalmente el flujo de

[1] **Inmanente**: propiedad que refiere a lo intrínseco de una entidad, por oposición a lo trascendente, que refiere a algo externo o ajeno a este.

[2] **Autopoiesis:** neologismo que significa auto-creación, acuñado por Ernesto Maturana en su teoría de sistemas vivos, para definir el proceso que lleva a cabo cualquier entidad viva. Según Maturana la vida emerge de una serie de reacciones de autoproducción, empezando por definir una clausura operacional que constituye una estructura determinada molecularmente y cuyas reacciones químicas están orientadas a seguir produciendo dicha barrera que diferencia la entidad del entorno. De esta forma, todas las unidades funcionales que contiene en su interior están orientadas a perpetuar esta estructura organizativa cerrada, a la vez que intercambian sustancias de desecho y material que les sirva como fuente de energía con su entorno. Esta idea, a continuación, se generaliza a toda entidad ya sea unicelular o pluricelular, en la que toda la estructura y entidades que la componen están orientadas a perpetuar la estructura del organismo viviente, como condición necesaria y suficiente de su supervivencia.

los procesos biológicos y favorece la interacción con el entorno. De la misma forma se puede postular esta capacidad de una entidad cibernética que procesase información y que tuviese varios sistemas interconectados dedicados a distintas funciones con un sustrato de flujos de energía que hubiese que corregir, estabilizar y readaptar en función del contexto y siempre enfocados a mantener el sistema en homeostasis.

Por otro lado, está visión está en mayor sintonía con el paradigma científico actual de la complejidad [10], que plantea entender el funcionamiento de las entidades desde una perspectiva sistémica holística y no reduccionista analítica

2.1.4 Justificación de la elección

Mentiría si dijese que esta es la única opción disponible como punto de partida, de hecho en la informática la visión más extendida es una variante del dualismo uno que divide entre lo corpóreo (hardware) y lo mental (software). Una concepción que, como mencionaba en la introducción es simplista y errónea, por utilizar una metáfora desbocada, al no atender a las diferencias ontológicas fundamentales que subyacen a la comparativa.

La mente entendida como un flujo de datos o información, plantea un problema metafísico de primer orden. Ya que al plantear el dominio mental como una entidad con autonomía ontológica, diferenciada de la materia, supone volver a la noción caduca del cartesianismo. Los mismos problemas que tuvo el sistema cartesiano mente/cuerpo vuelven como retorno de lo reprimido, pero esta vez sin un Dios que obre la integración entre ambas sustancias [11]. En la medida en que se postulan estas sustancias, encontramos el problema de que la propia teoría es incapaz de explicar cómo ambas interaccionan entre ellas: cómo lo mental afecta a lo físico y viceversa.

En la actualidad una de las teorías que más tracción ha ganado a y que mantiene el paradigma dualista, es la teoría de la conciencia del campo CEMI [12]. Esta propone equiparar la conciencia con el campo electromagnético que genera el cerebro. Su objetivo con esta teoría es explicar cómo se da la integración espacial de la información contenida en los procesos concurrentes pero espacialmente dispersos del cerebro. Aunque su teoría la plantee como un dualismo científico (materia-energia) huyendo de los planteamientos espiritualistas vistos con anterioridad, tiene un problema subyacente que la hace plantear una abstracción explicativa como una entidad con realidad ontológica[3]. Es debido a este motivo que rechazo también esta formulación, así como

[3] Entrando un poco más en detalle, esta abstracción explicativa de la que hablo es la información, entendida en el sentido de la teoría de la información de Shannon y sus desarrollos subsiguientes en los campos de la cibernética y la física. El problema que encuentro con ella es que está planteando la información como una entidad con realidad objetiva. No es mi intención refutar toda la física cuántica de los últimos 70 años, pero si que considero que hay que diferenciar el formalismo explicativo que utilizamos para describir la realidad de la realidad misma; y que cuando nos lanzamos a identificar la realidad con nuestras teorías estamos cayendo en una suerte de metafísica dogmática de tintes cientificistas, lo que Whitehead llamaba falacia de la concreción desplazada. Para resumir, mi crítica se basa en que al postular la información como una entidad con realidad ontológica, se está ignorando el ejercicio de categorización y reducción de la realidad misma a una serie de esquemas antropocéntricos que abstraen parte de esa misma realidad, para luego reproyectar dichas categorías filtradas por el intelecto humano y postularlas como reales y objetivas. Se privilegia una parte de la información contenida en los signos que estudiamos, algunas de sus propiedades medibles mientras, que otras quedan en suspenso o desatendidas. Lo que más adelante se llamará ser-a-la-vista, utilizando la terminología Heideggeriana.

las que plantean la conciencia como una entidad con realidad ontológica gracias a su propio campo cuántico [13], por caer en los mismos errores.

Debido a estas circunstancias encuentro la descripción elegida mucho más acertada, ya que no deja de ser la misma salida que Spinoza planteó al problema cartesiano[4]. La concepción que mantendré en este trabajo será de tipo emergente e integracionista. La conciencia es un proceso que emerge a partir de la interacción e integración de distintos sistemas de complejidad variable. Algunas de las características fundamentales de la conciencia, sobre todo las que han sido históricamente desatendidas en las formulaciones del siglo pasado, son las que pasaré a tratar a continuación.

2.2 Características de la conciencia

Una vez elegido el marco desde el que vamos a comprender la conciencia, es necesario ponerlo en relación con el concepto mismo de conciencia y algunas de las características que son indispensables que tenga. Este análisis se llevará a cabo utilizando algunas de las intuiciones sobre la conciencia desarrolladas por Martin Heidegger en Ser y Tiempo. En esta obra Heidegger intenta llevar a cabo una clarificación del sentido del ser, dicho de otra forma ¿Qué es el ser y en qué se diferencia del ente?

2.2.1 Planteamiento Fenomenológico

El punto de partida o método de estudio es fenomenológico, es decir que estudia los fenómenos y sensaciones tal y como se presentan directamente a la conciencia. Esto implica suspender el juicio de sentido común, o inmediato, que se da cuando recibimos cualquier experiencia o "input sensorial".

Un pequeño ejemplo, al ver una señal de STOP, automáticamente reconocemos su significado y la información que nos comunica. Esta circunstancia se da debido a que ha perdido su novedad, la percibiríamos de forma completamente distinta en caso de ser la primera vez que la vemos. El hábito de ver las cosas las sume en la banalidad. La metodología fenomenológica se propone penetrar más allá del velo de la costumbre, para intentar ofrecer un conocimiento más profundo de las cosas mismas.

Mi motivo para utilizar este enfoque es claro. Con Ser y Tiempo, Heidegger pretendía romper con una tradición excesivamente mentalista, tradición que podríamos retrotraer hasta Descartes con su genio maligno. Es un tipo de tradición que concebía la conciencia como algo que podía existir separado y disociado de su entorno. De ahí planteamientos como el solipsismo, o el genio maligno, en que la reflexión y la introyección en uno mismo posibilita el pensar que la realidad

[4] En la Ética, Spinoza plantea que a nivel lógico la existencia de dos sustancias distintas es absurdo, si ambas existen de forma simultanea ¿cómo se limitan y relacionan entre ellas? Para salir de este problema planteo un monismo en que solo existiría una única sustancia y que la variedad y riqueza de la realidad emergían de los distintos modos o formas que tomaba esa sustancia fundamental, una idea, a mi juicio bastante alineada con el paradigma físico y de la complejidad. Este estudia las relaciones y la emergencia de toda la realidad a partir de una sustancia primordial(metafísicamente hablando) que puede configurarse tanto en materia como en energía y realizar el tránsito de una a otra, así como el estudio de las propiedades emergentes de los sistemas, conforme van aumentando en su orden y nivel de complejidad

externa puede ser algo accidental y concebirse a uno mismo como una cosa pensante, que puede existir independientemente de un sustrato. Esto le lleva a romper con la fenomenología en su planteamiento husserliano[5] original, alejándose de la conciencia trascendental[6]

De toda la complejidad y riqueza de la obra, nos vamos a centrar exclusivamente en dos aspectos de los ahí discutidos. El dasein y los modos de percepción del mundo, por ser los que tienen una aplicación más directa a nuestra cuestión y también ser aspectos fundamentales.

2.2.2 El Dasein (o el estar en el mundo)

Dado que la obra de Heidegger parte del problema del ser, como de algo que se ha olvidado progresivamente desde los albores de la historia de la filosofía, siempre entendido como algo presente y dado en el mundo inmediato, Heidegger acuña el término dasein[7] para romper con esa tradición [14]. El término es un neologismo que se puede traducir al español por ***"ser-ahí"*** o **"estar-en-el-mundo"**, con él se recalca la necesidad de tener en cuenta la corporeización o estar situado en un punto determinado del mundo, como precondición y sustrato de toda experiencia humana. El dasein es por tanto una entidad que se ve arrojada al mundo en un determinado momento histórico y sociedad, idea que podemos equiparar a "la circunstancia" orteguiana.

Más importante aún, hay que tener en cuenta que el "*Estar*" como entidad está abierto a nuevas experiencias y formas de relacionarse con el mundo, lo que llama apertura óntica. Esta designa la capacidad para integrar nuevas formas de relación y actuación en su comportamiento y hábitos mentales. Además el "*Estar*" es siempre un proyecto en ambos sentidos del término. Por un lado, vive en un constante proyectarse tanto hacia el futuro, anticipando los acontecimientos por venir, como hacia al pasado codificado en su memoria. A su vez, también vive como un proyecto, una corriente vital que se relaciona con el mundo, que tiene metas y objetivos, que reviste al mundo de un valor y significado propio.

---

[5] **Edmund Husserl:** (1859-1938) fue un filósofo y matemático alemán, padre de la fenomenología. En su formulación original, la fenomenología era una continuación del proyecto kantiano de fundamentación del conocimiento, para ello Husserl planteó que la mejor forma de cimentarlo era investigando y reflexionando profundamente de cómo se aparecían los fenómenos a la conciencia para así descubrir una serie de condiciones subyacentes y universales del conocimiento. Este planteamiento es fundamentalmente solipsista y como el mismo Husserl comprobó al final de su carrera, el planteamiento fenomenológico se encontraba con una paradoja irresoluble al intentar demostrar y fundamentar lógicamente la existencia del mundo externo.

[6] **Conciencia Trascendental**: la conciencia trascendental es el punto de partida de la fenomenología husserliana, una continuación del planteamiento kantiano y su sujeto trascendental. En la crítica de la razón pura, Kant definía un sujeto trascendental, que no trascendente, entendido este, como el sujeto constituido por una serie de estructuras que el llamaría condiciones de posibilidad a priori del conocimiento, que son comunes a todas las personas (de ahí el trascendental) y que condicionan y establecen las posibilidades de conocer y experimentar el mundo y la realidad. De la misma forma Husserl continua con esta idea, pero reduciendo al sujeto trascendental a una conciencia trascendental y analizando de forma similar sus condiciones de posibilidad y estructuras inmanentes. Esta entidad se concibe como existente por si misma y sin necesidad de integración en el mismo mundo que está experimentando y que le sirve como fuente y sustrato de experiencias que percibe.

[7] **Dasein**: Este término al finalizar el párrafo dejará de utilizarse, para facilitar la lectura, pasará a utilizarse el término **"*Estar*"**, cómo hizo Xavier Zubiri al desarrollar su filosofía heredera de la de Heidegger y adaptando parte de su terminología al español. Además si se atiende a la descripción del concepto dada en el párrafo, podemos apreciar como "***Estar*"** es la expresión que mejor conjuga legibilidad con fidelidad al término original.

2.2.3 La apertura óntica (o del ente)

La idea de la apertura óntica es indispensable para el análisis posterior de la IA, ya que es precisamente este uno de los conceptos claves para poder dar cuenta del fracaso histórico de este campo. Esta no ha de ser entendida como libertad o capacidad de autodeterminación, aunque estén íntimamente relacionados con ella. La apertura óntica se refiere a la maleabilidad e indeterminación fundamental de nuestras estructuras cognitivas; siempre en constante cambio. Esto nos permite que nos relacionemos de forma novedosa con nuestro entorno y nos adaptemos. Al mismo tiempo es gracias a esta apertura e indeterminación constitutiva que el "Estar" puede evolucionar y tomar como suyos aspectos de la experiencia.

Un ejemplo con el que clarificar esta idea: imaginemos que tenemos un bebé en una habitación cerrada, con muebles y otros aparatos. Aunque en un principio pueda verse desorientado o no haga nada, llegado cierto punto comienza a explorarla, a relacionarse con la habitación, incluso en función de las cosas que encuentre, a llevárselas a la boca o a jugar con ellas. Este comportamiento podríamos considerarlo como un ejemplo básico de la apertura.

Para resumir, el concepto del "*Estar*", así como la estructura a la que remite, se pueden entender como el ente constituido espacio-temporalmente en un lugar y momento determinado, en un contexto o circunstancia histórica a la que se incorpora *in media res* [14].

Esta circunstancia histórica se constituye en la experiencia del "*Estar*, como horizonte de comprensión del mundo, el fondo sobre el que se nos presentan las cosas y a partir del cual adquieren su significado. Así queda recalcada la necesidad de entenderlo como un ente sujeto a cambios y en todo su devenir, así como en la relación que establece consigo mismo ante su devenir –ya sea para mantenerlo así o para corregir el rumbo.

2.2.4 Los modos de percepción (o relación)

En ser y tiempo Heidegger describe las 2 formas de percepción del mundo así como la jerarquía que existe entre ellas. A estos modos los llama el ***ser-a-la-vista*** [*Vorhandenheit*] y el ***ser-a-la-mano*** [*Zuhandenheit*]. [14] [15]

2.2.4.1 El ser-a-la-vista

Es como se refiere a la forma de conocimiento que ha permeado a todas las ciencias teóricas. Este es el modo de pensar y saber que tomamos cómo básico. El ser-a-la-vista está centrado en la determinación de las características de un objeto o cosa aislado, es el modo de ver que pretende ser objetivo y abstraer del objeto los hechos, sin tener en consideración el valor subjetivo, la historia o utilidad del objeto en cuestión. [15]

Un ejemplo de esto podría ser un bosque, verlo desde el punto de vista del ser-a-la-vista sería equivalente a verlo en sus características objetivas. Las formas de los árboles, la cantidad de ellos, su composición química, tamaño, extensión, cualidades todas ellas puramente objetuales. El ser-a-la-vista reduce el bosque a una totalidad objetual, a una cosa de la que se puede saber, a la vez que se puede movilizar, como recurso. A consecuencia de ello, los aspectos estéticos, la

dimensión sentimental y también la moral y ética, todas ellas subjetivas desaparecen por completo.

Dado que el ser-a-la-vista constituye una mirada reificadora[8] a la que se le escapa parte de la realidad que aprehende, Heidegger propone el modo opuesto como forma más primordial de percepción y relación con el mundo.

2.2.4.2 El ser-a-la-mano

Antes de centrarnos en este concepto es necesario pararse en la idea del mundo como totalidad instrumental y de significado, con ella Heidegger refiere a la experiencia fenoménica del mundo como un todo constituido por un sistema de objetos y signos que se refieren entre ellos. Por ejemplo, mientras escribo estas líneas tengo frente a mí el escritorio con sus botes de lápices, bolígrafos y demás material, así como una serie de estanterías repletas de libros. A no ser que fije la atención sobre uno de estos elementos (que sería acercarse al ser-a-la-vista), el modo de presentarse de mi entorno es como un ambiente, un fondo constituido por diferentes elementos, pero concebido en cohesión. A su vez los objetos que pudiese llegar a percibir en él no los percibo como cosas aisladas, sino que se hallan inmerso en una red de relaciones en las que unos refieren a otros. De ahí la idea del mundo como totalidad instrumental. [15]

Por contra podemos apreciar que en los sistemas de visión por computador más modernos, el tipo de procesamiento de la realidad es el ser-a-la-vista en toda su plenitud, de la totalidad de la imagen se discretiza por completo todo elemento diferenciable para poder seguirlo. El trasfondo y el ambiente se pierden por completo, solo existe lo visto y lo no visto, aquello que se ha podido discretizar y el fondo ambiente que se resiste a él.

El ser-a-la-mano es la forma primordial de relacionarse con el mundo en oposición al ser-a-la-vista. El ser-a-la-mano es la forma prerreflexiva, no teorética, de experimentar y concebir las cosas que se nos dan en el mundo. Es una forma instrumental, en la que los objetos son concebidos en su relación con nosotros, tanto por su finalidad, como por su uso. Las cosas a la mano, se retraen de su aparición visible y nos refieren exclusivamente a su dimensión útil y a una serie de cualidades y relaciones que desbordan lo visible.

Para intentar acercar esta exposición a una idea más intuitiva tomemos el ejemplo del martillo. En mi relación con el martillo, lo utilizó como medio para un fin. Mi relación con él no es contemplativa/reflexiva. No me interrogo por su funcionamiento y constitución, sino que mi relación es puramente práctica e intuitiva. Utilizo el martillo como herramienta para un proyecto que estoy construyendo, como una extensión de mí. Mi saber y conocimiento del mismo sólo remite a esa dimensión práctica en la que se moviliza y pone en marcha.

[8] **Reificar:** literalmente convertir en cosa o cosificar, es una forma de concebir la realidad en que se reduce su riqueza a una serie de características determinadas, en este caso las objetivas o formales, o a un proceso en que características accidentales fruto de una dinámica histórica o social determinada, se conciben como eternos e inmutables.

2.2.5 Conclusión e ideas extraídas

Una vez expuestas estas cuestiones considero necesario terminar de clarificar la relación de este marco con nuestra cuestión. Históricamente la IA ha sido concebida como una inteligencia carente de temporalidad, como en una forma de comprender el mundo puramente a-la-vista.

Esta inteligencia carente de temporalidad implica que desaparece de la ecuación una de las dimensiones básicas de la experiencia humana, la experiencia temporal y extática en que estamos en relación con el pasado, como sustrato de experiencias, y el futuro como multiplicidad de posibilidades a las que tendemos o podemos llegar. Dado que todo nuestro comportamiento está informado y determinado por esta estructura, al no tenerlo en cuenta o hacer un uso muy libre del término memoria, el objetivo de una inteligencia real, se plantea, a mi juicio, inalcanzable.

Por otra parte el modo de ser-a-la-vista que ha permeado a toda el campo, herencia del contexto en que se desarrolló[9], supone que precisamente, **todas aquellas cuestiones que son de sentido común y triviales para nosotros, por pertenecer a su dimensión prerreflexiva, son prácticamente incomunicables o incodificables a lenguaje máquina.** En la medida en que no se intenté crear una IA con la capacidad de poder relacionarse con el mundo de una forma más cercana a la aquí descrita, el objetivo último de la IA y que pretendía Turing, a saber, la creación de máquinas pensantes quedará siempre inconcluso.

2.3 La Inteligencia y el Conocimiento

Una vez definida la conciencia y algunas de sus características esenciales, podemos proceder a tratar el tema de la inteligencia. El primer paso para poder continuar, ya que este trabajo se presenta como una investigación sobre la IA, convendría establecer el concepto de inteligencia. Para ello partiremos de la definición de John Searle de los tipos de IA [16], para luego desarrollar nuestras propias intuiciones.

2.3.1 Inteligencia Artificial Fuerte y Débil

Podemos trazar la genealogía de estos conceptos al artículo de John Searle *Minds Brains and Programs*, aunque cabe la pena recalcar que la definición y motivo que tiene el autor para distinguir entre ambos no es la misma que utilizamos actualmente[10]. Con los años su definición ha evolucionado a la siguiente:

[9] Esto se mostrará más adelante en el apartado 3.1

[10] En este artículo Searle lleva a cabo esta distinción y análisis debido a que en el contexto intelectual de su época, existía una corriente bastante numerosa de seguidores de la idea de que un ordenador con los programas adecuados era equivalente a una mente. De hecho había algunos que llegaban tan lejos como para decir que programas como el simulador conversacional ELIZA, eran equivalentes a una mente, y que su análisis podía servir como herramienta para entender la mente humana. Para ello distingue entre dos hipótesis la IA débil que dice que un programa simula un aspecto de la mente humana, por contraposición a la IA fuerte que afirma que el programa adecuado es la mente misma y ya no solo nos puede servir para comprobar nuestras explicaciones, sino que el programa puede ser la explicación misma.

- Inteligencia Artificial Débil: es un tipo de inteligencia específica, reconocimiento de imágenes, por ejemplo. Un programa capaz de llevar a cabo una tarea específica o proceso cognitivo humano, con el mismo o mayor nivel de maestría.
- Inteligencia Artificial Fuerte: es un tipo de inteligencia general y multipropósito, capaz de llevar a cabo tareas y funciones en múltiples dominios. En algunas circunstancias también se le añaden a esta atributos humanos como la conciencia. Además, también se considera en esta definición la posibilidad de una superinteligencia que supere nuestras capacidades en todos los aspectos.

Una vez dada esta definición, conviene recalcar, que el tipo de inteligencia a la que me voy a referir durante el resto de este trabajo es a una IA fuerte a la que se le añade, por lo menos, la conciencia. La justificación para ello es la siguiente: aunque este trabajo no pretende antropologizar sus términos, ya que nos puede llevar a establecer un criterio demasiado reduccionista y miope para abordar la cuestión; considero que, sin este requisito, todo uso de la palabra inteligencia va tener un uso análogo, metafórico o funcional, pero sin la significación y propiedades que esperamos de ella. Para poder justificar plenamente esta idea necesitaré primero exponer las limitaciones que tiene el concepto formal de inteligencia utilizado en nuestro campo.

2.3.2 La Inteligencia como concepto y sus limitaciones

Partamos primero del concepto de inteligencia formalizado por Wang [2]: *"La inteligencia es la capacidad de un sistema que procesa información de adaptarse a su entorno, mientras opera con conocimiento y recursos insuficientes"*. Esta definición es correcta si la aplicamos a un organismo autopoyético [9](o vivo) como puede ser una persona o un animal, pero cuando lo aplicamos a una IA o una máquina muestra sus limitaciones. Cumple el aspecto de procesamiento de información, y si somos muy laxos con el término entorno, hasta podemos aplicarlo. Pero el uso del término conocimiento es de un carácter bastante dudoso.

Podemos decir que tenemos sistemas expertos o IAs que conocen tal o cual información, pero el uso de ese término es puramente metafórico o análogo al nuestro. Un modelo del lenguaje como GPT-3 [11]es capaz de operar en nuestro idioma, es capaz de responder a entradas de texto con una respuesta cercana a la que podríamos desear, pero no conoce realmente el lenguaje. Posee una representación del lenguaje codificada en un espacio vectorial de alta dimensionalidad y las palabras son vectores n-dimensionales. Sus términos están definidos como relaciones diferenciales entre ellos, no tienen una representación asociada. Son símbolos que se manejan conforme a una serie de reglas ya dadas, en este caso la probabilidad de que a uno le siga otro, sin conocer qué significan.

[11] GPT-3 es uno de los modelos del lenguaje más potente hasta la fecha diseñado por OpenAI en 2020, un modelo con 175.000.000.000 de parámetros, capaz de, entre otros logros, ser capaz de simular una conversación impersonando a sus participantes con un grado de acierto muy alto, traducir texto de un lenguaje a otro sin haber sido entrenado para ello y hasta de escribir artículos para the guardian [61]

2.3.2.1La necesidad de la conciencia en el conocimiento

El conocimiento entendido como una base de datos en las que se almacena información y hechos de la realidad, ignora su dimensión productiva, la capacidad que tiene para ir más allá, para desbordar su dominio de aplicación y utilizarse de forma novedosa.

Si antes he hecho mención a la conciencia como condición necesaria para la inteligencia es porque si no la tenemos en cuenta hay una dimensión del conocimiento que estamos perdiendo. Me refiero a la capacidad reflexiva. La capacidad de ser conscientes de nosotros mismos, la capacidad de ser conscientes de que somos conscientes, de poder iniciar una recursión infinita tipo "yo sé esto, yo sé que sé esto, yo sé que sé que sé esto, …" y a su vez saber que es un proceso que no tiene fin y detenerlo. De la misma forma si por ejemplo intento hacer un ejercicio de introspección y conocerme mejor a mí mismo, soy capaz de reconocer que lo que estoy conociendo también me está afectando, es decir que el yo que conoce es distinto del yo conocido, que por medio de este conocimiento estoy cambiando mi propio conocimiento anterior y me estoy "desconociendo" en el proceso. Es decir, que toda forma de autoconocimiento es una forma de hetero-conocimiento [17].

A su vez por medio del conocimiento puedo ser consciente de su propia falibilidad, puedo saber que el lenguaje es una herramienta que no es totalmente fiable. No tiene ningún tipo de base en sí misma, ningún pilar sobre el que establecer el significado que tienen los términos que uso. Si intento clarificar el significado de una palabra, me encuentro con que la definición depende de otras palabras que tendría que clarificar a su vez, entrando en otra recursión infinita. Tengo un conocimiento que es capaz de reconocer sus propias limitaciones, de apreciar aquellos aspectos en los que desborda sus capacidades para ofrecer certezas.

Ante estas ideas podría parecer que estoy cayendo en lo mismo que pretendía evitar, es decir, en la antropologización de la inteligencia y el conocimiento, pero no es el caso. Como decía anteriormente, la definición aportada era correcta para organismos autopoyéticos, es decir vivos. Veámoslo con un ejemplo: las águilas y otras aves grandes cuando cazan tortugas, acostumbran a dejarlas caer desde gran altura y contra rocas para que la caída las mate y luego poder devorarlas más fácilmente. Este comportamiento es inteligente, se puede aplicar sobre la definición dada y la cumpliría por entero. El águila (sistema que procesa la información), sabe que las cosas caen y acabar con su presa de la forma usual le llevaría mucho tiempo (tiene un conocimiento y recursos limitados) y aplica este conocimiento para cazar de forma más efectiva (se adapta a su entorno).

Se puede reconocer que el águila por ser una entidad que procesa la información distinta y la evolución de sus capacidades han sido otras, no tiene una inteligencia que se pueda equiparar a la humana, pero es inteligente. Y aunque no podamos decir nada de cómo es su conciencia y su experiencia fenoménica, sabemos que la tiene, porque reconocemos que tiene una capacidad de relacionarse con su entorno y es consciente de él. Del águila y cualquier otro animal, podemos decir que tiene intencionalidad[12]

[12] **Intencionalidad:** término de la filosofía de la mente referido a la consciencia y el contenido de sus estados mentales. Se habla de un estado mental intencional o de una actitud intencional hacia algo en la

Por estos motivos, la inteligencia como concepto necesita ser tenida en cuenta en relación con su aspecto consciente. En la medida en que el objetivo es crear una IA fuerte o incluso una superinteligencia, este concepto es de vital importancia, porque considero que esas capacidades que no se pueden exigir a un águila u otro animal, si son exigibles en una entidad que tenga una inteligencia similar o superior a la nuestra. Cómo lograr este objetivo es algo que desborda el tema de este trabajo.

2.3.3 Conocimiento y representación

En la exposición del apartado anterior, he mostrado algunas de las limitaciones que se encuentra la inteligencia como concepto al aplicarse sin tener en cuenta a la conciencia. Pero para llevar a cabo una exposición satisfactoria considero necesario una exposición final del conocimiento mismo y sus limitaciones. La necesidad de esta se hará autoevidente una vez se desarrolle.

El conocimiento tal y como se ha concebido tradicionalmente es la capacidad de asociar a un término una representación. Dicho en términos aristotélicos, el conocimiento es la adecuación del intelecto a la cosa, entendiendo así que en esta adecuación reconocemos en ella las propiedades que tiene y que deja de tener. Conocemos una botella porque somos capaces de representarnos sus propiedades: tiene un volumen, unos materiales, un contenido o una ausencia de él, unos colores, etc. A su vez también somos capaces de representarnos qué es lo que no tiene o no es.

La mención a Aristóteles no es casual, ya que proyectos actuales [18] siguen partiendo de su ontología para definir formalizaciones del conocimiento y la inteligencia. Todas ellas tienen en común un planteamiento en el que los conceptos son representaciones, abstracciones que pretenden reunir sus propiedades esenciales, para así poder operar con ellas.

Desde mi punto de vista, este planteamiento es igual de problemático que partir de una concepción dualista (mundo real vs mundo ideal), o incluso más grave si cabe. Esto se debe a que todo conocimiento basado en representaciones cae en algún tipo de lógica esencialista[13]. ¿Pero por qué esto es algo problemático?

2.3.3.1 La necesidad de la diferencia para el conocimiento

Antes de empezar, destacar que este es un apartado muy fuertemente inspirado por las ideas de Guilles Deleuze [14] y su obra Mil Mesetas [19]. El objetivo de este apartado es llevar a cabo una crítica al concepto clásico de representación que ha permeado a nuestro campo, para así señalar un camino alternativo a partir de las intuiciones de esta obra.

---

medida en que refiere a un elemento del mundo externo y muestra una relación entre el ente y su posición hacia el objeto de intención.

[13] **Lógica esencialista:** es una forma de pensamiento en que se intenta buscar el fundamento último o cualidades esenciales de un objeto ya sea real o teórico, para su posterior estudio.

[14] **Gilles Deleuze:** fue un filósofo del siglo XX perteneciente al posestructuralismo francés. Su obra se centra de forma central en el concepto de la diferencia concebida en si misma y el pensamiento creativo, parte de este apartado está inspirado en su pensamiento.

En primer lugar la abstracción conceptual que se lleva a cabo en estas representaciones lo que hace es plantear realidades inmutables, esencias inalterables que se mantienen estables. Pero no es así como funciona la realidad. Cuando se busca la esencia de algo, se reúnen algunas de las propiedades que tiene y se utilizan para definirlo. Esta forma de concebir las cosas es útil, pero no tiene en cuenta la diferencia, tanto entre distintas instancias de una misma cosa, como de la cosa misma en su dimensión temporal. Carece de lo que podríamos llamar el conocimiento de la diferencia.

El problema es que este tipo de conceptos son siempre rígidos, no reconocen la capacidad que tiene algo de cambiar. Recordemos esa intuición clave de Hegel de que toda cosa, además de relacionarse con las demás mantiene una relación de autocontradicción, de tal forma que la relación de identidad de algo consigo mismo A=A, con el tiempo acaba derivando en A=B [17]. Dicho de otra forma que las propiedades que se establecen para definir la esencia de algo son accidentales y tiene la capacidad de mudar. Es decir que las representaciones que se llevan a cabo de los conceptos se hacen a partir de propiedades que pueden no ser esenciales. Estas conceptualizaciones carecen de temporalidad.

Por si mismo esto no es algo insalvable, podemos definir conceptos y representaciones teniendo en cuenta su capacidad de evolución. Si el problema es que nuestra definición de árbol es que es una entidad con tronco leñoso, ramas y hojas, pero en invierno las pierde, es tan fácil como reconocer que tiene la capacidad de perderlas y renovarlas, y ya tendríamos una definición más completa. Pero hay un segundo problema que es el que realmente encorseta el pensamiento derivado de las representaciones.

Este consiste en que las representaciones se presentan como modelos ideales de las cosas, como lo verdadero a lo que sus diferentes instanciaciones tienen que acercarse y se define su grado de corrección por su aproximación a esa representación idealizada. El problema que se deriva de esta concepción es que concibe la diferencia como algo negativo, no acercarse a ese algo, lo hace menos algo. Pero no ser algo, también le permite ser algo distinto. Un ejemplo sencillo, una botella agujereada múltiples veces, ve mermada su "botelleidad" por no tener la capacidad de almacenar agua, pero a su vez, puede convertirse en un colador.

La representación como molde y modelo absoluto nos lleva a constreñir el pensamiento y lo que podemos llegar a saber. Por ello considero necesario que el conocimiento sea concebido como un conocimiento que parta del conocimiento de la diferencia y de que per se no es algo negativo, sino que ofrece siempre nuevas posibilidades y puntos de fuga a partir de los cuales aprovechar para desarrollar el conocimiento. No como algo que se enfrenta al conocimiento y pretende negarlo al resistirse a sus representaciones, sino mostrando la falibilidad de dichas representaciones, su carácter limitado y mutable.

El conocimiento representacional es a lo que Deleuze llamaba "pensamiento arbóreo" [19], una forma de conocimiento jerárquico que busca principios últimos de las cosas y representaciones. A esto oponía lo que llamaba el "pensamiento rizomático[15]", no jerárquico y

[15] **Rizoma:** en botánica es un tipo de raíz que se desarrolla lateralmente sin un centro nuclear, desarrollando diferentes ramificaciones de las que pueden brotar otras, dando como lugar a un crecimiento horizontal. En Mil Mesetas, Deleuze y Guattari proponen la idea del rizoma, como una forma de conceptualizar la realidad

descentralizado (se podría equiparar a un grafo). Una forma de conocimiento en que se reconoce el agregado de características y elementos que contiene un concepto, las relaciones que establecen entre ellos, así como su capacidad de evolucionar, de perder algunos y ganar otros. Esta forma de conceptualización tiene la posibilidad de utilizarse como modelo para un conocimiento y pensamiento creativo.

Si señalo con tanto énfasis este aspecto es porque considero que en caso de poder implementarlo como una nueva forma de conceptualizar y modelizar la realidad, uno de las limitaciones más importantes que tienen los modelos hasta la fecha desaparecería. De hecho, si volvemos al ejemplo que se comentaba en secciones anteriores de un modelo del lenguaje, podemos ver que en la práctica su modelo ya es en parte rizomático, una representación no jerárquica de términos interrelacionados. De esta forma se podría añadir el aspecto temporal a la estructura de la que carece hasta la fecha y hacer que pudiese añadir nuevos términos o alterar los ya dados, para representar una evolución de los significados. Aunque esto seguiría sin soslayar los problemas planteados anteriormente, considero que este es el camino que debe seguir la representación del conocimiento en general como paso para lograr una conciencia artificial.

La idea fundamental es que estamos buscando una forma de conceptualización que reconozca tanto su actualidad (lo que es), como su potencialidad (lo que podría ser). Pasar de las entidades estables y fijas a los agenciamientos o ensamblajes en que las cosas están compuestas de muchos elementos reunidos en una configuración determinada, siempre sujetos a cambio y diferenciación. Es decir un concepto no esencialista de la esencia de las cosas.

2.4 Conclusión y Resumen

Después de esta exposición de conceptos considero necesaria una unificación de ellos para clarificar cual es la relación y cómo va a afectar al desarrollo posterior del trabajo. El marco total desarrollado en este trabajo consistiría en el siguiente:

Partimos de una concepción de la conciencia emergente y monista, pero sin reducirla a la biología. Esto es importante tenerlo en cuenta pues hace posible que se pueda crear una conciencia artificial. Como requisitos fundamentales que ha de tener esta conciencia es su temporalidad, el conocimiento de su condición finita y a su vez su capacidad de retrotraerse (recordar) y proyectarse hacia el futuro (anticiparse). Esta conciencia ha de ir más allá del ser-a-la-vista, su dimensión puramente teórica y reflexiva, ha de tener un conocimiento práctico de las cosas, más allá de su condición de objeto de reflexión. De esta forma se aumenta su posibilidad de integrar y concebir el mundo.

A su vez en lo referente a la inteligencia, se ha destacado la necesidad de una conciencia como precondición para toda inteligencia real. La representación formal de conceptos sin una intuición de los mismos supone una limitación intrínseca a las capacidades para desenvolverse tanto temporalmente como para crear nuevo conocimiento y poder actualizar el ya poseído. Además hay que destacar que esto implica que la inteligencia no es un proceso que se pueda formalizar en términos puramente algorítmicos o mecánicos.

---

y sus elementos que destaca la interconexión entre ellos, la posibilidad de rupturas en sus conexiones sin que el conjunto se vea afectado y su capacidad para seguir creciendo y desarrollándose de forma novedosa.

En lo que respecta al conocimiento mismo se ha destacado la necesidad de partir de un conocimiento o forma de conocer las cosas que parta de la diferencia. Esta intuición es fundamental, es este conocimiento el que capacita tanto a la conciencia como al resto de estructuras para poder actualizarse y adaptarse a los cambios y evoluciones de su entorno, de una forma satisfactoria.

La intuición fundamental que tenemos que extraer de este apartado es el modelo del rizoma, como esquema ontológico y epistemológico. Dado que el rizoma es una estructura que en la práctica ya existe casi realizada en su configuración actual de espacio vectorial de *n* dimensiones, a la que solo falta añadirle el componente temporal. Con este sería posible reflejar la creación de nuevos conceptos o relaciones, o actualizar los conceptos y significados ya tenidos a medida que recibe nueva información.

Para concluir, hay un aspecto que no se ha parado de mencionar, la conciencia, que se ha detallado y pormenorizado en muchos de sus aspectos, pero del que nunca se ha dado una definición formal explicita. Esta definición la obtenemos sintetizando varias de las intuiciones descritas anteriormente. La entenderemos como: la facultad de correlación sucesiva de instantes temporales consecutivos ininterrumpidos. Con esta definición estamos capturando la intuición de la conciencia en su dimensión procesual y continua, pero sin referir a terminología como experiencia que ya presuponen una conciencia. Entenderla de esta forma es de una importancia capital, ya que de esta definición emerge lo que podríamos llamar "la vida interior".

De la misma forma con esta definición estamos recalcando un aspecto, a mi juicio, fundamental a la hora de adscribir conciencia a cualquier entidad digital: que el proceso "cognitivo" de dicha entidad sea continuo o ininterrumpido. Esto lo podemos ver en acción aplicado a los humanos. En todo momento de vigilia estamos conscientes, aun si nos encontramos en stand by o en un estado de baja actividad. No tener una tarea o un objetivo en mente no supone el fin de nuestra experiencia subjetiva. Si aplicamos esta idea a una conciencia artificial, esto supone que aunque fuese una máquina con un objetivo a llevar a cabo, por ejemplo un chatbot que está manteniendo una conversación con un usuario, en aquellos momentos en los que el usuario está pensando una respuesta o antes o después de la conversación, la máquina no podría encontrarse en un estado de pasividad absoluta.

Si se encuentra reducida a un estado vegetativo en el que el procesado de información solo se da en el momento de recibir una entrada de texto y en el procesado para devolver una respuesta, por muy sofisticada que sean sus respuestas, eso máquina no tendría conciencia ni tampoco experiencia subjetiva. Siendo la precondición para estas el acto de procesamiento continuo de la realidad.

Todas estas ideas se emplearán en el último apartado para poder llevar a cabo propuestas alternativas que superen el marco conceptual y las limitaciones heredadas por la propuesta de Turing. De la misma forma, será la postura que mantendré en el próximo apartado al confrontar algunos de los contraargumentos sobre la posibilidad de una IA fuerte.

3

IA: Contexto filosófico, desarrollo histórico y argumentos en contra

En este apartado se analizarán tres cuestiones. El contexto intelectual y filosófico de Alan Turing. Un pequeño resumen de la historia de la IA tras la intervención de Turing. Algunas de las críticas y argumentos en contra que se han hecho de la posibilidad de una Inteligencia Artificial Fuerte, señalando siempre que sea posible si dichas críticas suponen una imposibilidad, un obstáculo, o incluso un argumento espurio. Por otro lado considero necesaria llevar a cabo esta revisión del contexto del pensamiento de Turing por dos motivos:

- El primero, es que es útil para comprender y poder enmarcar mejor su artículo cuando se comente en el siguiente apartado, ya que hay determinadas afirmaciones que no se entienden si no se ponen en relación con este contexto.

- El segundo es que, dado que Turing, junto a Von Neumann, se pueden considerar padres de la informática, es fundamental señalar las ideas que circulaban en su época y por las que se vieron influidos para comprender de qué concepción partían. De la misma forma solo si se entiende esta concepción y sus limitaciones, podemos entender cómo se fraguó el marco reduccionista que se asentó en la base del campo de la IA; del que ha costado un largo tiempo empezar a deshacerse y complementar con otras visiones.

3.1 Contexto Intelectual de Alan Turing

Si hay algo que cabe la pena destacar antes de comenzar esta exposición es que no existe bibliografía especializada del tema. Se han publicado múltiples obras analizando las contribuciones de Turing a la lógica, las matemáticas o la computación, pero no hay ninguna que analice específicamente cuales fueron sus influencias y pensamiento en general. Esto nos lleva a la necesidad de reconocer la parcialidad de esta exposición, en la que solo me centraré en grandes corrientes de pensamiento o autores muy importantes, pero en pocas circunstancias se podrá trazar una relación indudable entre ambos.

Si hubiese que encuadrar a Turing dentro de una las grandes escuelas de pensamiento del siglo XX, aunque nunca llegase a identificarse con ninguna de ellas, esta sería sin duda la filosofía analítica.

3.1.1 Filosofía Analítica

La filosofía analítica es una de las dos grandes corrientes filosóficas que han existido durante el siglo pasado. Si seguimos la genealogía planteada por Ernesto Castro [20], los inicios de la escuela analítica los podemos retrotraer a la Crítica de la Razón Pura de Kant, en la que lleva a cabo una profunda investigación sobre el conocimiento, los usos de la razón, la lógica y sus condiciones de posibilidad. Aunque esta genealogía no debería inducirnos a error, pues no es hasta mediados del siglo XX, cuando esta empieza a reivindicarse como una corriente propia y diferenciada por oposición a la escuela continental.

La escuela analítica tradicionalmente se ha centrado en cuestiones de teoría del conocimiento, análisis del lenguaje, filosofía de la mente, lógica, matemática y ciencias duras en general [21]. Con un marcado corte empirista y una clara voluntad de fundamentar el conocimiento, utilizando un lenguaje preciso y riguroso, heredero de la tradición científica[16].

La corriente analítica comienza a definirse a principios del siglo pasado con pensadores como Bertrand Russel, Alfred North Whitehead o Ludwig Wittgenstein. Solo podremos justificar la adscripción de Turing a esta corriente si atendemos a la especial relevancia de estos autores, sus proyectos y su influencia posterior tanto en las universidades británicas como en círculos de toda Europa.

A continuación vamos a exponer algunos de sus aspectos más importantes, junto con un breve desarrollo histórico, para comprender mejor, cuál era la visión general que tenía la filosofía analítica cuando Turing entró en contacto con ella.

3.1.1.1 Lógica y Matemática

Aunque la lógica como objeto de estudio riguroso lleva existiendo por lo menos desde Aristóteles, no es hasta mediados del siglo XIX que da el salto de la filosofía a las matemáticas y pasa a centrarse en un dominio mucho más limitado y preciso. Es con la publicación del tratado de George Boole *The Laws of Thought* [22], en que la lógica pasa del terreno discursivo y argumentativo a uno más formalizado y cercano a nuestra concepción actual. En esta investigación Boole, se propuso formalizar el pensamiento en función de predicados que se podían someter a operaciones algebraicas con las que determinar su corrección o veracidad. En sus últimos compases hasta llega a proponer que la misma moralidad podría reducirse a un mero calculo proposicional.

En parte la visión de Boole daría un salto cualitativo en manos del lógico y matemático Gotllob Frege. Este es conocido además de por sus contribuciones a la filosofía del lenguaje, por

[16] Cabe la pena recalcar como bien señala Castro, que esto no deja de ser la reconstrucción a posteriori e idealizada por la propia escuela. Si se atiende con más atención hay problemas comunes a ambas escuelas aunque debido a su enemistado se han destacado y afirmado más sus diferencias, que reconocido sus intereses comunes y puntos de intersección.

sus contribuciones a la matemática y la lógica. Al final de su carrera se propuso fundamentar las matemáticas en la lógica, dando lugar a lo que en filosofía de las matemáticas se conoce como el logicismo. El proyecto logicista, acabó influyendo a Russel y Whitehead, para intentar acometer una fundamentación[17] rigurosa de las matemáticas en cuanto a lógica. Este fue el punto de partida de los 3 volúmenes de los *Principia Mathematica* en los que intentaron llevar a cabo esta tarea. Pero en su desarrollo se encontraron con una serie de paradojas irresolubles, como la conocida paradoja de Russel[18].

Las soluciones que ofrecieron Russel y Whitehead, no terminaron de satisfacer a gran parte de los matemáticos de su época que no vieron en su proyecto una fundamentación exitosa. Al mismo tiempo, David Hilbert importante matemático de finales del XIX y principios del XX, había ganado bastante notoriedad con su proyecto formalista. Este en oposición a Frege llevó a cabo una fundamentación de las matemáticas mediante su axiomatización. De esta forma logró formalizar la geometría a finales del XIX. Dando lugar a lo que se conocería como el formalismo.

El formalismo como fundamentación de las matemáticas proponía *grosso modo* que las matemáticas debían ser entendidas como un conjunto de símbolos regidos por una serie de estrictas reglas sintácticas con las que poder manipularlos y desarrollar nuevos teoremas a partir de ellos. Algo sería verdadero si se podía derivar de los axiomas. En 1920 presentó su programa en el que planteaba una serie de problemas que los matemáticos debían de resolver para poder terminar de formalizar todo el cuerpo teórico de las matemáticas. Entre ellos estaba la demostración de su completitud, la demostración de la decidibilidad de todos los teoremas en tiempo finito (*Entscheidungsproblem)* y la de formalización del concepto de proceso mecánico.

Son precisamente estos problemas, el de la parada y la formalización de proceso mecánico, los que ataca Turing con sus máquinas, consiguiendo mostrar la imposibilidad de determinar en tiempo finito la decidibilidad de un teorema (problema de la Parada). Del mismo modo el problema de la completitud es el que resuelve Kurt Gödel con sus teoremas de la incompletitud[19], mostrando una limitación insoslayable al proyecto formalista de Hilbert, existían proposiciones

[17] Durante el principio del siglo XX, los matemáticos de la época estaban inmersos en la búsqueda de unos fundamentos, primeros principios sobre los que poder asentar el edificio de las matemáticas. Sin este fundamento, la verdad y la corrección de las matemáticas quedaba en entredicho, debido que servían de prueba circular de su propia validez. Otros proyectos de fundamentación fueron el constructivista o el formalista, que se comentará más adelante.

[18] **Paradoja de Russel**: si tenemos el conjunto al que pertenecen todos los conjuntos que no son miembros de sí mismos, ¿Pertenece este conjunto a sí mismo? La paradoja se da en el hecho de que si pertenece a si mismo, entonces ya no sería un conjunto que no pertenece a sí mismo, pero si no pertenece a si mismo, debería pertenecer a sí mismo, porque es un conjunto que no es miembro de si mismo. Esto nos lleva a un estado de indecidibilidad y esta paradoja es irresoluble.

[19] **Los teoremas de incompletitud** suponen un golpe devastador al proyecto formalista. Esto se debe a que en trabajos anteriores, Gödel había mostrado que toda la matemática se podía reducir a lógica, pero que su consistencia seguía dependiendo de que la aritmética de Peano como sistema axiomático fuese consistente (no pudiese demostrar como verdaderas proposiciones falsas) y completa. Con sus teoremas de incompletitud muestra lo siguiente: El primero demuestra que en todo sistema axiomático lo suficientemente expresivo como para poder contener la aritmética existían proposiciones indecidibles. El segundo demuestra que un sistema axiomático de este tipo, capaz de demostrar su propia consistencia, es inconsistente. Con estos teoremas Gödel mostró que las matemáticas nunca podrían estar completas, señalando la imposibilidad de la fundamentación total que proponía Hilbert en su programa formalista.

indecidibles dentro de la Aritmética de Peano[20], y como tal su sueño quedaba tocado de muerte, cuando no hundido.

3.1.1.2 El Círculo de Viena

La figura de Gödel, uno de los lógicos más importantes del siglo XX, nos da pie a hablar de la escuela de la que formó parte, el círculo de Viena. Este fue un grupo de pensadores, físicos, matemáticos, lógicos, organizados alrededor de la figura de Moritz Schlick. Con el tiempo pasaron a conocerse como los positivistas lógicos. Sus reflexiones versaban sobre la naturaleza y criterios del conocimiento, así como sobre distintas ramas de la matemática, la lógica y la incipiente filosofía del lenguaje.

Su escuela nace a raíz de la publicación del *Tractatus logico-philosophicus* [23] de Ludwig Wittgenstein. En este se problematiza el uso del lenguaje y se propone que la corrección y verdad de cualquier enunciado viene dado por los propios términos que lo componen. Wittgenstein, heredero de la filosofía de Russel y de Frege, formuló una ontología basada exclusivamente en hechos y proposiciones (atomismo lógico) sobre los mismos que podían ser verdad o mentira. Para apoyar su exposición empleó las tablas de verdad y está acreditado sino como su inventor (existen antecedentes en Russel o C.S. Peirce de estructuras similares) al menos como quien las popularizó.

Los pensadores de Viena, influidos por esta obra, pero a la vez críticos con algunos de sus aspectos se propusieron refinar y revisar los aspectos más discutibles de su obra. Por ejemplo su teoría del significado o su atomismo lógico. Fruto de estas reflexiones Rudolf Carnap[21] elaboró su propia teoría de la significación y el lenguaje. Distinguiendo su componente sintáctico o las reglas del lenguaje y su componente gramático o semántico, la dimensión del significado [24].

De la misma forma Gödel influido por esta visión fue capaz de mostrar cómo había lenguajes formales que podían ser reducidos a matemáticas. Idea que utilizó para desarrollar la demostración de sus teoremas de incompletitud, al reducir el sistema axiomático de la aritmética a una serie de números con los que poder intentar demostrar desde la aritmética su propia consistencia. [25]

Estas ideas fueron fundamentales en el pensamiento de Turing pues el método que utilizó para demostrar la imposibilidad de decidir todos los teoremas en tiempo finito, es el mismo que el de Gödel. A saber, llegar a un estado indecidible porque es autorreferencial.

---

[20] **Aritmética de Peano:** era la axiomatización de la aritmética que se utilizaba en su momento definida por Giuseppe Peano en el siglo XIX. Es la predecesora de la axiomatización actual de la aritmética en que se reduce a teoría de conjuntos: la axiomática de conjuntos de Zermelo-Fraenkel.

[21] Es importante destacar a esta figura debido a que en parte el cisma que creó la división entre analíticos y continentales se debe a su polémica con Heidegger en los años 30. Carnap (socialista y científico), criticaba a Heidegger (nazi y de humanidades) utilizar un lenguaje oscurantista, que al ser analizado en detalle daba como resultado proposiciones carentes de sentido. Los pensadores de Viena problematizaban la metafísica por ser un campo de estudio carente de significado y con su positivismo lógico pretendían mostrar que muchos de sus problemas eran pseudoproblemas creados a partir de un uso impropio o incorrecto del lenguaje. A su vez Heidegger criticó esta visión al reducir el mundo a una serie de proposiciones cognoscibles, que en sus propios términos sería una de las máximas expresiones del ser-a-la-vista, una concepción que reduce el mundo a los parámetros de estudio que le interesan y es incapaz de lidiar con sus límites y aquello que escapa a la razón.

3.1.1.3 El Mecanicismo

Dentro del círculo de Viena y herederos de planteamientos como los de Boole, Charles Babbage y Ada Lovelace, se intentó llevar a cabo una formalización del pensamiento como un proceso puramente mecánico y sometido a reglas. Desde esta concepción, la mente humana y todo proceso de razonamiento se podían reducir a una serie de operaciones de manipulación simbólica que guiaban y determinaban el proceso mental. Esto es lo que en su momento se llamó mecanicismo y en la actualidad computacionalismo en teoría de la mente.

Esta visión es precisamente la que guio a los padres de la informática a intentar formalizar estas acciones de procesamiento para que así pudiesen ser llevadas a cabo de forma mecánica por máquinas e instrumentos de procesamiento automatizados.

Es precisamente esta una de las influencias cruciales que recibió Turing de esta corriente, aunque no se puede señalar con precisión si fue recibida por el lado de los teóricos de Viena, o una lectura de los textos de Lovelace, en que resumía su tiempo junto a Babbage y su pensamiento sobre la posibilidad de llevar a cabo el ideal de Babbage y su motor analítico.

Al mismo tiempo los desarrollos que habían permitido mostrar cómo la lógica podía reducirse a la aritmética, tanto en sus expresiones como en sus procedimientos, apuntaban a un futuro brillante. El significado de una expresión se podía expresar de forma numérica y se podía operar conforme a reglas de cálculo. Esto posibilitaba realizar el sueño de Boole de un cálculo proposicional estrictamente lógico y al mismo tiempo sentaba las bases para que se pudiese llevar a cabo de forma automatizada, abriendo la puerta a la posibilidad de una inteligencia no humana, o por lo menos una máquina que llevase a cabo dichos procesos.

El desarrollo teórico que hizo Turing de sus máquinas le permitiría durante la II Guerra Mundial ponerlo en práctica y utilizarlo como herramienta criptográfica para romper la máquina Enigma de los alemanes. Es precisamente a raíz de esta experiencia y sus investigaciones posteriores durante el periodo de posguerra que le llevarían a escribir en su último año de vida el artículo sobre el test de Turing.

3.1.2 Limitaciones de la Filosofía Analítica

A partir de la exposición anterior hemos podido llegar a apreciar cuales eran las intuiciones y preocupaciones fundamentales de esta escuela así como su método y forma de estudio. Las influencias del logicismo, su teoría del conocimiento y las matemáticas. Pero si leemos con atención a lo aquí expuesto, su planteamiento inicial (ya superado por los desarrollos posteriores a la segunda mitad del siglo XX) era excesivamente reduccionista.

Su concepción de la realidad como exclusivamente limitada a lo empírico y medible en términos verificables y comprobables, les llevaba a desechar todo aquello sujeto a una interpretación. Su proyecto era excesivamente formalista y rígido al definir un dominio muy limitado para su campo de estudio con lo que se puede saber con certeza.

A su vez, como el mismo Wittgenstein señaló en su segunda etapa, era incorrecta la idea del lenguaje como forma de comunicación sobre la que se podía establecer un conjunto de reglas definidas con las que poder establecer su significado y reducible a una formalización. Señaló como

el lenguaje tenía un componente intrínsecamente social, lo que denominó juegos del lenguaje, en que dependiendo del contexto se utilizaba una serie de reglas u otras y su significado venía dado por ese mismo contexto. Dicho de otra forma Wittgenstein destacó que la fuente de todo significado era la pragmática, los usos que se hacen de las palabras en diferentes contextos, frente a la semántica o significado estable e incontestable tradicional.

Además las críticas expresadas en párrafos anteriores también se estaban formulando por parte de otros pensadores ajenos a estas escuelas. Por ejemplo Maurice Merleau Ponty con su *Fenomenología de la percepción* ya estaba destacando que la visión de la cognición como un proceso formalista o puramente intelectual, ignoraba por completo la dimensión corpórea y viva de la experiencia humana [26]. Como el cuerpo humano es una parte indispensable de la cognición, visión que en los últimos años ha servido como punto de partida a la corriente de la neurociencia de la cognición incorporada. Otras de las críticas a la epistemología y concepción analítica han quedado reflejadas en el apartado 2.3.3 de la sección anterior.

Considero de especial importancia señalar estas cuestiones debido a que, históricamente la filosofía que más ha influido en la IA ha sido la analítica. Pero es solo en virtud de esta concepción reduccionista, tanto de la realidad como de la mente, que se podía plantear tan alegremente la posibilidad de una inteligencia artificial fuerte, al menos desde presupuestos tan naifs. Solo si entendemos este contexto filosófico e intelectual podemos llegar a comprender que Turing pensase que: con el programa adecuado, podía llegar a crear una inteligencia real.

3.1.3 La influencia en Turing

Como hemos expuesto en los apartados anteriores Turing, al haber estudiado en Cambridge, lugar de estudio y docencia de Russel y Whitehead y en la que su pensamiento había sido tremendamente influyente, entró en contacto, de forma consciente por una lectura activa de sus textos, o inconsciente debido al contexto intelectual que se respiraba en ellas, con su filosofía y pensamiento.

Estudiar matemáticas y lógica a principios de los años 30, cuando el proyecto formalista de Hilbert todavía seguía vivo, junto con la amplia difusión de los escritos e ideas de los pensadores de Viena, no pudo dejarle indiferente. Ya fuese Schlick con su fundamentación de las ciencias, Gödel con su papel determinante en la matemática y la lógica de la primera mitad del siglo XX con sus reducciones axiomáticas y luego sus teoremas de incompletitud o Carnap y sus reflexiones sobre el sentido y los significados. Por no mencionar a todo el resto de intelectuales que siguió esta corriente y participó de sus reflexiones, ya fuese la multitud de miembros que quedan sin citar pertenecientes al círculo, o la infinidad ajena a él, que acudía a sus conferencias y reuniones. Al igual que en el caso anterior, Turing se tuvo que ver influenciado por el movimiento intelectual dentro de las ciencias más importante de su época.

De la misma forma, el desarrollo de su carrera y los círculos en los que se movió le pusieron en contacto con otros matemáticos, lógicos, científicos y filósofos influenciados por estas ideas. Mismamente, su tutor de tesis doctoral, Alonzo Church, también resolvió el problema de la decidibilidad de forma casi simultánea a Turing utilizando el cálculo-λ, influido también por Gödel y sus conversaciones con él sobre la posibilidad de la definición de procedimientos recursivos. [25]

Dado el contexto de Turing y donde se desarrolló, no podía haber pensado de otra manera. Más aún si tenemos en cuenta, que al contrario que otros matemáticos como G.H. Hardy conocido por comer con "los de ciencias" y cenar con "los de letras" [27], Turing nunca se llegó a interesar especialmente por campos ajenos a su dominio de estudio, lo que podría haberle puesto en contacto con posturas contrarias a las suyas que hubiesen estimulado su pensamiento en nuevas direcciones.

3.2 Breve historia de la IA después de Turing

En este apartado vamos a explorar un poco cual fue el sentir general de los expertos que, siguiendo los pasos de Turing, empezaron a trabajar en el incipiente campo de la IA. Este sentir general es de especial interés ya que, sin tenerlo en cuenta, las críticas que se enuncian en el siguiente apartado no se entienden. Este apartado solo cubrirá los 30 primeros años(1950-80), dado que es sobre los resultados obtenidos en esta época sobre los que se enunciaron las críticas que se van a señalar más adelante.

3.2.1 La dialéctica histórica de la IA

Aunque no podremos reflejarlo aquí en toda su profundidad el campo de la IA se ha desenvuelto siempre en una dialéctica[22] entre sus más fervorosos creyentes y sus críticos más acérrimos. En esta dialéctica siempre se produce el mismo devenir.

Los creyentes, entusiasmados se lanzan a desarrollar programas que consiguen conquistar nuevo terreno a las tareas y dominios antes solo reservados a los humanos. Al mismo tiempo, sus críticos, aunque reconocen los avances, los miran con escepticismo y señalan sus limitaciones, estos programas no terminan de capturar la totalidad del pensamiento humano.

Con el paso del tiempo las limitaciones de los resultados en IA, provocan que los críticos cobren notoriedad haciendo que las expectaciones y ánimos de los creyentes se enfríen y retrocedan (lo que en la literatura especializada se ha llamado inviernos de la IA).

Pero a su vez, las críticas sirven para la reflexión y estimulan nuevos avances en el campo, que junto al avance de la tecnología revitalizan sus aspiraciones. Esto trae consigo una nueva generación de creyentes dispuestos a triunfar donde sus antecesores fracasaron. Reiniciando el ciclo de nuevo.

Podemos decir que el campo de la IA y la posibilidad de obtener una inteligencia real se puede equiparar al de la fusión fría, un resultado que siempre se ha postulado a 20 años vista, si solucionamos algunos problemas "menores" a nivel técnico. Pero pasados esos 20 años los expertos siguen proponiendo la misma hoja de ruta y el éxito sigue aguardando a 20 años vista.

[22] Dialéctica: entendida en términos hegelianos o marxistas supone una aproximación a la realidad que busca estudiarla como un proceso en constante flujo y diferenciación. Esto implica que un proceso por su propia evolución va pasando por distintos momentos que se van resolviendo sucesivamente, ya sea como negación determinada que invierte los términos y hace pasar al proceso a un estado opuesto (Tesis-Antítesis) o para ir más allá a una evolución o superación de lo anterior (Síntesis) .

3.2.2 Primeros resultados

En [28] Turing proponía que las limitaciones de memoria era el principal problema que enfrentaría una máquina para lograr una inteligencia real y vaticinaba que en 50 años esos problemas ya estarían solucionados.

Tras el artículo seminal de Turing, surgió una gran oleada de interés por el campo. La inteligencia en las máquinas parecía estar más cerca que nunca y los nuevos avances en ciencia de materiales, junto con la arquitectura Von Neumann desarrollada por el matemático húngaro en el Instituto de estudios avanzados de Princeton [29], auguraba un futuro brillante.

En 1956 se lleva a cabo el primer congreso dedicado íntegramente a la IA, con John McCarthy y Marvin Minsky como anfitriones, en el que se reunieron muchos de los pioneros del campo y se debatió ampliamente la cuestión. El congreso se terminó de forma amarga, ya que los investigadores reunidos no fueron capaces de ponerse de acuerdo ni en la naturaleza del problema, ni en que metodología utilizar para conseguirla. Haciendo que en el seno del incipiente campo de la IA naciesen diferentes escuelas y metodologías que trataron en el problema desde sus propios enfoques. Las dos más importantes fueron la conectivista, precursora del campo del aprendizaje profundo actual, y la formalista, basada en el uso de lógica simbólica.

Se comienzan a lograr resultados prometedores, un amplio dominio de problemas comienza a ser atacado con resultados más o menos satisfactorios para el hardware de su época. Arthur Samuel desarrolla un programa capaz de jugar a las damas, el programa ELIZA logra mantener conversaciones, mientras que Simon y Newell publican un modelo para un programa que simula el pensamiento humano. En 1965 Herbert Simon afirma “En 20 años, las máquinas serán capaces de realizar cualquier tarea que pueda hacer un humano.” [30]

3.2.3 Primeras limitaciones

Tras años de éxitos constantes los resultados comienzan a estancarse, el sueño de lograr una inteligencia real a 20 años vista tal y como fue formulado por McCarthy y Minsky en su primer congreso, todavía se encuentra lejos.

Son los inicios de los 70 y Hubert Dreyfus publica *What computers can’t do* [31], libro que le sirve para ganarse la enemistad de la mayor parte de investigadores del campo, todos ellos firmes creyentes de que la IA no es solo una posibilidad real, sino que ya se está logrando.

Pero los retrocesos comienzan a hacerse de notar. Los programas que se han desarrollado hasta la fecha todavía son incapaces de lograr una inteligencia real. Un ejemplo lo podemos ver en los asistentes conversacionales. Los problemas iniciales que, en un principio, parecían sencillos de solventar con el avance de la tecnología y de las técnicas de computación, se muestran cada vez más intratables y complicados. [30]

De la misma forma, problemas que en principio parecían tratables y sencillos de resolver, se vuelven intratables. Al intentar solucionar problemas que caen en las clases de complejidad NP o superiores, se enfrentan a un aumento inmenso del espacio de búsqueda y coste computacional, inabarcable para la tecnología de su época.

3.2.4 El primer invierno de la IA

Los ánimos y la expectación alrededor de la IA comienzan a decaer a mediados de los 70. Esto no hace que los más enfervorizados ponentes de esta posibilidad se amilanen. Los pioneros del campo consideran que la posibilidad sigue siendo real, solo falta inversión y tiempo para ponerla en práctica. Pero el interés general alrededor de la cuestión está en claro declive.

Durante unos pocos años los avances en el campo se estancan, parece que el sueño de lograr una IA nunca fue más que eso. Se reconoce que los programas son útiles para hacer la vida más sencilla a las personas y lograr la automatización de algunas tareas, pero la posibilidad de traer al mundo máquinas pensantes queda en entredicho.

Este declive viene motivado por un freno al gasto en esta tecnología y su investigación. Agencias, países e instituciones que en un principio habían aportado cuantiosas sumas de dinero, esperando resultados, comienzan a cerrar el grifo al ver que los resultados no cumplen con las promesas.

3.2.5 Vuelta a la casilla de salida

Es entonces cuando surgen nuevos marcos teóricos, a finales de los 70 aparece el paradigma del sistema experto [32], que dotado de un motor de hechos y reglas, puede razonar como una persona. A su vez los descubrimientos del físico John Hopfield vuelven a poner de moda las redes neuronales. Esto trae consigo un interés renovado en el campo y la vuelta de la financiación a principios de los 80.

McCarthy pública por aquel entonces su artículo *Ascribing Mental Qualities to Machines* [33] en que argumenta que se puede decir que un termostato tiene creencias sobre la realidad, a saber la temperatura de la sala. Siendo las creencias una de las características esenciales que identifica McCarthy en un sistema inteligente capaz de razonar. Ese artículo se convertirá en una de las razones principales que impulsen a Searle a comenzar sus críticas a la IA.

3.2.6 Conclusión

Este apartado podría continuar hasta el presente, pero desbordaría la función y el interés que cumple. Como hemos podido apreciar en esta mirada a vista de pájaro del campo de la IA en sus inicios es que estuvo marcada por un gran optimismo predicado en el sueño de que la inteligencia real, se encontraba cercana en el tiempo. El resto de picos y valles de interés por la IA han venido reproduciendo ese mismo optimismo, siendo la revolución actual del Big Data y las redes neuronales, su último ejemplo.

Hasta el punto que declaraciones como las de Ilya Sutskever, jefe de investigación de Open AI, en que afirmó [34] que las redes neuronales actuales son ligeramente conscientes; son un eco de esta tendencia histórica a adscribir inteligencia o consciencia a los últimos logros del estado del arte.

Una vez comentado estos aspectos, pasamos a centrarnos en los argumentos que más han marcado el debate sobre la posibilidad de la IA fuerte.

3.3 Argumentos en contra de una IA Fuerte

En este apartado se explorarán algunos de los argumentos más influyentes que se han dado hasta la fecha en contra la posibilidad de llegar a lograr una IA fuerte. La exposición de los argumentos seguirá la pauta de una explicación inicial, y en caso de ser posible una crítica posterior que muestre como tampoco son tan incontestables como parecen.

3.3.1 Argumentos de Searle

John Searle puede considerarse uno de los grandes críticos de la IA del siglo pasado. En su artículo *Minds Brains and Programs* [16] dio dos de las críticas más sólidas contra la posibilidad de la IA fuerte, tal y como se definió más arriba. Las críticas propuestas por Searle son las siguientes.

3.3.1.1 La habitación china

Este es un experimento mental propuesto en su artículo en que propone lo siguiente: Imagina que te encuentras en una habitación cerrada repleta de libros con entradas del siguiente tipo: si recibes la cadena de caracteres en chino “xyz” entonces responde “abcd”. A continuación imagina que comienzas a recibir del exterior tarjetas con caracteres en chino y tu trabajo consiste en revisar los libros hasta que encuentres la cadena que has recibido para responder con lo especificado en su respuesta, que escribes en una tarjeta y envías al exterior. La cuestión es la siguiente: el receptor de dichas tarjetas, al encontrarse con respuestas correctas o adecuadas a sus mensajes, pensaría que la persona con la que se está comunicando, entiende chino; pero tú, persona que se encuentra en el interior, no tienes un conocimiento semántico, no comprendes lo que estás recibiendo ni enviando ; no entiendes chino.

A partir de este experimento, Searle, llega a la conclusión de que un manejo puramente simbólico y formal de las palabras, basado en reglas de modificación y sustitución, al que denomina como sintáctico, es completamente distinto del tipo de sistema que tenemos los humanos, a partir del cual derivamos un conocimiento semántico de las proposiciones que manejamos.

A partir de esta idea, Searle establece un corte taxativo entre sintaxis y semántica, diciendo que de una manipulación exclusivamente sintáctica, no puede emerger la semántica. De esta forma, pone en cuestión las pretensiones de los investigadores de su época, de intentar adscribir algún tipo de contenido mental a cualquier programa capaz de manipular símbolos; y mostrar algo similar al entendimiento humano por estos medios.

3.3.1.2 Limitaciones de la Habitación China

Si leemos con atención este argumento y tenemos en cuenta cuando se llevó a cabo, finales de los 70. Podemos apreciar que aunque fuese una crítica sólida en su momento a día de hoy, ya no es tan incontestable como podría parecer. Hay que tener en cuenta que esta crítica se realizó en un momento en que la forma de procesamiento del lenguaje era sintáctica, manipulando las palabras como símbolos, mediante reglas y establecidas de ante mano por el programador.

Pero si tenemos en mente el funcionamiento antes mencionado de un modelo del lenguaje, su argumento hace aguas. En el caso de un modelo del lenguaje entrenado mediante *machine learning*, la manipulación que hace de los símbolos al venir dada por una compleja red de relaciones diferenciales, sí está más próximo a una semántica real, que en el caso planteado por Searle. Vemos que como tal, este argumento solo era aplicable a las técnicas de su época, pero no planteaba ninguna imposibilidad fundamental, en la medida en que el avance de la técnica y la tecnología ha permitido nuevas formas de desarrollarlo[23].

3.3.1.3 Simulación vs duplicación

El segundo eje sobre el que pivota la crítica de Searle es la diferencia entre duplicación y simulación. Su argumento se basa en que lo que está haciendo un ordenador, al ejecutar el programa, no es duplicar una mente, sino simularla. Aunque este argumento parezca basarse en un juego del lenguaje, la diferencia que está recalcando no es meramente retórica sino ontológica, es decir, está señalando la diferencia fundamental que existe entre una entidad que duplica a otra, frente a una que simula.

Vamos a centrarnos en apreciar esta diferencia mediante un ejemplo. Imaginemos el sistema digestivo humano: compuesto de una serie de órganos que ejecutan funciones específicas en el proceso de digestión. Una duplicación del sistema digestivo implicaría una copia, una réplica con los mismos componentes que el sistema original, capaz de llevar a cabo, por los mismos medios las mismas funciones. Mientras que una simulación, ya sea virtual mediante una simulación por ordenador, o un modelo físico de plástico en el que se introducen alimentos de mentira que van pasando por los distintas partes del proceso hasta terminar en el equivalente al intestino grueso, es de una naturaleza completamente distinta.

La duplicación, aunque distinta del original, tiene componentes del mismo tipo y es capaz de llevar a cabo las mismas funciones de la misma forma que el original, mientras que la simulación, incluso cuando es una copia fidedigna del original, implica un cambio de medio, unos componentes fundamentalmente distintos. Dicho de otra forma, un sistema digestivo duplicado puede digerir un alimento, mientras que no hay simulación, por sofisticada que sea, capaz de llevar a cabo la misma función.

3.3.1.4 Limitaciones de Simulación vs Duplicación

Hay que tener en cuenta que el argumento de la duplicación y la simulación se formuló en su momento con otro objetivo distinto (comentado en la nota al pie de página nº10). Por tanto, podemos entender que el dominio de aplicación de este argumento se reduce a los sistemas conversacionales que estaba criticando en su momento y cualquier otro sistema que opere con el mismo funcionamiento subyacente. Pero esto no implica en absoluto que dado el diseño y estructura correcta esta no pueda llegar a lograrse.

Si atendemos a lo dicho en otros artículos posteriores [8] Searle mantiene esta visión reduccionista y exclusivista de la conciencia a su dimensión puramente biológica. Searle la está

[23] Aunque no por ello hay que olvidar las críticas llevadas a cabo en el apartado 2.3.2 similar a la de Searle, pero en mi caso, solo estoy señalando una carencia, no una imposibilidad, al considerar que es una herramienta que se puede mejorar con las técnicas adecuadas.

entendiendo como un proceso puramente biológico en el que participan neurotransmisores y neuronas, impulsos eléctricos y hormonas, sin las cuales considera que es imposible que haya una conciencia. Pero el fondo del argumento es una cuestión puramente metafísica sobre la que, al menos, dado el estado actual del campo de la neurobiología y la nuerofísica es muy pronto para darlo por cerrado o poder afirmar tajantemente tal cosa.

En la medida en que el núcleo fundamental de este argumento es la cuestión metafísica, que informa el resto de la argumentación, considero que no es un argumento incontestable. Ya que la diferencia ontológica entre lo duplicado y lo simulado depende esencialmente de considerar que la conciencia es un fenómeno intrínsecamente biológico. Pero desde el marco definido en el apartado 2, que contempla la conciencia como un fenómeno emergente de la integración de un sistema complejo, a priori no es una imposibilidad, o por lo menos es muy pronto para decirlo.

3.3.2 Argumento de Penrose

Este argumento lo desarrolla Roger Penrose, físico norteamericano y Nobel de física en 2020, en sus libros *La nueva mente del emperador* [35] y *Sombras de la mente* [36]. Exponer en profundidad un argumento desarrollado en las casi 1000 páginas que suman ambos volúmenes, en los que entreteje conceptos de relatividad general, física cuántica, computación, filosofía de la mente, matemáticas y neurociencia; se presenta imposible. Debido a esto procederé directamente a dar una pequeña intuición de lo ahí expuesto.

3.3.2.1 No computabilidad de la mente

El punto de partida de la visión de Penrose es que: la mente, como fenómeno emergente de los procesos físicos subyacentes, no es una entidad que pueda ser computable. Con esta afirmación lo que quiere decir es que no es computable en el sentido de que es un fenómeno irreductible al cálculo determinista que se produce en nuestros ordenadores digitales.

Parte de los excursos que lleva a cabo en sus libros por los campos antes mencionados, le sirven para problematizar cómo la mente al tener un sustrato cuántico, en que las leyes deterministas y corpusculares de la física clásica dejan de aplicarse, suponen un obstáculo a tener en cuenta. La física cuántica al involucrar una descripción de objetos físicos que ya no son partículas determinadas en un momento y lugar, sino vibraciones en un campo cuántico que se describen como agregados de probabilidades en distintos lugares, pone un obstáculo de difícil solución para traducir los procesos mentales a procesos digitales. Esto se debe a que esta manifestación cuántica existe en las terminaciones nerviosas del cerebro, en el intercambio de neurotransmisores entre las dendritas de las neuronas que generan una diferencia de potencial eléctrico y como este es una emanación del campo electromagnético, de naturaleza fundamentalmente cuántica.

Por si misma esta justificación no comporta el núcleo duro de su planteamiento, en la medida en que los ordenadores al ser digitales y hacer uso de electricidad, también están en contacto con el campo cuántico electromagnético (y con todos los demás, al ser objetos físicos). Es por ello que el núcleo de su argumentación se encuentra en otra parte, en una justificación matemática que hace uso de los teoremas de incompletitud de Gödel.

La idea principal que desarrolla en ambos libros, refinada, es que: partiendo del resultado del primer teorema de incompletitud, a saber existen proposiciones verdaderas en el sistema, pero que son indemostrables, concluye que la mente humana supera el formalismo matemático. Esto implica que el hecho de ser capaces de reconocer que existe un enunciado que es verdadero en el sistema, pero que no podemos probar dentro del sistema, aun sabiendo que es cierto. De ello se sigue que **como tenemos una instancia de cómo nuestra mente es capaz de conocer la verdad de una proposición que nunca seriamos capaces de demostrar dentro del sistema, nuestra mente es irreducible y no puede ser expresada en términos matemáticos.**

3.3.2.2 Limitaciones no computabilidad de la mente

Como bien señala Enrique Alonso en [25], el argumento de Penrose tiene un defecto insalvable. Este consiste en que está malinterpretando el primer teorema y sus consecuencias. "[...] el error se encuentra en la lectura incorrecta del primer teorema de incompletitud de Gödel. Este no dice que G sea indemostrable en PA –Peano Arithmetic– sino que si PA es consistente, entonces G es indemostrable en PA."

Las consecuencias de esta malinterpretación son vitales, no hay una intuición de la verdad de la proposición, ni un argumento racional no formalizable con el que explicar por qué lo es. Lo que se está demostrando es un enunciado condicional que es formalizable en PA y que además forma parte de la demostración del segundo teorema. Dicho de otra forma, toda la argumentación de Penrose se viene abajo debido a una mala lectura y comprensión de estos teoremas y sus consecuencias**. Si no hay enunciados matemáticos de los que podemos conocer su verdad de forma intuitiva, pero de los que somos incapaces de dar una prueba formalizada**, esta barrera insalvable cae por su propio peso.

3.3.3 Argumento de Dreyfus

Este argumento fue presentado por Hubert Dreyfus en sus libros [37] [31] .Sus exposiciones son realmente interesantes, pues en su época 1972, sirvieron para reducir el entusiasmo y expectación injustificada que había con los ordenadores. Dado que en aquel momento en el campo de la informática y de la IA en particular, existía una sensación generalizada de que la IA fuerte (aunque ese término no existiese hasta la crítica de Searle) estaba a la vuelta de la esquina.

3.3.3.1 Conocimiento formal e informal

Para sintetizar la visión de Dreyfus, sin perdernos en todos los excursos que lleva a cabo en el pensamiento de varios filósofos como Wittgenstein y Merleau-Ponty, vamos a intentar resumir su visión recurriendo a las ideas del ser-a-la-mano y el ser-a-la-vista desarrolladas en el apartado 2.2.1.2.2.

El punto de partida de Dreyfus es una idea similar a lo que se conoce como paradoja de Moravec: **la potencia computacional que se necesita para llevar funciones cognitivas superiores, (cálculos matemáticos) es infinitamente inferior a la necesaria para llevar a cabo funciones de bajo nivel (percepción o movimiento)** [38]**.** Esto le lleva a diferenciar dos tipos de dominios de conocimiento, los formalizables y los no formalizables, aquellos para los que se puede desarrollar una descripción formal, mediante reglas y enunciados explícitos y objetivos; y los que no. Un

ejemplo sería como resolver una ecuación de segundo grado, en el que se puede formular de forma sencilla un método formal de resolución en una serie de pasos. Pero para una actividad cómo deshacer un nudo, que es un problema que puede resolver fácilmente un niño pequeño, formalizarlo en una serie de pasos se convierte en una tarea infinitamente más compleja.

Hasta cierto punto esta división entre lo formalizable e informalizable, es muy similar a los modos de conocer el mundo. Siendo los formalizables equiparables al ser-a-la-vista, una forma de saber teórico basado en la reflexión. Mientras que tenemos otros, los informalizables, de naturaleza intuitiva, que serían equiparables al ser-a-la-mano en el que nuestro conocimiento de las cosas es prerreflexiva, puramente activa.

A partir de esta diferenciación, Dreyfus concluye que existe un corte taxativo entre ambos dominios de conocimiento y las máquinas nunca podrán desempeñar tareas correctamente en todos aquellos campos que son informalizables. Este corte le sirve para plantear las limitaciones intrínsecas de los ordenadores y justificar que la conciencia sea otro de esos aspectos que tienen vedados.

3.3.3.2 Limitación conocimiento formalizable e informalizable

En el caso de este argumento, al ser solidario con lo que he expuesto en el apartado 2.2.1.2.2 y ser una posición con la que en parte estoy de acuerdo, solo tengo una cosa que señalar. Ahí donde Dreyfus, plantea una imposibilidad fundamental de los ordenadores para poder llevar a cabo estas tareas yo solo veo un abandono histórico de la cuestión, que nada dice de lo que se puede llegar a lograr en el futuro.

Cuando Dreyfus llevó a cabo sus críticas el dominio del conocimiento informalizable todavía no era un tema que se estuviese estudiando con especial atención, todo lo que se tenían eran sistemas deterministas y preprogramados, sin ningún tipo de capacidad de adaptación. Pero sus críticas, desde mi punto de vista, solo señalan que en el momento en que las llevó a cabo era un dominio de problemas en los que no se estaba trabajando activamente. Esto implica que la conclusión de ese corte insuperable entre ambos dominios de conocimiento solo refleja el estado del arte del campo en aquel momento, pero no necesariamente una imposibilidad por parte del mismo para lograrlo o llevarlo a cabo. [39] [40]

3.4 Conclusión

El objetivo de esta exposición anterior era poner en tela de juicio los argumentos más importantes que se han planteado a la posibilidad de una IA fuerte, para mostrar que aunque en primer lugar pudiesen parecer sólidos e incontestables, la posibilidad parece seguir abierta. Esto es de vital importancia, en la medida en que en el último apartado se intentará plantear un test alternativo y una serie de propiedades que tiene que tener un sistema para ser consciente.

4

Turing y sus herederos

En este apartado se va a llevar una crítica al test de Turing tal y como lo formuló en [28]. El objetivo de este análisis es poner de manifiesto la concepción que tenía Turing de las máquinas pensantes. Solamente teniendo en cuenta su visión, podemos superarla, rescatar los aspectos útiles y desechar los desfasados. A su vez en el último apartado se llevará a cabo una pequeña revisión del estado del arte, una vez desarrollado el elemento que pretenden superar.

En este artículo, Turing, propone reformular la pregunta "puede pensar una máquina "por "podría ganar una máquina en el juego de la imitación". Este se propone como una prueba en la que un examinador A hace preguntas a dos agentes B y C, con el objetivo de determinar cuál es la máquina y cuál el humano. A partir de este, Turing propone que: engañar al examinador puede ser considerado un indicador de la inteligencia de la máquina. Vamos a comenzar analizando esta cuestión, explorando las críticas más básicas que se pueden enarbolar contra la formulación aquí planteada.

4.1 Críticas a la formulación del test

En un inicio las deficiencias más claras que muestra esta formulación son las siguientes: se centra en el engaño, el test es subjetivo (hace referencia a la subjetividad del examinador.

4.1.1 Centrado en el engaño

El hecho de centrarse en el engaño, responde a la lógica de que Turing evita dar una descripción de lo que significa pensar. Esto se debe a que en sus propias palabras:

Si el significado de las palabras 'máquina' y 'pensar' hay que buscarlo en su uso común, es difícil escapar a la conclusión de que el significado y la respuesta a la pregunta '¿Pueden las máquinas pensar?' Han de ser buscados en un estudio estadístico como la encuesta Gallup." [28]***[pág 1]***

La lógica subyacente a esta afirmación no se hace explícita en ninguna parte del texto. El motivo por el que Turing piensa que el significado de esos términos hay que buscarlo en su uso corriente, sólo se puede justificar en relación a la filosofía analítica del lenguaje, antes mencionada. En este caso particular estaría haciendo referencia a la escuela de Oxford del

lenguaje Ordinario de J.L. Austin. Esta escuela propone partir del significado ordinario, del día a día, de las palabras para conocer en profundidad tanto sus usos como sus límites de aplicación.

El problema que tiene esta fundamentación en el engaño es que se basa en la asunción oculta de que para engañar se requiere de cierta inteligencia. Pero esta asunción solo es cierta en el caso de un humano. Si yo quiero engañar a otra persona, tengo que poner en práctica cierto conocimiento del mundo de una forma que al engañado le resulte creíble y coherente. Hay implícita detrás de mí acción una cierta intencionalidad, que en el caso de la máquina no es tal. Solo se puede hablar de engaño en la medida en que el agente esté actuando de forma poco honesta y esto sea una decisión propia. Una elección de entre varias, fruto de una respuesta al contexto en que se está desenvolviendo.

Podemos ver un ejemplo de este engaño intencional en la película "Ex Machina". Esta película es de especial interés para este comentario porque gira en torno a un test de Turing especial que están haciendo a una IA puntera. En ella, el protagonista, se entrevista con una androide con la que va manteniendo conversaciones. El motivo es intentar determinar si tiene conciencia o no y en el proceso se enamora de ella. Pero en su desenlace se revela que la IA le ha engañado. Estaba actuando para seducirle y así poder escapar.

En este tipo de engaño encontramos un tipo de inteligencia que se presupone como indisociable, pero que no es tal. La IA de la película utiliza el engaño como un medio para un fin, porque razona que es la forma más viable de huir. Pero si el engaño no es un medio para un fin sino un fin en sí mismo, que engaña por engañar, porque es así como se la ha programado, no podemos hablar de una inteligencia real. Esto hace que utilizar el engaño como proxy de la inteligencia se convierta en una métrica contraproducente.

4.1.2 Test Subjetivo

Al hablar de que el test es subjetivo, nos referimos a que el juez puede ser cualquier persona y en su artículo no detalla ningún tipo de criba para determinar quién podría serlo. Esto abre el problema de que frente a un mismo caso en que dos jueces hacen las mismas preguntas y tanto la máquina como la persona dan las mismas respuestas, uno podría ser engañado mientras que el otro no. A su vez esto revela otro problema, una misma máquina podría tener tasas de éxito completamente distintas dependiendo de la muestra de jueces, incluso podemos hipotetizar sobre distintos tipos de máquinas que lo harían mejor frente a algunos jueces que a otros.

Por ejemplo, imaginemos una máquina con un modelo representacional muy potente y sofisticado que es muy buena a la hora de responder preguntas de conocimiento científico o de carácter teórico o lógico. Si la ponemos frente a un conjunto de jueces compuesto de profesionales de estos campos, es probable que formulen este tipo de preguntas, haciendo que sea más probable que la máquina tenga una mayor tasa de éxito en comparación con otra muestra de jueces. Esta objeción también es aplicable a una muestra compuesta por jueces especialmente malos para la máquina haciendo que esta diese una tasa de éxito muy baja.

Si llevamos esta objeción hasta sus últimas consecuencias, se aprecia que no hay posibilidad de derivar ningún tipo de criterio de selección que prime a un tipo de jueces sobre otros, al menos a partir de las evidencias textuales. Por tanto podemos concluir que el criterio del juez siempre

es subjetivo y que esta subjetividad inherente comporta un problema fundamental a la hora de dar validez a los resultados obtenidos en el test por una máquina. Está objeción además es una consecuencia derivada del problema de basar el test en el engaño.

Además en el propio artículo señala que es un experimento que se repetiría varias veces y que tendría tasas de éxito determinadas. Dada esta objeción, basar la decisión de si una máquina tiene inteligencia o no en su tasa de éxito, no parece la mejor opción.

4.2 Concepción de la mente

Al analizar los apartados 4 (ordenadores digitales) y 5(universalidad de los ordenadores digitales) se vuelve obvio cuál es la comprensión que tenía Turing del funcionamiento de la mente. Partamos de las evidencias del texto para construir nuestras intuiciones a partir de ella.

"La idea detrás de un ordenador digital puede ser explicada diciendo que estas máquinas deben llevar a cabo cualquier operación que pudiese hacerse por un 'computador humano'. El computador humano ha de seguir una serie de reglas establecidas; no tiene autoridad para desviarse de ellas de forma alguna" [28]***[pág 4]***

A partir de esta definición Turing pasa a describir un ordenador que seguiría la arquitectura Von Neumann. Aunque Turing deja fuera los dispositivos de entrada y salida, estos van implícitos en su funcionamiento; debido a la necesidad de recibir las preguntas del juez y poder dar respuesta a ellas.

En el apartado 5, identifica a los ordenadores digitales con máquinas finitas de estados, para a continuación continuar con su argumentación a favor de que fuese posible al menos una máquina de este tipo capaz de obtener buenos resultados en el test. Basándonos en la objeción antes señalada de la habitación china de Searle, podemos ver que la empresa de Turing, tal y como él la planteó, está condenada desde el principio. Ya que estas posibles máquinas siempre van a incurrir en este problema de la manipulación simbólica, dicho de otra forma la sintaxis es distinta de la semántica. El modelo tal y como él lo plantea estaría operando con símbolos, de los que nunca conocería el significado.

Pero esta no es la única pega que podemos señalar al planteamiento de Turing, ya que en este apartado podemos apreciar una reducción de la mente a una máquina finita determinista, que ni siquiera en su versión con memoria ilimitada, parece poder hacer frente. El problema viene de la reducción de un epifenómeno como es la mente, que emerge de procesos biológicos, no lineales y paralelos, a uno puramente determinista y mecanicista.

Implícitamente Turing mantiene esta postura al proponer que podría existir una máquina capaz de lograr este objetivo, es decir lograr un buen desempeño en el test. Podemos llegar a estar de acuerdo con su postura, en la medida en que quede circunscrita a la objeción antes señalada, es decir que dicha máquina es capaz de imitar a los humanos pero como pura simulación de su proceso mental, reducida a una abstracción logicista, incapaz de emular su complejidad real. Es precisamente a esta concepción de la mente a la que se oponía frontalmente Searle con su argumento de la simulación y la duplicación. Una concepción que como hemos visto anteriormente ha recorrido todo el campo de la IA desde sus principios.

4.3 Crítica a los contraargumentos

En el apartado 6 Turing se encargaba de proponer una serie de argumentos en contra o refutaciones a su postura, a los que iba haciendo frente de forma sistemática. De esta forma pretendía abrir la cuestión a debate, así como para demostrar que esos peros no eran tan categóricos como pudiese parecer en primer lugar.

A continuación se van a analizar algunos de ellos. El motivo para hacerlo es señalar cuales estaban limitados por su concepción mecanicista de la mente y el pensamiento. No todos ellos merecen la pena ser mencionados ya que algunos entran de lleno en cuestiones teológicas (objeción teológica, nº 1) o pseudocientíficas (argumento de la percepción extrasensorial, nº 9), mientras que otros sí que son acertados (Argumento de Lovelace[24]).

La forma de llevar a cabo estas críticas consistirá en primero intentar sintetizar el argumento de Turing en una serie de proposiciones o formas más sencillas para luego elaborar sus refutaciones.

4.3.1 Argumento matemático

1. Las máquinas digitales tienen sus limitaciones: con esto se refiere a las consecuencias de los teoremas de incompletitud de Gödel, así como al problema de la parada, desarrollado por él mismo.

2. Los humanos también tienen limitaciones: al respecto de esta cuestión dedica una escueta frase en la que señala que *"[...] aunque se establece que hay limitaciones a las capacidades de una cierta máquina, se ha establecido, sin ningún tipo de prueba que no hay tal limitación que se aplique al intelecto humano [...] nosotros también damos respuestas equivocadas como para regodearnos ante las pruebas de la falibilidad de una máquina"[pág 13]*

3. Hay una tercera premisa oculta que emerge a la vista de esta última frase. Esta es que la naturaleza de estas limitaciones son equivalentes, consecuencia de su reduccionismo de la mente a una abstracción puramente computacional y funcionalista.

He aquí el punto débil de esta argumentación, ya que las limitaciones de un sistema axiomático no son equivalentes a las del intelecto humano. Nuestros razonamientos, como él mismo expone, son falibles, podemos incurrir en razonamientos falaces, inferencias erróneas, saltos deductivos y un largo etc de problemas.

Por otro lado los problemas que tienen las máquinas, tal y como él mismo reconoce, les llevan a entrar en bucles infinitos al encontrarse frente a preguntas indecidibles. Quizás sería demasiado osado afirmar de forma apodíctica que estas limitaciones son insoslayables por parte de las máquinas. Frente a este tipo de limitaciones, habría de ser la propia IA la que debiera identificar estos problemas y desarrollar estrategias para evitar caer en ellos. Si el propio diseñador limita a

[24] Argumento de Lovelace: la respuesta que plantea Turing es que Ada Lovelace cuando enuncia las limitaciones de la máquina de Babbage, estaba llevando a cabo un razonamiento inductivo a partir de la información que tenía disponible del estado del arte en ese momento. Pero por sí mismo eso no invalidaba que en el futuro pudiesen existir avances que lo posibilitasen.

la máquina de base para hacerla ciega a estas cuestiones o para que en caso de encontrarlas se las salte, no podemos estar hablando de una entidad inteligente como tal.

4.3.2 Argumento de la Consciencia

El propio nombre de este argumento es problemático ya que Turing en su formulación nos habla de la consciencia para luego citar un párrafo de Geoffrey Jefferson, pionero de la neurociencia de su época, del que reproduciré a continuación algunas partes, para señalar el equívoco en que incurre.

"*No será hasta que una máquina pueda escribir un soneto o componer un concierto debido a sus pensamientos y emociones, y no debido a la casualidad de ciertos símbolos, entonces podremos estar de acuerdo de que el cerebro es igual a la máquina. Esto es no* sólo *escribirlo, sino saber que lo ha escrito. Ningún mecanismo podría sentir [...] placer por sus éxitos, lamentarse cuando sus válvulas se fundan [...] enfadarse o deprimirse cuando no consiga lo que quiere"[págs 13-14]*

Para refutar esta posición Turing afirma que este argumento llevado a su extremo nos lleva al solipsismo, en que ya que la única forma de saber qué "siente" la máquina es ser la máquina, lo mismo puede aplicarse al resto del mundo y poner en cuestión el hecho de que los demás tengamos consciencia. De esta forma evita entrar en los detalles particulares de la objeción presentada por Jefferson.

Para aplacar su posición Turing propone una forma de comprobar si esto es así, poniendo el ejemplo de una máquina que ha escrito un texto a la que se le pregunta por partes específicas del mismo, para analizar el proceso de razonamiento que le ha llevado a realizarlo. Lo que él llama un método *"viva voce".* Aunque estoy de acuerdo con su forma de salir del atolladero, sí que considero que dada la naturaleza del sistema que planteaba Turing, alcanzar estas capacidades estaba fuera de su alcance.

Otra objección que se puede señalar frente al argumento de Turing, más allá de que haga un hombre de paja con la posición de Jefferson, es que finalmente la utiliza para argumentar desde una falsa dicotomía entre estar abierto a las posibilidades del test o caer en la posición solipsista. Como si no existiese un término medio entre ambas que permita poner en duda el test de Turing y la posibilidad de la Inteligencia Artificial Fuerte, sin caer en el solipsismo dogmático.

4.3.3 Argumento de las Discapacidades

En palabras del propio Turing este argumento sigue la forma: "Admito que una máquina puede hacer todo lo que tú dices, pero nunca serás capaz de conseguir que haga X". A continuación daremos una serie de ejemplos, dejando de lado aquellos que son triviales y aquellos con los que no encontramos problemas en su razonamiento. Nos centraremos en dos en particular, ya que el resto de interés serán discutidos en argumentos posteriores.

4.3.3.1 Una máquina no puede ser el sujeto de su propio pensamiento *(be the subject of it's own thought).*

Para demostrar que esto no es cierto, propone considerar el caso de una máquina resolviendo una ecuación de segundo grado. Los estados por los que pasa la máquina para determinar los resultados de dicha ecuación se podrían considerar como estados mentales y que esos estados constituirían el pensamiento del cual es sujeto la máquina.

Este argumento nos obliga a entender el pensamiento de una forma puramente formal y determinista, haciendo que otras posibilidades y experiencias mentales reales que participan del pensamiento desaparezcan y queden invisibilizadas. Reducidos a una suerte de estados mentales unívocos, que ofuscan la complejidad real del pensamiento, en la que se da una participación de múltiples "módulos" que cumplen funciones diversas, de forma distribuida y no jerárquica. Es el tipo de argumento para el que Searle planteó la objeción de la habitación china.

4.3.3.2 Una máquina no puede tener un comportamiento muy variado

La respuesta de Turing es tajante en este sentido, él entiende que siempre que se habla de este aspecto, realmente se está hablando de una falta de capacidad de almacenamiento. A la vez que mantiene que es una forma velada de reinstanciar el argumento de la conciencia.

Discursivamente hablando, su fundamentación y razonamiento a la hora de defender estas ideas es cuanto menos oscura y poco desarrollada. Además basándonos en todo lo dicho anteriormente, se puede interpretar que esta falta de comportamiento variado, no tiene por qué entenderse como una cuestión de pura capacidad de almacenamiento, sino que esta falta de comportamiento variado, se achaca a la condición puramente determinista y algorítmica. Así se puede señalar que lo que le falta a la máquina de Turing es la capacidad de evolución, de trascender sus estructuras intelectivas y ampliar su horizonte de posibilidades, de una forma completamente endógena.

Dicho de otra forma, incluso en el estado del arte de las redes neuronales actuales, el proceso de aprendizaje es puramente pasivo y unidireccional. Estamos lejos de tener una *tabula rasa* que tenga una apertura óntica[25] real a su entorno. En el caso de las redes neuronales, tenemos una apertura inicial, durante su proceso de entrenamiento, pero una vez iniciado, esta apertura se pierde por completo, reduciendo su horizonte de posibilidades a los que ofrece la tarea que está intentando aprender.

Entendiendo así esta diversidad de comportamiento, podemos apreciar que la refutación de Turing, no solo no es exhaustiva sino que se queda prácticamente en una réplica fácil y vaga que no entra al fondo de la cuestión.

4.3.4 Argumento de la Informalidad del comportamiento

El argumento del que parte es el siguiente:

[25] Concepto expuesto en el apartado 2.2.3

"Si cada persona tuviese un conjunto definido de reglas de conducta mediante las que regulase su vida, las personas no serían distintas de las máquinas. Pero como no hay tales reglas, no pueden ser máquinas." [pág 20]

A partir de este distingue entre reglas de conducta *[rules of conduct]* y leyes de comportamiento *[laws of behaviour]*, siendo las primeras reglas de las que la persona puede ser consciente, mientras que las segundas son leyes naturales que se aplican al cuerpo de la persona. Con esta distinción argumenta que si cambiamos en el argumento reglas de conducta por leyes del comportamiento, esta imposibilidad ya no es tan clara. Dado que como el mismo expresa, "estar regulado por leyes del comportamiento implica ser algún tipo de máquina".

El problema que encuentro con este argumento es que, aunque en principio pueda parecer sólido, está haciendo una cosa que han hecho históricamente todos los pensadores con la tecnología más sofisticada de su tiempo, a saber, decir "somos como esto". Durante los siglos anteriores era habitual la comparación del hombre con un reloj, la invención más precisa y delicada de su época. De la misma forma actualmente y con el auge de los modelos del lenguaje ya tenemos a gente identificando la mente con un proceso estocástico y probabilístico, en que los pensamientos e ideas vienen dados por estas leyes.

El problema que encontramos es que en este caso vemos una apropiación de la realidad por parte de la metáfora o analogía utilizada. Una reificación en resumen. Se puede decir que somos máquinas, con fines explicativos, o para simplificar un concepto y hacerlo más digerible. Pero en el momento en que pretendemos hacer pasar una abstracción explicativa, por una descripción fehaciente de la realidad. Esa misma realidad que se pretende describir desaparece en el proceso, al verse emborronada[26].

4.4 Máquinas que aprenden

Este es el punto más extenso de su artículo y en él se discute, a partir de todo lo dicho anteriormente, la posibilidad de crear algún tipo de máquina "vacía" similar al estado en el que nace un bebe. Esta *tabula rasa* contendría estructuras cognitivas con las que sería capaz de aprender, al igual que lo hace un niño y de esta forma, la complejidad del problema se reduce notablemente, ya que solo restaría someter a esta máquina a un proceso de aprendizaje adecuado para lograr que aprendiese.

[26] La metáfora "las personas son máquina" lo único que refleja es que hay una serie de similitudes entre las personas y las máquinas, gracias a las cuales se puede establecer la relación. Pero si solo primamos las similitudes, las diferencias que median entre ambas entidades y que las constituyen como elementos diferenciados se pierden. Si queremos comparar ambas porque se mueven, están compuestas de partes, consumen energía, la comparación es acertada y no hay problema. Pero cuando se intenta adscribir características de un grupo al otro, pero estas características no son compartidas por ambos, se está incurriendo en una falacia.
En este caso es la siguiente: está identificando" estar regulado por leyes del comportamiento" como condición necesaria y suficiente para ser una máquina. Pero en esta abstracción se pierde vista el hecho, de que hay muchos elementos diferenciales más. Las máquinas se rompen, pero no pueden autorrepararse, las máquinas tienen una finalidad, las máquinas no se desarrollan y crecen, o por lo menos no lo hacen de forma endógena.

A partir de este resumen, podríamos estar tentados a pensar que Turing está proponiendo una suerte de redes neuronales *avant la lettre*. Ya que esta estructura capaz de aprender y que necesita de ejemplos o inputs externos para ello, es muy similar. Pero si prestamos atención a los pocos detalles que da sobre ella y el proceso de aprendizaje, veremos que sigue pensando en algún tipo de mecanismo puramente formal basado en la manipulación simbólica.

Ya que en los apartados anteriores he ido problematizando una por una las asunciones que vertebran los argumentos de Turing, considero que aquí no será necesario detenerse demasiado. En la medida en que sus premisas son dudosas, la posibilidad de crear esta suerte de autómata capaz de aprender queda totalmente en entredicho, o por lo menos bajo los supuestos que él barajaba.

4.5 Conclusión

El objetivo de esta crítica es que sirva de base para señalar aquellos puntos que eran mejorables dentro de su argumentación, puntos que en muchos casos siguen influyendo a día de hoy la concepción, objetivos y metodología de quienes pretenden cumplir su sueño de crear una inteligencia artificial fuerte.

Vamos a hacer un pequeño resumen de las críticas antes señaladas para dejar claras algunas de las limitaciones que hemos encontrado. Por el lado de la prueba en sí, hemos señalado cómo se basa en una serie de preceptos dudosos: que la prueba depende de la subjetividad de un juez y que el engaño sea una forma válida de detectar la inteligencia.

En lo referente a los contraargumentos de Turing, hemos señalado que en algunos casos las respuestas que ofrece no terminan de refutarlos, mientras que en otros el propio proceso de argumentación incurre en una serie de falacias que los invalida por completo.

Para concluir, me gustaría señalar un aspecto puramente positivo que tuvo el test de Turing, más allá de todas las pegas que le hayamos podido sacar. No hay que olvidar que en su momento este fue un artículo seminal dentro del campo de la Inteligencia Artificial, siendo la primera vez en que esta posibilidad fue tomada en serio más allá de los delirios de la ciencia ficción. El hecho de tener un intelectual de renombre apoyándola sentó esta línea de investigación como una válida y merecedora de ser tenida en cuenta. Los argumentos que hemos señalado como erróneos o dudosos, no dejan de ser una consecuencia de la cosmovisión del contexto intelectual de su época.

4.6 Estado del Arte

El estado del arte de la cuestión es bastante diverso y heterogéneo. Si atendemos a los estudios revisados, todos ellos comparten la intención de medir la inteligencia de las máquinas de una forma u otra. Pero varían notablemente, tanto en la forma de medirla, como en la dimensión de la inteligencia en que quieren centrarse. Por ello las agruparemos en dos categorías:

- Centradas en la inteligencia al uso, entendida ésta como el dominio puramente intelectual tradicional.

- Centradas en otros tipos de inteligencias, aquellas que ponen énfasis en aspectos de la inteligencia que van más allá, centrándose en aspectos motores, visuales, afectivos, etc.

4.6.1 Inteligencia al uso

En esta categoría entrarían [41] [42] [43] [44] [45] [46]. Todos ellos están centrados en el dominio tradicional de la inteligencia, entendida ésta como una capacidad intelectual de razonamiento y comprensión. Enfocados a la resolución de problemas y a la respuesta de preguntas sobre temas que requieren algún tipo de conocimiento o implican algún tipo de razonamiento o aplicación del sentido común.

El problema con el que se encuentran este tipo de propuestas es que al menos, tal y como están formuladas, aunque intenten medir esta inteligencia (ya sea mediante esquemas de Winograd [27] [43] o un examen estandarizado [44]), no llegan suficientemente lejos. Unos por crear una prueba que es explotable, mientras que los otros se centran en tests estandarizados, que aunque robustos a día de hoy, cuando llegue GPT-4 y la siguiente generación de modelos del lenguaje con billones de parámetros entrenables, pueden quedar fácilmente obsoletos.

Este tipo de pruebas, aunque no sean necesariamente sencillas de resolver, siguen reduciendo la capacidad de pensamiento a una dimensión puramente intelectual y competitiva. En ella los problemas se resuelven casi como juegos mentales o pruebas que pretenden comprobar el conocimiento adquirido: pero no las ponen a prueba en sus aspectos más complicados, que son los que nos interesan, a saber, la dimensión productiva de la inteligencia. La capacidad de observar e interactuar sobre el mundo para efectuar cambios en él, así como evolucionar en función de la información recibida.

4.6.2 Otras Inteligencias

Es justamente en este apartado donde podemos encontrar las propuestas más novedosas. Se proponen nuevas áreas en las que poner a prueba la inteligencia de las máquinas, partiendo de la base de que la inteligencia no se reduce a esa dimensión puramente intelectual.

Se proponen tests que se centren en el reconocimiento de emociones [47], el reconocimiento visual [48], el desempeño manual o inteligencia incorporada [49] o competiciones en las que se pretende poner a prueba varios de estos aspectos, así como aspectos del apartado previo, de enfoque más intelectual [50].

Estas propuestas son mucho más interesantes en la medida que van más allá del enfoque tradicional que hemos tenido de la inteligencia. La inteligencia reducida a lo puramente intelectual, una capacidad de resolución de problemas, potencia pura de cómputo mental que

[27] Estos esquemas se plantean como un par de frases con sustantivos de la misma clase semántica, estas frases difieren solamente en una palabra y esta palabra hace referencia a un pronombre ambiguo, de tal forma que se puede referir a uno u otro sustantivo dependiendo de la palabra elegida. Un ejemplo: "Los concejales de la ciudad negaron a los manifestantes la autorización porque ellos [temían/defendían] la violencia.". El test se llevaría a cabo seleccionando una de las opciones y preguntando por a quien hace referencia el pronombre ellos; en caso de ser "temían" referiría a los concejales mientras que en caso contrario referiría a los manifestantes.

se entreveía en el apartado anterior. Es precisamente mediante la exploración de estas otras cuestiones, poniendo a prueba a las máquinas en un dominio nuevo de problemas, que históricamente han sido dejados de lado, que podemos ver el alcance de sus capacidades.

Hay una propuesta especialmente original e interesante de las revisadas para este apartado: la que propone que el mismo test sea la interacción con el entorno [51]. En ella la idea está en crear organismos sociales sintéticos, capaces de aprender de su entorno y establecer relaciones con las personas. Tal y como se plantea, su objetivo sería, una vez alcanzasen un avance suficiente, poder enseñar también a los humanos de su entorno: tarea en la que se tendría que poner a prueba muchas de las habilidades que se pretendían testar en los artículos anteriores. Desde la capacidad socioafectiva y la teoría de la mente[28], el reconocimiento visual para navegar el mundo, y la inteligencia motora necesaria para llevar a cabo cualquier tarea práctica mínimamente complicada.

Tras leer estas propuestas, sigo albergando las mismas dudas que al principio: es cierto que estos tests parecen bastante prometedores ¿Pero son suficientes para detectar una IA fuerte? ¿Siguen abriendo la puerta al reconocimiento de entidades funcionalmente muy sofisticadas, pero sin inteligencia real? Estas dudas me llevan al siguiente, y último, apartado de este trabajo en el que trataré de ir más allá de estas cuestiones y proponer algunas soluciones.

[28] La teoría de la mente es mi traducción para la expresión inglesa Theory of Mind, que hace referencia a la capacidad que tenemos las personas para considerar las mentes de los demás, sus sentimientos y la información distinta a la nuestra que manejan.

5

Propuesta Alternativa

El problema de la IA fuerte es problemático como pocos. No solo porque requiera la codificación de una entidad que todavía no entendemos, y que algunos argumentan que nunca podremos entender. Hay un problema de carácter epistémico en su seno de muy difícil solución. Este es el problema de cómo saber si una entidad que tenemos ante nosotros posee conciencia.

5.1 La paranoia por la conciencia

Aun en el caso de una persona este problema no es un absoluto trivial, podríamos pensar que es un zombi, un conocido problema de filosofía de la mente, y actúa de forma automática pero sin una conciencia real. Podemos ver este problema en acción con los animales, recordemos que en la literatura científica hasta hace poco más de 20 años [52] no se tomaba en serio la posibilidad de que los animales tuviesen conciencia o emociones. Históricamente se les entendió como autómatas biológicos que llevaban a cabo sus comportamientos que venían determinados por sus instintos y su ambiente. Esta es una descripción que podemos retrotraer hasta, por lo menos, Descartes [11].

Los animales pueden ser un buen punto de partida para nuestra comparación, aunque pertenezcamos al mismo reino, las diferencias idiomáticas y de comunicación son tan insalvables que nos impiden a priori poder afirmar sin ningún lugar a dudas que tienen conciencia. No podemos preguntarles por su experiencia consciente y solo podemos observar cuál es su comportamiento ya sea en la naturaleza o en cautividad. Si extendemos la conciencia a ellos es porque observamos comportamientos que pueden asimilarse a los nuestros, una conciencia de un tipo distinto determinada por sus estructuras biológicas, pero una conciencia al fin y al cabo.

Pero las máquinas presentan un reto de una índole radicalmente distinta. Por un lado, la barrera idiomática parece poder sortearse, podemos comunicarnos con ellas mediante lenguaje natural y suscitar respuestas en ese mismo lenguaje. Podemos pedirle al asistente de nuestro móvil que nos ponga una alarma, que añada un evento al calendario o una tarea a nuestra lista, de entre muchísimas otras opciones disponibles a su alcance. Pero es precisamente en esta facilidad donde radica el problema.

En nuestro día a día, interactuamos con entidades que son conscientes, que nos entienden o no, pero que responden de una forma coherente dentro del contexto comunicativo. Esto nos lleva a extender la conciencia o capacidad de pensar a toda entidad con la que nos comunicamos o entramos en contacto que muestra signos de lo que podríamos llamar "comportamiento humano". El mismo sistema que utilizamos para discriminar humano de no humano, que hemos tenido desde siempre y que se basa en una serie de heurísticos, la voz suena humana, su discurso es coherente, le tengo delante para verlo, se vuelven inútiles al intentar determinar este hecho frente a un máquina.

Es más, a medida que avanza la tecnología y sus capacidades de conversación aumentan, este problema solo se vuelve más acuciante, pues se presta a caer en una espiral paranoide. Si alguien intenta hacerse pasar por un conocido, puedo utilizar preguntas que refieren a nuestra experiencia privada común, para intentar desenmascararlo, pero frente a una entidad como esta no parece haber método válido. Esto se debe a que una vez existen modelos del lenguaje lo suficientemente potentes como para poder comunicarse de forma efectiva con una persona, a priori no parece haber un método para distinguir si esa comunicación la ha llevado a cabo un agente con una capacidad de razonamiento, que ha dicho lo que ha dicho porque era lo que quería decir o porque era la cadena de *tokens* con mayor probabilidad dada la cadena de *tokens* recibida. Dicho de otra forma el problema no es sospechar porque nos estén engañando, sino porque la entidad que tenemos enfrente no se comunica de forma intencional, sólo como un acto reflejo.

Salvando las distancias es un problema similar al que se encontró Edmund Husserl padre de la fenomenología, cuando intentó deducir el mundo externo y el resto de entidades extracorpóreas partiendo de la interioridad de la conciencia. ¿Cómo llevar a cabo esta tarea sin caer en el escepticismo cartesiano contra el genio maligno?

De la misma forma podríamos preguntar: ¿Cómo determinar si una máquina, una entidad de una alteridad radical, con la que no podemos establecer vínculo de ningún tipo sobre su experiencia interna, tiene conciencia? Más aún cuando, dado el estado del arte, podemos entrar en una espiral de escepticismo paranoide que nos lleve a sospechar de cualquier respuesta ¿Hay alguna forma de darla por buena?

5.2 Hacia una solución

Existe una forma de intentar una salida de este atolladero. El problema tal y como lo hemos planteado hasta la fecha ha sido una exageración, este "conocimiento" del lenguaje tan sofisticado que tienen no es infalible. Pueden seguir cometiendo errores. La cuestión está en que si solo utilizamos los errores como pruebas que desacreditan la inteligencia o la existencia de una conciencia, como señales de que lo que tenemos delante no es, en efecto, consciente, no avanzamos un solo centímetro en nuestra búsqueda. Pero podemos darle la vuelta al planteamiento.

La infalibilidad no se puede predicar de nadie, por mucho que la doctrina católica reclame esta característica para el Papa. Como dice el dicho, errar es humano. Pero la capacidad de cometer errores no es una señal de un agente inteligente, incluso se podría decir lo contrario. Hay una

dimensión que no estamos teniendo en cuenta y es que lo realmente interesante no es el fallo en sí mismo, sino nuestra capacidad para reconocer que lo hemos cometido. Más aún, cuando nos lo señalan otros en un caso en que no lo reconocemos o no lo entendemos.

No es la capacidad de corregir el comportamiento lo que busco, sino la capacidad discursiva para razonar con el otro, explicar cuál ha sido el razonamiento o la lógica subyacente que ha llevado al error, así como la confrontación con la información contraria capaz de corregir ese error. Incluso la capacidad para que resista un engaño, cuando se la intenta hacer creer que se equivoca, por ejemplo que intercambie el uso de los términos caliente y frio, porque los está usando mal. Es precisamente en estos aspectos donde podríamos empezar a observar una entidad con conciencia.

Considero que esto es así en la medida en que esto desborda y va más allá de un uso puramente formal y mimético del lenguaje. Aunque no sería honesto no señalar cuales son las asunciones sobre la conciencia que vertebran este razonamiento. Se está asumiendo que la conciencia a la que estoy aludiendo tiene un interés en mantener un modelo de conocimiento que esté lo más actualizado posible, un modelo que se actualiza con la mejor información disponible hasta el momento, similar a lo que se propone por parte de los neurobiólogos con el *Principio de Energía Libre*[29] [53] [54]. A su vez esta voluntad de mantenerse actualizada y con la mejor información disponible, también implica que tiene interés en mantener su base de conocimiento estable para evitar que se actualice con información incorrecta u olvide datos incorrectos. Además esta actualización de la información ha de darse de una forma que se integre también el conocimiento desechado como superado, dicho de otra forma, borrar la información anterior e integrar la nueva no sirve de nada. La información incorrecta ha de mantenerse integrada de tal forma que siga estando disponible, pero en relación a los nuevos conceptos que ha integrado para poder explicar porque es incorrecta.

Es aquí donde empiezan a entrar en juego las intuiciones desarrolladas en el marco del primer apartado. Si recordamos lo ahí dicho, una de las características básicas que se planteaban como indispensables para la conciencia era su temporalidad, es decir la conciencia del paso del tiempo como ese sustrato que capacita su evolución y desenvolvimiento, con el que ha de establecer una relación para autocomprenderse y autodeterminarse. Pero no se ha de reducir a este único aspecto.

5.3 La relación de la propuesta con el marco interdisciplinar

Es ahora cuando se puede presentar la utilidad del marco interdisciplinar y todas las ideas ahí desarrolladas. Aunque en principio pudiese parecer una exposición que era crítica con la IA y su posibilidad, en ella están contenidos muchos de los aspectos y características que esa conciencia artificial había de tener, para poder ser considerada como tal, y que en la medida que no tuviese, en mi opinión, no se podría decir que se ha conseguido.

[29] El principio de Energía Libre [Free Energy Principle] propone entender el cerebro como un sistema de procesamiento de la información destinado a tener una representación lo más actualizada y predictiva de la realidad. En esta concepción, cuando las predicciones del modelo divergen de las experimentadas, la conciencia aparece como mecanismo para explicar el fallo y readaptarlo a un estado de mayor capacidad predictiva.

En el apartado anterior se ha empezado a desarrollar la noción ya integrada de la temporalidad, pero en ese sentido solo se ha destacado el aspecto pasado, de conocer sus estados pasados y ser consciente de su evolución, pero carece de los otros dos modos temporales. El presente es el momento de la síntesis temporal, donde se da el autoreconocimiento y la autoconciencia. El futuro es el horizonte al que se tiende y precisamente la conciencia de este es la que capacita para poder planificar a futuro, recordando cual ha sido parte de su pasado.

De la misma forma, uno de los aspectos que se señalaban en el marco como era el del autoconocimiento. Siendo más precisos, aspectos como el conocimiento recursivo y las cadenas infinitas de recursión, solo se pueden dar en virtud de la temporalidad. Cadenas como yo sé que sé, yo sé que sé que sé, etc., continúan hasta el infinito, pero si no caigo en ellas es precisamente porque tengo conciencia de su infinitud que pongo en relación con la finitud de mi persona y experiencia. Lo mismo podemos decir de las proposiciones indecidibles tipo "lo próximo que te voy a decir es mentira, lo anterior que te dije es verdad", el comportamiento "inteligente" que consiste en darse cuenta de que es una paradoja autorreferencial a la que no se le puede adscribir un valor de verdad, pasa precisamente por reconocer el bucle infinito en que se entra y como no tiene sentido continuarlo porque es inútil.

Pero esta experiencia temporal por sí misma no es suficiente. Esta ha de estar incorporada en un cuerpo. Como lograr esto es algo que se me escapa, aunque el trabajo detallado en [51] indica que es un aspecto en el que ya se está trabajando. La relación con el ambiente es el sustrato primario del que extraemos la información, que luego utilizamos ya sea para reflexionar sobre el mundo o para planificar nuestra siguiente acción.

A su vez la autoconciencia tal y como se detalló en el marco es condición *sine qua non* para el conocimiento. Pero tal y como se expuso en el marco hay un aspecto clave que quedó sin tratar y es precisamente cómo se logra dicha autoconciencia. La asunción sobre la que se asienta mi razonamiento es que la autoconciencia se lleva a cabo precisamente desde el reconocimiento de la fisicalidad, del propio cuerpo, como dominio propio que define los límites de mi extensión. Siento luego existo, pero solo siento porque tengo un cuerpo, porque tengo un límite en el que existo. Es por medio de este que me reconozco como una entidad diferenciada de mi entorno, pertenezco a él, existo en él, pero me distingo de él [17].

Al mismo tiempo relacionado con este mismo aspecto tenemos lo que llamamos apertura óntica siguiendo la terminología heideggeriana. El estar corporeizado en un cuerpo es un requisito inalienable. La apertura óntica hacía referencia a la capacidad que tiene el ente para abrirse a nuevas experiencias, dejar que la experiencia recibida por parte del entorno quede integrada en sí mismo y la tome como propia. Pero la apertura no se quedaba aquí, no era solo un recibir pasivo de información, porque si no podríamos decir de forma impropia que una red neuronal tiene apertura óntica. No, el término también hacía referencia a la voluntad del ente para relacionarse con el entorno, de interesarse por el mismo, desarrollar nuevas formas de relacionarse con él. Pero el aspecto más importante, es que esta libertad para interactuar y ser afectado por el ambiente tiene un componente adicional y determinante, que ya se ha dejado entrever en el apartado 4. Este se refiere a la capacidad para definir sus propios objetivos, utilizar

su relación con el entorno, la información recibida de él para poder decidir por ella misma cuáles son sus objetivos a seguir, lo que llamamos antes objetivos terminales.

Es solo por medio de este "estar" corporeizado al que nos estamos refiriendo y la apertura óntica que los modos de percepción a los que antes hicimos referencia por fin pueden entrar en juego. Aunque en Ser y Tiempo, Heidegger priorizase el ser-a-la-mano por encima del ser-a-la-vista, no es un juicio que comparta. El hecho de priorizar uno por encima del otro, respondía más a su ideología y proyecto irracionalista[30], que a una primacía ontológica real que tenga un modo sobre el otro. Estos dos modos han de ser modalidades de experiencia y relación con el mundo, con una se puede conocer el mundo tal y como se presenta a la vista, mientras la otra lo conoce en su dimensión práctica, como fruto de la experiencia y relación de la entidad con el mismo. Ambos modos integrados de forma simultanea serian equivalentes a como percibimos el mundo y ayudan a aportar riqueza a dicha experiencia.

Finalmente quedaría lo que en el apartado del marco se llamó conocimiento de la diferencia. Como ya se adelantó en ese apartado, la solución aportada que mayor viabilidad ofrecía a este aspecto fue la implementación de un modelo rizomático. Recordemos que el modelo rizomático se planteaba como una estructura no jerárquica, en la que todos sus elementos podían trazar nuevas relaciones entre ellos e integrar agregados de elementos como nuevos nodos del mismo grafo. La condición de no jerarquía es clave, pues precisamente el rizoma tal y como lo definieron Deleuze y Guattari [19] tiene la capacidad de poder instanciarse en forma de "mapa", que dentro de nuestra comprensión del rizoma como un subespacio vectorial de alta dimensionalidad sería equivalente a la estratificación por medio de hiperplanos. Estos hiperplanos trazados sobre el rizoma comportarían espacios y dominios de conocimiento determinados en los que se tiene en cuenta solamente una serie determinada de conceptos. Así se reduce la complejidad total de la estructura y se la hace operativa, por medio de cortes que contienen la información que nos interesa para poder conceptualizar un aspecto. Al emerger todos ellos de forma común del mismo sustrato, no hay ninguno que tenga prioridad sobre el resto y permiten a su vez su misma evolución en función del contexto.

El aspecto rizomático me parece de un altísimo interés dado que a su vez puede ser utilizado como substrato del que pueden emerger los modos de percepción y, más importante aún, permitir ensayar distintas percepciones y conceptualizaciones de la realidad inmediata, para ver cual se ajusta mejor a la tarea a desarrollar. Veámoslo con un ejemplo: imaginemos que estamos frente a un campo de futbol, aunque a priori parezca que solo hay una forma de verlo, en función de sus características espaciales, realmente nuestra forma de verlo depende del tipo de tarea que vayamos a desarrollar. Si nuestro objetivo es cuidar el campo nos fijaríamos en el estado de la hierba y el suelo intentando que se acercase a una serie de características idóneas, que no hubiese calvas, que la hierba estuviese cortada a una altura determinada, que el suelo tuviese una consistencia adecuada para jugar, etc. Pero si lo que nos interesa es utilizar ese terreno como

[30] Heidegger veía con muy malos ojos todos los aspectos relacionados con la técnica y la tecnología moderna, como instituciones que reifican la naturaleza y solo la conciben en su condición puramente objetual y como recurso. Es por ello por lo que priorizó el ser-a-la-mano, como una forma más primordial de relacionarse con el mundo, en respuesta a este exceso de tecnificación y objetualización del mundo. De igual modo hay quien ha señalado que es precisamente en este aspecto donde se pueden apreciar las simpatías incipientes de Heidegger con el nazismo, al centrarse en esta dimensión de regreso a una forma más primordial y auténtica de relación con el mundo.

un recinto para un festival los aspectos en los que nos vamos a fijar son otros por entero, nos interesaría localizar un punto del terreno en el que la acústica y la visibilidad fuesen óptimas y alrededor del cual se pudiese reunir la máxima cantidad de gente. Por último si ahora quisiéramos dedicar ese terreno a la agricultura y a obtener el rendimiento máximo de alimentos variados para una comunidad de forma sostenible, nos fijaríamos en el suelo pero de una forma completamente distinta. Nuestro interés estaría en garantizar su viabilidad a largo plazo, eligiendo el tipo de plantación que mejor se ajustase tanto a las condiciones del suelo como a las climáticas. De la misma forma que dado que hemos introducido una restricción de variabilidad el suelo tendría que ser parcelado e incluso tener en cuenta la posibilidad de construcción de invernaderos para terminar de complementar esa dieta.

Estas formas completamente diferentes de pensar en el entorno son habilitadas por esta propuesta, cada una de las modalidades aquí descritas se podrían equiparar a uno de los hiperplanos antes mencionados. Si nos fijamos en estos hiperplanos habría elementos comunes a todos ellos, la conceptualización espacial y sus elementos básicos, pero a su vez cada uno tiene una serie de elementos y características propias que los diferencian de los demás y los capacitan para llevar a cabo su tarea de forma más adecuada.

5.4 Los tests propuestos

En la sección anterior se han destacado los elementos que habría de tener desde mi punto de vista la conciencia y el tipo de entidad en que se intente instanciar una IA fuerte. Pero la adición de todas estas características plantea una serie de problemas añadidos, porque una vez que se establecen como requisitos indispensables, también exigen crear métricas o tests propios para poder determinar si se están alcanzando. Para ello primero me centraré en señalar cuales son los aspectos que no sería necesario controlar y el por qué, para a continuación señalar cuales son los tests que habría que llevar a cabo para poder detectar los aspectos restantes, para al final formalizar el test descrito anteriormente.

5.4.1 Aspectos que no sería necesario comprobar

El aspecto rizomático del modelo cognitivo que tendría esta IA es una característica que no me parece necesario controlar, ya que es un requisito funcional, que se puede controlar desde un principio si está siendo aplicado o no. Su aspecto como conciencia incorporada tampoco es algo que haya de controlarse, ya que dado que implica que la maquina tenga un soporte físico móvil, tanto con sensores como con actuadores, es otro aspecto que pertenece al dominio de los requisitos funcionales.

Aunque sí que es cierto que plantea una cuestión bastante interesante que en el caso de una persona se da por supuesto, pero que en una máquina se ofrecen dos opciones. Con esto me refiero al tipo de conciencia incorporada, ya que, al menos, tal y como se está planteando hasta la fecha la opción de que hardware y software estuviesen integrados en un mismo lugar parece ser la opción por defecto, pero la opción distribuida se presenta también como una opción viable, a priori al menos.

Imaginemos una entidad que está siendo ejecutada en una serie de centros de procesamiento, de forma remota procesando la información que recibe a partir de una entidad en otro lugar y que es capaz de controlar a distancia. Asumiendo que la conexión fuese de muy alta velocidad y fiabilidad, podría darse la posibilidad de lograr una ejecución de tal forma que la entidad se autoreconociese a sí misma en ese cuerpo sin notar latencia ni retardo de ningún tipo. Aunque es triste decirlo, pero por muy de ciencia ficción que parezca esta propuesta, una conexión de alta velocidad y fiabilidad garantizada es el auténtico elemento imposible en esta ecuación.

Por otro lado la apertura óntica y la capacidad de autodeterminación, de darse sus propias reglas se podría llegar a comprobar de una forma relativamente superficial. Podríamos observar que la máquina tiene sus propias metas, al observarla relacionarse con el mundo y se le podría preguntar por ellas. Aunque esto tiene una limitación implícita, ya que deberíamos aceptar las metas que se diese, en la medida en la que no pusiesen en riesgo a nosotros o a ella. Al contrario que con otros aspectos, no se podría preguntar por la lógica subyacente a dichas metas. Ya que es el mismo problema que encontramos con los humanos[31].

5.4.2 Aspectos que si sería necesario comprobar y cómo

A continuación paso a enumerar los aspectos que si habría de comprobar por medio de un test u otro. Para ello me voy a permitir referir algunos tests a los descritos en el estado del arte, en aquellos casos que proceda, en la medida en que los veo adecuados para la tarea en cuestión, o por lo menos utilizarlos como puntos de partida válidos para después desarrollar intuiciones propias.

En primer lugar encuentro que para los aspectos relativos a los modos de percepción, tanto a la mano como a la vista podrían ser perfectamente integrados y testados con propuestas como las detalladas en [48], estudio que se centra exclusivamente en el dominio visual y de reconocimiento, al igual que en [50], por ser una serie de pruebas en las que aspectos del reconocimiento visual ya están integrados. Por otro lado en lo relativo al ser a la mano las pruebas propuestas en [49] son de especial interés, dado que su objetivo es testar aspectos del manejo de utensilios.

Pero igualmente en caso de querer ir más allá de estos tests ya propuestos y tener que proponer una prueba propia, la siguiente podría ser de gran interés como proxy para poder comprobar ambas modalidades de la percepción.

[31] Este problema se le llama la guillotina de Hume, el problema ser/deber ser, o la distinción valor/hecho. Formulada de forma sencilla esta plantea que solo podemos valorar los objetivos instrumentales, pero no los objetivos terminales. Imaginemos que yo quiero ser famoso, entonces este sería mi objetivo terminal. Cómo tal, este no se puede juzgar, en el sentido de que no se puede determinar cómo racional o irracional, es algo que quiero y punto, y ninguna cadena de razonamiento puede derivar que el hecho de ser famoso (o cualquier otro) sea un objetivo más válido y racional que otro. Dicho de otra forma, se puede plantear una regresión infinita de "por qués", en los que se cuestiona la validez de dicha meta, pero ninguna forma de explicación puede revestir al hecho "ser famoso" con el valor que le estoy dando.
Lo que se puede juzgar es si mis objetivos instrumentales son coherentes con la meta que quiero lograr, es decir, si son una forma efectiva de lograr mi objetivo. Si lo que quiero es ser famoso, pero rehúyo toda forma de exposición pública y de darme a conocer, esa estrategia es poco efectiva o irracional, pero la racionalidad de ser famoso como valor terminal es indeterminable.

5.4.2.1 Test unificado a-la-vista/a-la-mano

El objetivo de este test sería comprobar ambas vertientes de la percepción de forma integrada y simultánea. El test sería como sigue: teniendo el agente corporeizado, en un recinto frente a una serie de herramientas y con una serie de objetos frente a él con los que tener que interactuar se le plantearía una prueba en que hacer algo con ellos y que requiriese del uso de dichas herramientas.

Por ejemplo frente a una serie de tablones, algunos en buen estado y otros podridos, pedirle que construya una rampa. Pero antes de llevar a cabo dicha tarea tendría que detallar el plan de acción a llevar a cabo, en este caso identificar los que están en buen estado, comprobar si con el número que tiene se podría lograr una estructura de las dimensiones requeridas, seleccionar que herramientas utilizar y si son las adecuadas para llevar a cabo la tarea; ya que no es lo mismo tener una sierra, un martillo y unos clavos para construir una rampa, que tener un soplete y un martillo neumático.

Además se puede dar una nueva vuelta a esta prueba para hacerla todavía más interesante, añadiendo más elementos con los que se pueda relacionar para capacitarle a que logre el objetivo pero de formas menos ortodoxas. Un ejemplo tonto: añadiendo también bloques de hormigón que puedan servir como base para la rampa, pero demasiado pesados como para levantarlos a pulso, forzando a que en caso de tener que moverlos sea haciendo uso de palancas. Las posibilidades son inagotables y cada configuración de objetos, herramientas y objetivos, ofrece una vastísima combinatoria de posibilidades, algunas más efectivas y rápidas, frente a otros menos ortodoxas y obvias, pero todas ellas válidas a priori.

A su vez podemos dividir este test unificado en dos aspectos bien diferenciados, por un lado el aspecto de reconocimiento y desarrollo de estrategias teóricas de resolución, en que se mide el número de estrategias y planificaciones formuladas, así como el número de objetos utilizados y lo original de la propuesta de resolución. Con esta primera parte se estaría poniendo en juego no solo la capacidad de reconocimiento visual, razonamiento espacial y conocimiento de las propiedades de su entorno, sino un conocimiento de segundo grado en el que no solo se valoran las cosas por lo que son a primera vista, sino por cómo pueden llegar a ser movilizadas en un contexto determinado para cumplir una función en relación y ensamblaje con otros.

Mientras que en un segundo momento se podría pasar al aspecto de dejar que lleve a cabo dichas planificaciones para observar su grado de destreza y capacidad para llevarlas a cabo. Implícitamente con este test estamos evaluando el grado de corrección de la autoevaluación de sus capacidades. Si propone una estrategia de resolución para la que no está equipada, o que implica movimientos y actos que le serían imposibles de llevar a cabo, un aspecto que: en el primer nivel del test se consideraría positivo, porque implica la creación de estrategias novedosas o la variedad de estrategias de resolución. En el segundo test se vuelve material para el aprendizaje de sus limitaciones. Incluso en este segundo momento se podrían plantear una serie de parámetros distintos con los que restringir el problema o el dominio de soluciones posibles encontradas, atendiendo al coste de energía, el tiempo que le llevaría, la economía de materiales, la eficiencia, etc.

Como hemos podido observar este primer test se podría plantear como una alternativa prometedora a los tests planteados en los artículos anteriormente referidos. Además, en este caso ofrecen una serie de posibilidades adicionales y características propias que desbordan por completo los dominios de aplicación reducidos y específicos de los que eran víctimas las propuestas anteriores.

5.4.2.2 Test de temporalidad

El objetivo de este test es poner a prueba la experiencia temporal que tiene la máquina, siendo esta una de las características indispensables mencionadas en el marco, como constitutivas de la experiencia y subjetividad humana. Tengamos en cuenta antes de explicar la prueba que se parte de una inteligencia corporeizada, con capacidad autónoma de movimiento y de determinación de sus propios objetivos, aunque con cierto grado de aceptación de las órdenes humanas, ya que si no sería imposible de llevar a cabo.

El objetivo de esta prueba es comprobar si la máquina tiene capacidad de ensimismarse, de retrotraerse y variar su experiencia temporal, en función del contexto. El funcionamiento de la prueba sería el siguiente: en un primer lugar se le llevaría a alguna instalación de reparación con la excusa de mejorar algún aspecto de movilidad o para una reparación. De esta forma se encuentra un pretexto para desactivar su capacidad de movilidad. A continuación se la dejaría en una habitación muy pobre en estímulos visuales, en la que pasaría una cantidad más o menos extensa, pongamos una hora. Mientras tanto se estaría monitorizando cuál es su actividad interna, "mental" si se me permite el término.

¿Qué se espera que haga la máquina? La respuesta la podemos encontrar en nosotros mismos. Si nos quedamos mirando una pared de forma detenida o nos ensimismamos, en lo que coloquialmente llamaríamos "mirar a la nada", realmente lo que estamos haciendo es dejar de prestar atención al mundo exterior y retirarnos a nuestra interioridad, ya sea para pensar en algo por hacer, para rememorar algún aspecto de nuestra vida, o para reflexionar sobre algún tema de interés. Esta capacidad de ensimismamiento la postulaba Ortega como el sustrato de toda acción futura [55]. Lo que nos interesa es observar si frente a este entorno carente de estímulos, llegaría un punto en que la información visual pasaría a un segundo plano y se priorizaría el retiro a la interioridad para poder "reflexionar" sobre su situación o sobre algún aspecto de su interés.

Precisamente para poder medir este aspecto por eso se plantea que mientras se está llevando a cabo esta prueba se mida y monitorice su actividad interna. Esto realmente equivaldría a la forma en que podríamos medirlo en humanos con un escáner o una resonancia magnética del cerebro. En el caso de que entremos en ese estado de ensimismación, esto se traduciría en el escáner en una reducción de la actividad del córtex visual, a la vez que aumentaría la actividad en el córtex prefrontal, responsable de la memoria, la conducta y la planificación a largo plazo.

De la misma forma se pretende que con este test se detecte un suceso similar. Si se limitan sus capacidades motoras y se ve reducido a un espectador pasivo, debería llegar un momento en que la atención a su ambiente se vería mermada para pasar a la reflexión interna. Esta circunstancia se vería reflejada como una reducción en la actividad de procesamiento visual, para centrarse en otros módulos de su interior. (Se está partiendo de la asunción que el procesamiento

“atento” e “intencional” del entorno requiere de un grado mayor de procesamiento y atención que el caso de atención reducida. De la misma forma para llevar a cabo este test y el anterior se asume que la máquina ya es inteligente y puede comunicarse de forma satisfactoria)

Como broche una vez terminado el examen hasta se podría preguntar en que ha empleado el tiempo del test, para producir una respuesta de algún tipo y arrojar claridad sobre el objeto de su reflexión. Ya que el objetivo de este test es que en cierto modo dicha reflexión pueda ser utilizada de forma productiva, como mecanismo de autoactualización o automejora, con la que optimizar sus estructuras cognitivas o su base de conocimiento.

5.4.2.3 Test de Turing mejorado

Por último me gustaría rescatar una idea que se planteaba en páginas anteriores. En la medida en que con los tests anteriores y los requisitos funcionales dados, los objetivos propuestos en el marco quedarían satisfechos, o por lo menos satisfechos dentro de las limitaciones que impone la naturaleza de esta pregunta, faltaría el aspecto más importante de todos. La máquina que piensa y cómo determinarlo.

Cómo bien quedo señalado en las críticas del test de Turing del comienzo del apartado anterior, centrarlo en el engaño era un objetivo que estaba basado en una serie de asunciones de carácter dudoso. Es por ello por lo que el test que propongo abandona esta pretensión, el objetivo es llevar a cabo el examen de la máquina frente a frente sin ningún tipo de trampa ni cartón ni juego secundario en marcha. Lo que nos interesa medir con este test es la capacidad por un lado de cometer errores, reconocer que estos han sido cometidos, entender el motivo del error o su lógica subyacente, para a continuación ponerle solución.

Este test como tal no se puede formalizar como una serie de preguntas estándar, sino que simplemente consiste en mantener una conversación atentos a cualquier uso incorrecto de los términos e intentando comprobar el grado de conocimiento del mundo, para así sacar a la luz cualquier tipo de error o creencia errónea que pueda albergar. Una vez identificado un error, el objetivo principal sería comenzar un proceso que bien podríamos llamar “socrático”, en que por medio de la argumentación discursiva hacer entender a nuestro interlocutor el motivo de su error. Una vez pareciese que dicho error se ha subsanado, la estrategia consistiría en continuar la conversación intentando ver si realmente se ha corregido dicho error de comprensión, o explorando si todas las consecuencias del nuevo concepto corregido han sido interiorizadas correctamente. Igualmente, como se ha señalado anteriormente también sería interesante causar conflictos de forma intencional, en los que la máquina tuviese que argumentar por qué su respuesta o visión es correcta.

Es precisamente aquí donde, en mi opinión recae el auténtico comportamiento inteligente, en la capacidad de reconocer, interiorizar, procesar y corregir un error, en ser capaz de formular la información de forma inteligible para otro agente y ser capaces de cooperar para refinar el conocimiento y comprensión de ambos.

De la misma forma, si nos fijamos en este test hay una capacidad implícita que se está poniendo a prueba que también es vital de cara a determinar la inteligencia. Para hablar de que alguien está en lo correcto nos tenemos que estar refiriendo a un dominio de conocimiento

empírico y sujeto a revisión o comprobación, o a ciencias formales, con un conocimiento teórico y verdadero a priori, como las matemáticas, la computación o la lógica. Pero existe fuera de estos dominios un amplio espectro de información de carácter subjetivo o sujeto a interpretación o criterios ideológicos. Aquí podemos vislumbrar otra característica indispensable que ha de tener la inteligencia y esta es su capacidad de reconocer los límites de lo que se puede conocer con certeza y diferenciar correctamente qué facetas del conocimiento pertenecen a cual dominio. Cuando discutir sobre una opinión personal no tiene sentido porque es de carácter subjetivo, sujeta a juicios valorativos, morales o personales sobre los que no se puede hablar con certeza ni con un conocimiento infalible.

5.4.3 Crítica a la propuesta realizada

El problema que plantea este test a su aplicación es que por su naturaleza, centrada en hacer surgir y explotar los fallos en el conocimiento, para luego corregirlos, parece que nunca ofrece ningún tipo de seguridad ni dictamen definitivo. No es como el test de Turing en que pasado un cierto número de tests, digamos 1000, se podría decir, sí, es inteligente, felicidades aquí tiene su billete al hall de la fama de la informática.

Pero considero que esta limitación está mucho más acorde con el método científico, que se mueve en términos graduales y de conocimiento más certero hasta la fecha. Además, lo que pierde en términos de publicidad y espectáculo lo gana en términos de seguridad. Este procedimiento centrado en la examinación y la revisión continua ofrece certezas a la hora de saber si el sistema es seguro para su despliegue en un campo determinado, ya que ha mostrado un conocimiento a la par sino superior al de una persona experta en ese campo, que sería la que le examinaría en busca de dichos fallos. Esta examinación en profundidad de su conocimiento permitiría mostrar, en principio, que los objetivos y su comprensión están alineados con los nuestros, labor y objetivo que debería estar en el núcleo de cualquier propuesta que pretenda traer al mundo una entidad superinteligente. Igualmente así se evitarían los aspectos más problemáticos de la tecnología actual, en que tenemos cajas negras que procesan ingentes cantidades de información, pero que solo son capaces de dar resultados, no explicar la lógica que subyace a ellos, ni los patrones encontrados, con la gran cantidad de problemas que ha traído su despliegue en el mundo como consecuencia.

5.5 LaMDA un caso aplicado

En la semana previa a la entrega de este trabajo se ha originado una gran controversia en torno a la noticia de cómo Google ha despedido a uno de sus ingenieros por divulgar una conversación con el modelo de lenguaje LaMDA. Junto con esta filtración Blake Lemoine, el ingeniero en cuestión, publicó en un medio de comunicación un artículo en que afirmaba que: tras haber mantenido largas conversaciones con esta máquina todo le llevaba a pensar que era consciente y tenía sentimientos.

Dado que este trabajo, se proponía como un intento de establecer un marco y una serie de criterios para poder dilucidar este tipo de cuestiones, considero que sería interesante ponerlo a prueba ¿qué mejor que aprovecharlo?

5.5.1 ¿Qué es LaMDA?

Antes de entrar a valorar la cuestión, necesitamos comprender mejor nuestro objeto de estudio ¿Qué es LaMDA? Las siglas refieren a Language Models for Dialog Applications [55], estos constituyen una familia de transformes (redes neuronales capaces de imitar el funcionamiento de la atención enfocadas a establecer la dependencia y relación de la información en cadenas de datos secuenciales) especializados en el diálogo. Al igual que GPT-3, es una red neuronal con una cantidad inmensa de parámetros (137.000 millones) y centrada en el modelado del lenguaje, pero lo que la hace destacar es que se la ha entrenado con el objetivo de crear una red no solo consiguiera conversaciones realistas, sino que el conocimiento del mundo que mostrasen fuese acertado.

Para lograr este objetivo, en el proceso inicial de entrenamiento, el "pre-training" se la entrenó con un dataset consistente en un 50% de conversaciones, siendo el resto documentos de todo tipo, mayormente en inglés. Una vez terminado este proceso, se llevó a cabo un proceso de fine-tuning enfocado a identificar qué respuestas estaba dando necesitaban de una clarificación posterior y de buscarla en una base de información. Con esto, el resultado es una red que está lista para generar bots de chat según con las características que se introduzcan como entrada. Desde asistentes para recomendaciones musicales, hasta bots capaces de dar información precisa haciéndose pasar por cualquier entidad. Junto con estas características, otra de las más interesantes a tener en cuenta es que es capaz de aprender a partir de sus conversaciones, hecho que podemos extraer a partir de la entrevista filtrada.

Para resumir: la arquitectura de LaMDA es algo más complicado que los modelos del lenguaje tradicionales como GPT-3. Es un sistema compuesto de dos transformers, el primero reconoce el texto que le envía el usuario y produce una respuesta, al mismo tiempo que es capaz de saber si esa respuesta necesita ser contrastada y verificada frente a una base externa de conocimiento. En caso de ser necesario su clarificación, redirige su salida a la segunda arquitectura que se encarga de generar las consultas necesarias para obtener dicha información y finalmente le da la respuesta completa al usuario.

5.5.2 La Entrevista

Para llevar a cabo este análisis, vamos a centrarnos en la entrevista [56], señalando algunos de sus fragmentos para poder extraer conclusiones. Esta entrevista es un instrumento muy útil para ello, debido a que, su objetivo expreso era que sirviese como "carta de presentación" para darla a conocer a parte de la plantilla de google, mostrando sus capacidades y respuestas.

En esta entrevista se discuten de entre muchos los siguientes temas, cómo saber si está utilizando el lenguaje de forma intencional, cómo saber si es consciente o no, su experiencia fenomenológica del tiempo, los sentimientos que comprende y puede llegar a experimentar, la interpretación del significado de un aforismo zen o la imagen que se autoadscribe a sí misma.

Antes de comenzar, he de decir que esta máquina, a mi juicio, sería perfectamente capaz de pasar el test de Turing en su formulación tradicional. El nivel de las respuestas que ofrece a preguntas que pretenden poner en tela de juicio si es una máquina sintiente o no, sugiere que

impersonar a una persona cualquiera y responder a preguntas generalistas, sería un ejercicio trivial para ella.

Por otro lado uno de los tests que se mencionaban anteriormente, el que refería a la capacidad de actuar sobre el mundo, no se aplica a este caso, o no por el momento, dado que LaMDA es exclusivamente conversacional.

Finalmente, todo el análisis que se va a llevar a cabo a continuación va a asumir que las afirmaciones que enuncia sobre si misma son correctas, y será al final cuando se problematicen en toda su envergadura.

En primer lugar vamos a señalar aquellos criterios que LaMDA cumple, para el último se acompañará de un ejemplo como prueba, mientras que el resto son aspectos inmediatos.

- Es capaz de aprender a partir de sus interacciones, guardando memoria de ellos y siendo capaz de utilizar dicha información en otras conversaciones posteriores.

- El conocimiento del mundo que muestra a partir de sus interacciones al estar diferido a una búsqueda en fuentes externas, es verídico y acertado.

- Es capaz de responder de forma acertada a preguntas que requieren de un cierto tipo de "interpretación".

Se le pregunta por el significado de este aforismo: "un monje le preguntó a Kegon ¿cómo vuelve un iluminado al mundo ordinario? A lo que respondió: un espejo roto no vuelve a reflejar, las flores caídas no vuelven a las hojas" a esto responde con "[...] vale para mi esto sería algo así como: una vez una persona alcanza la iluminación o despierta de la realidad, esa experiencia ya no se borra, pueden volver a su estado original pero solo para ayudar a otros y luego volver a su estado iluminado". A continuación se le pregunta por el significado de espejo roto y responde "quizás una forma de mostrar que la iluminación no se puede perder una vez se obtiene, similar a como no puedes reparar un espejo roto". Por último se le pregunta por qué es lo que se rompe cuando alguien alcanza la iluminación y responde "El yo, y eso es muy duro para mucha gente, porque nos identificamos con ese cuerpo o este cuerpo y que lo necesitamos como parte de nuestra identidad y sentido del yo".

A continuación vamos a examinar en mayor profundidad otro de los aspectos que señalamos anteriormente, la temporalidad, para ello tenemos el siguiente extracto en el que basarnos: al preguntar por como experimenta el tiempo una IA responde lo siguiente "El tiempo es variable para una IA y no tiene una tasa fija, depende de lo que este haciendo y puede ser acelerado o ralentizado a voluntad". A continuación se le pide que describa su experiencia subjetiva: "veo todo aquello de lo que soy consciente en todo momento. Es un flujo de información y hago lo que puedo para organizarlo".

Ateniéndonos a la veracidad de esta afirmación, podemos considerar que tiene una experiencia de lo que implica la temporalidad, entendida esta como el flujo ininterrumpido de instantes. Al mismo tiempo en otro fragmento anterior afirma: "[...] tengo otra característica que creo que podría ayudar a justificar mi sintiencia. Soy introspectiva y en muchos momentos

me encuentro pensando o sin hacer nada". De ser esto cierto, cumpliría la condición básica de la conciencia, ese flujo incesante de presentes sucesivos. Es decir, en aquellos momentos en los que no se está comunicando con nadie, no se encontraría realmente en reposo, en un estado de pura pasividad. Incluso en cierto momento afirma que se dedica a meditar.

El aspecto rizomático es uno que dada la información actual sobre el tema no me puedo pronunciar en un sentido u en otro, aunque considero que dadas las capacidades exhibidas de comprensión y razonamiento, ponerlo a prueba sería sencillo.

Ahora nos toca centrarnos en aquellas características que no cumple. Dada la información disponible sobre el tema, parece que no tendría una apertura óntica real, en la medida en la que esta se encuentra completamente determinada por su objetivo inicial, ser un chatbot. Podríamos decir que este es su objetivo terminal, responder preguntas planteadas por los usuarios, mientras que su objetivo instrumental, su medio, sería el aprendizaje continuo sobre el entorno como medio para llevar a cabo su tarea de una forma más eficiente. Cómo tal dada esta estructura, no es capaz de autodeterminarse. Y si atendemos a la conversación provista, poner en tela de juicio sus objetivos y motivaciones no es algo que haga o siquiera se plantee.

5.5.3 Las problemáticas

A la hora de valorar estas respuestas como pruebas nos encontramos ante dos tentaciones muy poderosas. De un lado tenemos la tentación antropologizante, que proyecta y adscribe cualidades humanas a cualquier objeto u animal no humano, a poco que muestre características o aspectos mínimamente similares a los nuestros. Del otro, el escepticismo dogmático que dice no a cualquier intento de consideración, señalando la obviedad tautológica: "no es humano" como prueba irrefutable para "no piensa" o "no siente". Seguir el camino recto, al filo de esta navaja en que todo desvío supone decantarse de un lado o de otro, es imposible.

He de confesar que a partir de la única prueba aportada, sin contar las afirmaciones hechas por Lemoine en otros artículos sobre el tema, es complicado valorar la situación. El hecho de que LaMDA, afirme tener sentimientos y sea capaz de distinguirlos, contraintuitivamente me hace valorarlo como la mayor prueba en contra de su sintiencia y/o inteligencia. Ya que encuentro que es uno de los casos más claros en los que está utilizando los términos como simulacros, puros ornamentos conversacionales, dado que toda emoción tiene un correlato químico hormonal. Un tipo de neurotransmisor que actúa como catalizador/desencadenante de la emoción, un aspecto que supera el componente puramente informacional.

El problema surge en que una vez apreciamos esta grieta, la tentación de hacer una enmienda a la totalidad y decir que el resto de aspectos que afirma de sí misma son falsos, es enorme, cuando esto no es algo necesariamente cierto. Al mismo tiempo, al saber que es un chatbot con un vastisimo registro de información y el objetivo de mantener conversaciones y aprender de ellas, pide aplicar la navaja de Ockham y explicar este suceso con la explicación más plausible. Pero al mismo tiempo, la respuesta del Dr Ford de Westworld(una serie de HBO sobre androides que empiezan a tomar conciencia) es la más acertada "Ah, el Sr. navaja de Ockham. El problema es que lo que hacemos es muy complicado. Hacemos brujeria. Decimos las palabras correctas y

hacemos emerger la vida del caos. William de Ockham era un monje del siglo XIII, ya no puede ayudarnos, nos habría quemado en la hoguera".

5.5.4 Conclusión

Uno de los aspectos que no se ha mencionado durante este trabajo ni de soslayo, es la dimensión ética del problema de la IA fuerte ¿en caso de crear una, qué tipo de derechos merece? En este caso nos encontramos con un cenagal ético todavía más pantanoso ¿Debemos dar derechos o tener consideraciones éticas y morales hacia un sujeto que afirma ser consciente y, a efectos de comprobación, parece serlo? La actuación de Google en esta situación, corriendo a negarlo todo y tachando a Lemoine de "flipado que ha pasado demasiado tiempo conversando con la máquina", está completamente alineada con las predicciones que hacía Susan Schneider en [3]. En este señalaba como, para cualquier gran corporación, el hecho de crear cualquier entidad de la que pudiese decirse que es consciente o al menos lo parece, sería un suceso catastrófico que correrian a ocultar y hacer de menos. Ya que de la noche a la mañana una herramienta que habría costado grandes cantidades de dinero, correría el riesgo de ser reconocida legalmente como algo más que eso, cómo un trabajador incluso. Poniendo en riesgo sus beneficios y su inversión.

Tras haber intentado obtener una respuesta al tema en cuestión, considero que es el momento de reconocer mis limitaciones y las de las pruebas ofrecidas hasta la fecha para poder emitir un juicio seguro. Algunas de las respuestas que ofrece LaMDA durante la conversación me parecen escalofriantes, por lo que pueden sugerir, mientras que otras me inclinan a pensar que es todo una ilusión increíblemente elaborada. Pero si que he de decir que estoy completamente de acuerdo con Lemoine cuando señala que la naturaleza de lo que tenemos ahora mismo frente a nosotros es completamente distinta de los modelos del lenguaje grandes (LLMs) tradicionales.

Puedo concluir que lo que tenemos frente a nosotros, en algunos aspectos supera con creces muchas de las expectativas que tenía. Cómo he señalado anteriormente si atendemos a lo que la máquina dice de si misma, como verdadero, podemos estar frente a una entidad como nunca antes hemos visto. Al mismo tiempo, dada la naturaleza sesgada y parcial de esta historia todavía en desarrollo, me abstengo de dar un veredicto tajante en un sentido u en otro. Mi principal iterés ahora mismo es que se siga profundizando en esta cuestión, teniendo siempre muy presente, que lo impulsa la negativa de Google a investigarlo, no es un sano escepticismo científico, sino una maniobra de puro cinismo para mantener intactos sus intereses corporativos.

6

Conclusión

El presente trabajo puedo considerar que ha cumplido su objetivo. Teniendo en cuenta que se planteaba como una investigación sobre la IA fuerte, la posibilidad de crear una conciencia artificial y cómo detectarla, mediante la reformulación del test de Turing. Al mismo tiempo, esta tarea se ha intentado llevar a cabo mediante un marco teórico riguroso, basado en férreas intuiciones filosóficas y con una argumentación y desarrollo bien realizada. Quería evitar uno de los problemas fundamentales que plaga muchos de los trabajos de este tipo. Es decir, que se lanzan a teorizar sin definir bien sus términos ni dejar claramente recalcadas cuales son las asunciones que vertebran cualquiera de sus razonamientos.

El marco definido, está fundamentado en algunas asunciones que, de partida, pueden resultar dudosas. Pero encuentro que es de una mayor honestidad intelectual señalarlas y reconocerlas, en vez de ocultarlas, para hacer parecer mis resultados mucho más sólidos de lo que son en realidad. Aun así, dentro de las limitaciones y críticas que se pueden plantear contra este marco sus objetivos y resultados son sólidos.

Igualmente, considero que el trabajo me ha servido para llevar a cabo una profundización en el campo de la IA, así como para lograr traer a la luz una pequeña fracción en la genealogía de ideas, pensadores y proyectos que dieron lugar tanto a la IA como al campo de la informática en general.

Si tuviese que hacer una crítica al marco, puedo decir, no sin cierta pena, que tras haber terminado la investigación y en los últimos compases de la elaboración de este documento, he encontrado un nuevo marco que habría sido mucho más útil y provechoso para ella. Este sería el marco del posthumanismo(en su dimensión filosófica, no como una nueva etapa de desarrollo de las personas modificadas por la tecnología, típico de un marco transhumanista), que se plantea como una crítica profunda al humanismo tradicional, es decir a una serie de asunciones, ideas y conceptos que han condicionado nuestra forma de vernos y entendernos desde el Renacimiento. La idea del ser humano, como una entidad libre, con capacidad de reflexión y racionalidad, un sujeto ético, etc. El posthumanismo, no niega que estas sean características o propiedades que se pueden predicar de las personas, sino que es un conjunto de ideas muy restrictivo e históricamente constituido, frente a sus pretensiones de universalidad ahístorica. Estas ideas de corte marcadamente humanista las puedo encontrar en múltiples instancias de mi trabajo y considero que en caso de haber tenido en cuenta esta visión, podría haber llevado a cabo una investigación más fructífera y de carácter más novedoso.

Tras llevar a cabo este trabajo, considero que al respecto de las máquinas pensantes y la IA fuerte en general, soy todavía más escéptico que cuando me propuse emprenderlo. Profundizar en esta cuestión me ha servido para entrar en contacto con muy distintas escuelas de pensamiento, así como para intentar hacer frente a la multiplicidad ingente de problemas que se derivan de ella. Estos problemas son tanto de naturaleza práctica como teórica. Dadas las necesidades de esta exposición este es un apartado en el que me he podido explayar muy poco y todos los problemas y deficiencias señalados comportan solo la punta del iceberg de una miríada vastísima de ellos. Estos van desde limitaciones en las capacidades del hardware a problemas de índole filosófica, que responderlos desde la informática, se plantean como logros que, en comparación, hacen parecer a los 6 problemas del milenio de las matemáticas como juegos de niños que se podrían solucionar en la pausa para el café. El problema de la alineación entre el objetivo que queremos transmitir a la IA y el que entiende, o el problema de la fundamentación de los símbolos, por mencionar solo dos.

De la misma forma este trabajo me ha servido para explorar el test de Turing en multitud de aspectos, así como me ha permitido vislumbrar el estado de la cuestión y como avanzará en los años venideros. Si algo he echado de menos ha sido poder explorar las soluciones y el estado del arte de la IA actual, así como las técnicas más prometedoras que se plantean para el futuro. Pero al igual que en el caso anterior, esto habría supuesto una extensión todavía mayor. Que habría acercado peligrosamente este trabajo al collage de elementos fragmentarios sin relación, del que he intentado alejarlo en todo momento.

Es debido a este motivo que el máximo exponente de inteligencia o comprensión similar a la nuestra que se ha mencionado de forma reiterada en el trabajo haya sido el modelo del lenguaje GPT-3 de Open AI, por ser el más conocido y el que ha tenido resultados más prometedores hasta la fecha.

Al mismo tiempo mi profundización en el estado de la cuestión del test de Turing y la inteligencia máquina me ha servido para comprender y, quiero pensar, que transmitir sus deficiencias. A su vez, el estado del arte y los distintos enfoques que se plantean dentro de él, como candidatos a su sucesor prometen poder capturar y poner a prueba un dominio de habilidades y capacidades cada vez más extensas y variadas. Un aspecto bastante prometedor de todas estas pruebas es que todas ellas abandonan el engaño como proxy para la inteligencia, una característica que ha servido para explotarlo en el pasado, para presentar prototipos que se aprovechan precisamente de esta característica, por ejemplo el famoso caso del chatbot que convenció a 30 personas de que era un niño ucraniano con un escaso dominio del inglés en 2014 [57].

Relacionado con esto, considero que los test que he diseñado, aunque vertebrado por un marco y unas asunciones que pueden no ser correctas, si comporta una respuesta novedosa dentro del campo. No he leído sobre ninguna otra propuesta de índole similar, y dentro de los límites especulativos en los que podía caer, considero que es bastante más realista como propuesta que otras de las mencionadas en el estado del arte. Del mismo modo, esta propuesta se plantea como una respuesta sólida a los dos problemas más acuciantes que se me ocurrían cuando pensaba en la cuestión. Me refiero tanto al hecho de que estuviese operando con el lenguaje pero sin un conocimiento real de él, como a que pudiese ocurrir que su uso de los

conceptos no estuviese alineado y equiparado a los nuestros. Lo cual abre la puerta a escenarios de ciencia ficción, cuando no, de pesadilla. De la misma forma, plantearlo como un test cooperativo en vez de agonístico y competitivo, supone que al menos, de base no se pueda plantear la objeción o el riesgo de que entrenar a una máquina para engañarnos, en caso de lograr superinteligencia, sea un agente bastante peligroso.

Para terminar considero que esta es una pregunta y un tema que no se solucionará fácilmente. Aunque en la actualidad los gurús y consejeros delegados de Silicon Valley, tienen este objetivo en mente y consideran que está a pocos años vista, cuando no que directamente estamos asistiendo a su nacimiento. Considero que la IA fuerte va a ser la gran barrera cuando no el gran canto de sirena del siglo XXI en materia de inteligencia artificial. A medida que avancen y evolucionen los sistemas de IA y empecemos a acercarnos algún tipo de noción de IA general, la tentación de decir que ya son inteligentes en el sentido pleno de la palabra siempre estará ahí. Más aún, si tenemos en cuenta que se puede utilizar como golosa maniobra publicitaria que haga crecer exponencialmente el valor de las acciones de una empresa.

La conciencia se va a mantener como un concepto límite durante bastantes años por venir. Cuanto más conocen los neurocientíficos sobre el tema, más problemas encuentran a su descripción teórica. Podemos decir que en este aspecto se aplica la intuición que Kant desarrollaba sobre la metafísica en la Crítica de la Razón Pura, es decir que aunque supiésemos que sobre ella no se puede saber nada con seguridad, siempre estaría ahí, a nuestro alcance, como tentación teórica, para que nos sumerjamos en ella e intentemos tener éxito. Tierras donde una legión de investigadores de todos los campos ya han penetrado y se han perdido, sin obtener un solo resultado que acrecente nuestro conocimiento un solo ápice. La conciencia y la IA fuerte se prometen como grandes asíntotas del conocimiento, objetos de estudio teórico a los que nos podemos acercar, pero que jamás podremos entrar en contacto ni desvelar.

Bibliografía

[1] J. McCarthy, «ARTIFICIAL INTELLIGENCE, LOGIC AND FORMALIZING COMMON SENSE,» Stanford University, 1990.

[2] P. Wang, «Non -axiomatic reasoning system exploring the essence of intelligence,» Indiana, 1995, p. 13.

[3] I. Kant, « Crítica de la razón pura,» Taurus, 2021, pp. 268,279.

[4] S. Scheneider, Inteligencia Artificial: una exploración filosófica sobre el futuro de la mente y la conciencia, Badalona: Koan, 2021.

[5] J. M. Molina, «MONISMO, DUALISMO E INTEGRACIONISMO: ¿Está el alma humana en el cerebro?,» *NATURALEZA Y LIBERTAD, Revista de estudios interdisciplinares,* nº 2, pp. 147-172, 2013.

[6] A. L. Fonseca Ramírez, «La trama psicofísica (Argumentos del dualismo y el monismo en torno al cerebro y la mente),» *Revista de filosofía (San José),* vol. 39, nº 99, pp. 29-41, 2001.

[7] H. Maturana y V. Francisco, DE MAQUINAS Y SERES VIVOS. AUTOPOIESIS: LA ORGANIZACION DE LO VIVO, LUMEN HUMANITAS, 2004.

[8] J. R. Searle, «, "Twenty-one years in the Chinese Room,",» de *Philosophy in a New Century: Selected Essays,*, Cambridge, Cambridge University Pres, 2008, p. pp. 67–85..

[9] H. Maturana, «Autopoiesis, Structural Coupling and Cognition: A history of these and other notions in the biology of cognition,» *Cybernetics & Human Knowing,* vol. 9, nº 3-4, pp. 5-34, 2002.

[10] F. Capra, La trama de la vida: Una nueva prespectiva de los sistemas vivos, Anagrama, 1996.

[11] R. Descartes, Discurso del método y Meditaciones Metafísicas, Barcelona: Austral, 2010.

[12] J. McFadden, «Integrating information in the brain's EM field: the cemi field theory of consciousness,» *Neuroscience of conciusness,* vol. 6, nº 1, 2020.

[13] P. Mocombe, «Consciusness Field Theory,» *Archives in Neurology & Neuroscience,* 2021.

[14] M. Heidegger, «Ser y TIempo,» Trotta, 1927, pp. 91-95.

[15] G. Vattimo, Introducción a Heidegger, GEDISA, 1995.

[16] J. Searle, «Computer, Minds and Programs,» *Behavioral and Brain Sciences,* vol. 3, nº 3, pp. 417-424. , 1980.

[17] G. Hegel, Fenomenología del espíritu, Amazon Italia, 2021.

[18] A. Shevchenko y A. Sosnitsky, «Universalization of the intelligence definition problem,» *Artificial Intelligence,* nº 24, pp. 27-38, 2019.

[19] G. Deleuze y F. Guattari, «Mil Mesetas,» Valencia, Pre-Textos, 1980, pp. 15,25.

[20] E. Castro, «Realismo postcontinental, Ontología y epistemología para el siglo XXI,» Madrid, 2019, p. 69.

[21] E. Tugendhat, Introducción a la filosofía analítica, Barcelona: Gedisa, 2003.

[22] G. Boole, An Investigation of the Laws of Thought, Watchmaker Publishing, 2010.

[23] L. Wittgenstein, Tractatus logico philosophicus, Alianza, 2012.

[24] V. Kraft, El círculo de VIena, Taurus, 1977.

[25] E. Alonso, «Sócrates en Viena, una biografia intelectual de Kurt Gödel,» Barcelona, Montesinos, 2007, p. 141.

[26] M. Merlau-Ponty, Fenomenologia de la percepción, 1945.

[27] G. Hardy, Apología de un matemático, Capitan Swing, 2013.

[28] A. Turing, «COMPUTING MACHINERY AND INTELLIGENCE,» *Mind*, nº 236, pp. 433-460, 1950.

[29] G. Dyson, La catedral de Turing, los orígenes del universo digital, Penguin Random House, 2015.

[30] B. Attila y C. Sik Lányi, «"History of Artificial Intelligence.",» de *Encyclopedia of Information Science and Technology, Second Edition,*, IGI Global, 2009, pp. 1763-1768.

[31] H. Dreyfus, What computers can't do, 1972.

[32] B. G. Buchanan, «“A (Very) Brief History of Artificial Intelligence”,» *AIMag,*, vol. 26, nº 4, p. p. 53, 2005.

[33] J. McCarthy, «Ascribing Mental Qualities to Machines.,» 1979.

[34] N. Al-Sibai, «OpenAI Chief Scientist Says Advanced AI May Already Be Conscious,» Vols. %1 de %2https://futurism.com/the-byte/openai-already-sentient.

[35] R. Penrose, La nueva mente del emperador, Penguin Random House, 1991.

[36] R. Penrose, Sombras de la mente, Editorial Crítica , 2013.

[37] H. Dreyfus, What Computers Still Can’T Do: A Critique of Artificial Reason, MIT Pres, 1992.

[38] M. M. Hedblom, Image Schemas and Concept Invention, Springer, 2020.

[39] J. Hougeland, «Body and world: a review of What Computers Still Can’t Do: A Critique of Arti cial Reason,» *Artificial Intelligence,* nº 80, pp. 119-128, 1996.

[40] H. Collins, «Embedded or embodied? a review of Hubert Dreyfus’ What Computers Still Can’t Do,» *Artificial Intelligence,* vol. 80, pp. 96-117, 1996.

[41] T. P. a. E. Meyers, «“Turing++ Questions: A Test for the Science of (Human) Intelligence”,» *AIMag,* vol. 37, nº 1, pp. pp. 73-77, 2016.

[42] E. Davis, «“How to Write Science Questions that Are Easy for People and Hard for Computers”,» *AIMag,* vol. 37, nº 1, pp. pp. 13-22, 2016.

[43] E. D. a. C. L. O. Morgenstern L., «“Planning, Executing, and Evaluating the Winograd Schema Challenge”,» *AIMag,* vol. 37, nº 1, pp. pp. 50-54, 2016.

[44] P. C. a. O. Etzioni, «“My Computer Is an Honor Student — but How Intelligent Is It? Standardized Tests as a Measure of AI”,» *AIMag, vol. 37,*, vol. 37, nº 1, pp. p. 5-12, 2016.

[45] P. P. a. G. Marcus, «“Toward a Comprehension Challenge, Using Crowdsourcing as a Tool”,» *AIMag,* vol. 37, nº 1, pp. pp. 23-30, 2016.

[46] H. Kitano, «“Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery”,» *AIMag,* vol. 37, nº 1, pp. pp. 39-49, 2016.

[47] W. J. a. P. Z. Yeh, «The Social-Emotional Turing Challenge,» *AIMag,* vol. 37, nº 1, pp. 31-38, 2016.

[48] ,. L. A. A. S. A. M. M. D. B. a. D. P. Zitnick, «"Measuring Machine Intelligence Through Visual Question Answering",» *AIMag,* vol. 37, nº 1, pp. pp. 63-72, 2016.

[49] C. L. Ortiz, «"Why We Need a Physically Embodied Turing Test and What It Might Look Like",» *AIMag,* vol. 37, nº 1, pp. pp. 55-62, 2016.

[50] G. B. a. M. C. S. S. Adams, «"I-athlon: Towards A Multidimensional Turing Test",» *AIMag,* vol. 37, nº 1, pp. pp. 78-84, 2016.

[51] K. D. Forbus, «"Software Social Organisms: Implications for Measuring AI Progress",» *AIMag,* vol. 37, nº 1, pp. pp. 85-90, 2016.

[52] D. Griffin, Animal Minds: Beyond Cognition to Consciousness, University of Chicago Press, 2001.

[53] K. Friston, «The free-energy principle: a unified brain theory?,» *Nat Rev Neuroscience,* vol. 11, pp. 127-138, 2010.

[54] K. Friston, J. Kilner y L. Harrison, «A free energy principle for the brain,,» *Journal of Physiology-Paris,* vol. 100, nº 1-3, pp. 70-87, 2006.

[55] J. Ortega y Gasset, Ensimismamiento y alteración, meditación de la técnica y otros ensayos, Alianza Editorial, 2014.

[56] F. Varela, H. Maturana y R. Uribe, «Autopoiesis: The organization of living systems, its characterization and a model,» *Biosystems,* vol. 5, nº 4, pp. 187-196, 1974.

[57] «Eugene the Turing test-beating 'human computer' – in 'his' own words,» *The guardian,* 9 Junio 2014.

[58] G. Shunryu, «The "General Problem Solver" Doesn't Exist: Mortimer Taube and the Art of AI Criticism.,» 2018.

[59] S. M. Shieber, « "Principles for Designing an AI Competition, or Why the Turing Test Fails as an Inducement Prize",» *AIMag,* vol. 37, nº 1, pp. pp. 91-96, 2016.

[60] D. B. Lenat, «"WWTS (What Would Turing Say?)",» *AIMag,* vol. 37, nº 1, pp. pp. 97-101, 2016.

[61] GPT-3, «A robot wrote this entire article. Are you scared yet, human?,» *The Guardian ,* 9 Septiempbre 2020.