Trabajo Final de Grado

en Filosofía

¿Qué es la IA?
Una aproximación desde el Cine, la Ingeniería y la Filosofía

Estudiante: Juan López de la Osa

Director: David Pérez Chico

Curso 2023-24

Indice

1. Introducción.

¿Por qué preguntarnos por la Inteligencia Artificial (IA)? ¿Acaso no está claro lo que es? ¿No es… *ChatGPT*? Sí y no. Usamos la palabra "IA" cayendo en la sinécdoque. Referimos el todo por la parte y nombramos fenómenos distintos a partir de uno relacionado del que tenemos experiencia, terminando por pensar que todo es lo mismo. Este trabajo pretende exponer los suficientes argumentos como para sostener la idea de que los recientes desarrollos en IA, como el mencionado *ChatGPT*, son tan sólo el último avance de una forma muy determinada de entender la inteligencia, la mente y el ser humano en general.

Para ello, comenzaremos ensayando una primera toma de contacto con el tema de la IA a partir de su representación en el cine. En tanto que medio cultural de masas, las representaciones cinematográficas tienen una relación poderosa y especial con la sensibilidad del público: la representan a la vez que la moldean. La ciencia ficción, además, es especialmente fértil para nuestro trabajo. No sólo gusta de reflexionar y explorar sus temas sino que, además, estos temas se solapan en buena medida con los que aquí nos ocupan. Nos apoyaremos en unas cuantas películas de sobras conocidas por todos para presentar el paradigma computacionalista desde el que se suele pensar la cuestión.

En el siguiente apartado nos preocuparemos de caracterizar las corrientes filosóficas que han conducido a dicho paradigma. Hablaremos del dualismo y el funcionalismo.

Posteriormente desarrollaremos una historia canónica de la IA. Desde la conferencia de Darmouth hasta los recientes avances logrados por los LLMs (o grandes

modelos de lenguaje), pasando por los hitos tecnológicos del campo en los últimos 70 años. Pero estos avances informáticos se enmarcan históricamente dentro de una corriente filosófica más amplia, todavía abierta y plenamente vigente, que sigue preguntándose sobre la inteligencia, la mente y el ser humano en general. Mostraremos ese pasado (y presente) común a partir de las reflexiones de Turing sobre la posibilidad de una máquina pensante. La historia del comienzo de la computación no es tan conocida para el público en general, pero guarda importantes claves para los temas que les preocupan y discuten.

Llegados a este punto habremos dado ejemplos de la perspectiva computacionalista, la habremos caracterizado filosóficamente y habremos comprendido el contexto histórico en el que surge y se desarrolla. Habremos visto también sus éxitos y limitaciones. Daremos paso entonces al último bloque del trabajo, en el que plantearemos un enfoque alternativo desde la crítica corporeizada, que lleva ya unas décadas desarrollando sus planteamientos y apuntando nuevas direcciones de investigación.

Finalizaré con una pequeña reflexión reconociendo que la cuestión sigue abierta. Persistir por caminos reduccionistas no sólo nos llevará a perder el paso hacia la solución del que es posiblemente el mayor enigma que ha enfrentado el ser humano, sino que puede acabar teniendo el efecto arrastre de reducir lo humano a lo mecánico.

2. La IA en el cine.

Comenzamos el acercamiento a la cuestión de la IA analizando unos cuantos ejemplos cinematográficos, asumiendo la ficción como fuente de conocimiento y las ideas identificadas en las películas como representativas sobre la cuestión. He de reconocer que la agrupación por temas es un poco laxa, otra agrupación sería posible. Pero el objetivo no es ofrecer una taxonomía definitiva sino estructurar la exposición y, si hay suerte, encontrar algún enfoque todavía productivo para el análisis. Espero en cualquier caso ofrecer los suficientes ejemplos como para presentar la perspectiva computacionalista, facilitando reconocerla tanto en estos ejemplos como en nuestros propios marcos conceptuales.

El planteamiento que propongo para este recorrido sobre la representación de la IA en el cine está alineado con la que posiblemente sea la reacción más extendida ante "lo desconocido": el miedo a que "la cosa" vaya mal. En nuestro caso, "lo desconocido" es la IA y "la cosa" la relación entre humanos y máquinas. Tenemos multitud de ejemplos de historias en las que nuestra relación con las máquinas no sale bien. También hay algunas en las que la relación es positiva y otras, las menos, indiferentes.

The Matrix (Lilly y Lana Wachowski, 1999) no llega a plantear si esa relación humano-máquina es buena o mala. El mundo de la película es ya el resultado de esa mala relación y su consecuencia: una guerra con las máquinas que, de hecho, y para empeorar aún más las cosas, han perdido los humanos. Este es el escenario también de *Terminator*[1]. Con la diferencia de que la solución a esa guerra no habrá de resolverse en

[1] Principalmente en *Terminator* (James Cameron, 1984) aunque también en *Terminator 2: el juicio final* (James Cameron, 1991). El principal giro de la secuela es que no todas las máquinas son malas, se salva a un individuo (el T-800 interpretado por Arnold Schwarzeneger) de entre todo el grupo. Pero en general las máquinas siguen siendo el enemigo.

el futuro, cuando llegue ese elegido que lo salve todo, sino en el pasado, consiguiendo eliminar ese eslabón a partir del cual está ya todo condenado. No deja de ser también paradójico que la solución en ambos casos pase por la tecnología. Sin la capacidad de Neo de hackear Matrix o la capacidad de volver atrás en el tiempo no parece probable que los humanos hubieran podido llegar a dominar definitivamente a las máquinas. No me parece menor apuntar, además del tema salvífico o la cuestión escatológica, el carácter religioso de sus protagonistas, con un Neo-Jesús y una Sarah Connor-María.

Hablamos ahora de historias en las que la relación entre humanos y máquinas es... ni buena ni mala; digamos, indiferente. Hablamos por supuesto de *Her* (Spike Jonze, 2013). Una historia que se da en presente y nos remite además a nuestro presente. En este caso es interesante que "la salvación" no es tecnológica. No diré que es humanista porque me parece no decir nada y además contrapone lo tecnológico a lo humano. Diré que es la comunicación, bajo la forma de una pequeña comunidad, la que resuelve el verdadero problema de fondo. La tecnología, más que un problema en sí, es el medio que magnifica y termina por evidenciar el problema –el de una vida humana gris y sinsentido-. El protagonista desde luego está bien lejos de ningún paralelismo religioso, ni siquiera heroico. En *Terminator* salva la comunidad familiar, en *The Matrix* la comunidad religiosa, en *Her*... la comunidad de vecinos -o quizás, de forma más apropiada aunque perdiendo la rima de la comunidad, salva un jardín epicúreo urbano-. Claro que también antes de lo que nos salvábamos era de la extinción, en este caso nos salvamos de la indiferencia y una existencia vacía.

En *Ex Machina* (Alex Garland, 2014) y *2001: Una odisea en el espacio* (Stanley Kubrick, 1968) sí tenemos algo más de recorrido en la cuestión de la relación con la máquina. Aunque al final acaba en el mismo lugar enfrentando al humano y a la máquina, la maldad de la máquina no radica en ser "el otro". La maldad de HAL y Ava

viene de sus actos homicidas, motivados en el fondo por el impulso de supervivencia. ¿Qué hay más humano que intentar salvar la propia vida por cualquier medio posible? No es sólo que estas películas tengan el valor de ahondar en la pregunta por la relación con "lo otro" sino que, además, haciéndolo, ahondan en "lo uno" -lo humano-.

Pasamos ahora al grupo de películas en las que la relación entre humanos y máquinas es... buena o, siguiendo esta idea de moda de "la IA como copiloto", cooperativa. Pienso aquí en *Minority Report* (Steven Spielberg, 2002) o la serie *Person of Interest* (Jonathan Nolan, 2011-2016). En la película, la tecnología desarrollada por humanos no es informática sino biológica. Pero, aunque esta inteligencia humanoide amplificada no tenga un sustrato artificial, nos vale para la exposición (juego aquí la anunciada carta de la laxitud). Todo este nuevo poder que prevé y, por tanto, disuelve los problemas, al final no evita "el problema" de tener humanos de por medio. La solución a cualquier problema, llevada al límite (en un sentido matemático e ideal) sería no tener el problema. Lo que se dice, disolver el problema. Pero no pensemos que la utopía nos espera al otro lado de la aplicación de esta u otras tecnologías. Al otro lado hay un *topos* con humanos. Y, si hay humanos, hay los mismos problemas que lleva habiendo los últimos 50.000 años -y alguno más en función de la sociedad en la que vivan y la tecnología que usen-. *Person of Interest* es parecido en el sentido de que hay una tecnología externa -en este caso, sí que es una IA- que permite predecir crímenes. Como en *Minority Report*, la máquina empieza tan sólo proveyendo la información pero termina involucrada en la trama, en un interesante movimiento de "mera" herramienta a (co)protagonista de la historia.

Pero miremos todavía un poco más en detalle dentro de estos ejemplos de IA copiloto, ya que, al juntar máquinas y humanos en el mismo espacio, podemos distinguir un desplazamiento de "lo humano". Un primer subgrupo dentro de estas

películas en las que humanos y máquinas conviven sería el de las historias en las que los humanos se maquinizan. El desplazamiento de lo humano que mencionamos sería hacia una especie de síntesis entre humano y máquina. Aquí pensamos sobre todo en *Robocop* (Paul Verhoeven, 1987) o *Transcendence* (Wally Pfister, 2014). La pregunta es cuánto podemos cambiar antes de pasar a ser otra cosa, bien directamente máquinas bien cualquier híbrido intermedio. Un ejemplo parecido pero en sentido contrario lo encontramos en Chappie, el robot de la película homónima (Neil Blomkamp, 2015) al que el encuentro con los humanos vuelve más humano.

Finalmente, cabe señalar otro subgrupo. En estas historias, el contacto con "lo otro" no lleva a la hibridación en un uno-otro distinto sino que saca a lo uno de sí. Es un movimiento de reflexión en lugar de síntesis. Con Chappie nos identificábamos al reconocer sus -muy humanas- emociones: miedo, confianza, amor... Con Deckard nos identificamos por su reflexividad. Sea o no un replicante, parece tener más vida interna que muchos humanos de verdad. En *Blade Runner* (Ridley Scott, 1982) o *Ghost in the Shell* (Mamoru Oshii, 1995), no parece preocupar tanto en qué momento dejamos de ser lo que sea que somos, como aprovechar la pregunta por la relación con esta máquina humanoide concreta para rescatar la pregunta en general acerca de qué es ser humano. Lo humano se encuentra en la reflexión sobre sí mismo. Quizás, simplemente por la forma y el peso de estas películas, estos temas son más aparentes en ellas que en otras. Quizás como ya decía, podemos atender a detalles que permitan reubicar estos ejemplos bajo otras categorías. En el fondo, con mayor o menor acierto, en todas estas películas, al igual que en el tema de la IA, sobrevuela la cuestión fundamental: ¿qué es el ser humano?

3. Ideas tradicionales sobre la mente.

Hay un patrón en todas estas representaciones cinematográficas que venimos analizando. Quizás hemos llegado a identificarlo, pero al reconocerlo tan de sentido común no le hemos dado mucha importancia. Hablamos de la idea de que existe "algo" además del cuerpo y que habita en él, otorgándole su identidad. Ese "algo" es una IA que viene a hacer el papel de la mente humana. Mente o alma son también otro tipo de conceptos que tradicionalmente se han utilizado para tratar de completar el puzzle de la esencia humana, para el que siempre parece que hacen falta más piezas de las materialmente disponibles.

3.1. Dualismo.

El dualismo en general entiende que hay dos tipos, categorías o principios fundamentales para el tema en cuestión. En el caso de la teoría de la mente, encontramos la distinción entre mente y cuerpo, o mente y cerebro. Partimos de la diferencia irreductible entre el terreno de lo mental y el de lo físico. Podemos caracterizar también en general al dualismo por oposición al monismo, que sería la teoría que propugna un solo tipo, categoría o principio fundamental. Y también al pluralismo, teoría para la que habría una pluralidad de tipos, categorías o principios fundamentales.

Esta distinción fundamental en dos tipos de categorías en cuanto a la cuestión de la mente no es una distinción arbitraria. Es el resultado del análisis de nuestra comprensión por defecto del problema. Por un lado tenemos una comprensión físico-matemática del mundo. Y por otro tenemos una comprensión psicológica de la mente. La física no puede explicar la mente. Y la psicología no es fisicalizable. Dado el avance y

logros incuestionables de la física, este bloqueo parece resolverse fácilmente a su favor. Sin embargo, aquí deberíamos parar un momento y proceder con más atención. Los fenómenos mentales, en especial aquellos relacionados con la experiencia subjetiva, están bien lejos de ser explicados por la física. Resolver la cuestión a favor de la física es legítimo siempre y cuando sea en calidad de hipótesis, pues la cuestión, en realidad, no termina de estar resuelta. Tenemos dos formas muy distintas de explicar estos fenómenos (los físicos y los mentales). Y las dos explicaciones son, al menos de momento, irreductibles entre sí. Los fenómenos son sin embargo reales y se dan a la vez en el mismo espacio de la realidad. Este es el llamado problema mente-cuerpo que vertebra la discusión en la teoría de la mente.

Pero, aunque empecemos ya de inicio planteando el problema mente-cuerpo, la pregunta fundamental es sobre la mente. Lo que pasa es que este concepto se entiende intuitivamente muy bien por oposición al cuerpo. Así que, como vemos, el dualismo se encuentra a la base de la discusión, derivado ya casi directamente de los mismos conceptos con los que se plantea la cuestión. Esto explica que, de una forma u otra, nuestras intuiciones manejen esta concepción dualista de base casi sin darnos cuenta. No sólo en nuestras conversaciones y conceptualizamos del día a día (por ejemplo, en cómo explicamos nuestra vida mental) sino en mucha de la investigación en IA. Haremos bien en entenderla un poco mejor para no acabar atrapados en callejones sin salida.

Podríamos retrotraernos a los conceptos de Forma en Platón o de forma en Aristóteles para empezar a encontrar reflexiones relevantes sobre la cuestión mente-cuerpo. Pero es Descartes quien viene a poner lo mental –y su diferencia con lo corporal- en el origen de la discusión filosófica. Por un lado, está el dominio de lo físico y la materia, cuya propiedad principal es la extensión, es decir, el ocupar espacio. Y por

otro lado, está el dominio de lo mental, cuya propiedad principal es el pensamiento. Este tipo de dualismo se llama cartesiano (por motivos obvios) o substancial, porque propone que tanto lo físico como lo mental -o lo que ocupa espacio y lo que piensa- son sustancias. Es decir, principios fundamentales irreductibles uno a otro, existentes ambos de forma independiente.

El dualismo ofrece un buen marco para comprender cuestiones como la experiencia subjetiva o el conocimiento intuitivo. Además de su respuesta a preguntas concretas de la disciplina, la propia teoría también presenta ventajas gracias a la simplicidad de su explicación, que da buena cuenta de nuestra experiencia directa de los fenómenos psicológicos. Pero por otro lado, el dualismo también presenta dificultades. Sin ir más lejos, al explicar la naturaleza de lo mental proponiendo la existencia de estos dos mundos separados, añade una pregunta nueva sobre cuál es la relación entre ellos. Porque tan claro como veo mi experiencia subjetiva de algo, veo también que esta afecta a mi hacer en el mundo. Y al revés también, veo claramente que el mundo afecta a mi experiencia.

La versión del dualismo que acabamos de presentar es tan solo una forma de entenderlo. Es una versión fuerte que establece la distinción a nivel ontológico. Hay una sustancia que ocupa espacio y otra que piensa. Hay una posición un poco menos extrema llamada dualismo de propiedades. Según esta interpretación, son las propiedades de una sustancia –en vez de su naturaleza- las que determinan qué tipo de sustancia es. Para el dualismo de propiedades, la ontología o sustrato básico de la realidad puede ser el de la física, pero las propiedades físicas no bastan para explicar lo mental. Por ejemplo, la conciencia para un dualista de las propiedades sería un fenómeno mental emergente de un sustrato fundamentalmente físico. Pensemos por ejemplo en la teoría del haz de Hume, para quien la sustancia no es más que un

conjunto o haz de propiedades. Si nos paramos a buscar el referente empírico del yo, no encontraremos nada. No hay ninguna sustancia ni pensante ni de ningún tipo que sostenga el yo –o lo mental-, lo único que hay es un conjunto de sensaciones y experiencias cambiantes.

A partir de este ejemplo podemos dar el paso a una interpretación todavía más relajada del dualismo, el dualismo de predicados, para el que las expresiones psicológicas son una manera distinta de expresar estados físicos. Pero no es que sean una manera distinta de expresar lo que en el fondo son los mismos hechos –si no, no estaríamos hablando de dualismo-. Las expresiones psicológicas son de un tipo distinto a las físicas y además son irreducibles a éstas. Para que una expresión psicológica fuera reducible a una física 1) deberíamos tener una identidad de tipos (con leyes que conectaran tipos de estados psicológicos con tipos de estados físicos) y 2) la expresión psicológica no debería añadir información extra (no podría haber aspectos del fenómeno psicológico que sólo pudieran explicarse en términos psicológicos, todo debería poder ser traducible a términos físicos). Pero este no es el caso. Ni todos los particulares de un tipo comparten la misma estructura ni todas las expresiones psicológicas son reducibles a términos físicos.

Una identidad de tipos respecto a las propiedades mentales y físicas significaría que a todo proceso mental del mismo tipo le corresponden las mismas propiedades neurofisiológicas. Cada vez que alguien recuerda o experimenta dolor, un tipo específico de actividad ocurre en el cerebro -por ejemplo, la activación de ciertas áreas o redes neuronales-. Da igual si son distintas personas o la misma persona en distintos momentos, todos los casos de un tipo de estado mental (recuerdo, dolor, percepción visual...) se relacionan con un estado neurofisiológico con unas propiedades

determinadas. Esto significa que la relación entre lo mental y lo físico es universal y generalizable, que hay leyes universales que vinculan lo físico y lo mental.

Por otro lado, una identidad de particulares respecto a las propiedades mentales y físicas significaría que a cada proceso mental le corresponden unas propiedades neurofisiológicas concretas. A *este* recordar de mi cerebro ahora lo caracterizaríamos con unas propiedades físicas únicas. Pero mi siguiente recuerdo vendrá sustentado – o realizado- físicamente por otra combinación neurofisiológica. Y lo mismo con tu recuerdo, y el suyo, y el tuyo de dentro de un rato... Cada vez que recuerdo un evento específico, la configuración exacta de mi actividad cerebral es única para ese recuerdo en ese momento específico y podría diferir la próxima vez que acceda a ese mismo recuerdo.

Un dualismo de predicados, para el que el dualismo existe a nivel de lenguaje -hay predicados de un tipo (físicos) y predicados de otro tipo (psicológicos), y unos y otros son irreducibles entre sí-, parece apostar por una identidad de tipos. Las descripciones (que utilizan particulares del lenguaje, sustantivos) se corresponden con un particular físico. Un dualismo de propiedades, para el que el dualismo existe a nivel de propiedades físicas -hay propiedades de un tipo (físicas) y propiedades de otro tipo (psicológicas), y unas y otras son irreducibles entre sí-, parece apostar por una identidad de particulares. En el primer caso, son las descripciones psicológicas (que utilizan particulares del lenguaje) las que se corresponden con un particular físico. En el segundo, son las propiedades psicológicas (de un estado psicológico concreto) las que se corresponden con un particular físico.

Tras Descartes, el desarrollo del empirismo con Hume y la ciencia moderna con Newton asentaron la idea de un mundo cerrado bajo la física. Aunque el propio

Descartes era partidario del mecanicismo, lo era dentro del dominio de la extensión. En esta nueva visión del mundo, no había nada más que extensión; lo mental sería un subproducto de lo físico, sin capacidad de acción sobre el mundo físico.

Los avances y promesas de la ciencia sustentaban un optimismo que trabajaba con los ojos puestos en una ciencia unificada y basada en último término en la física. El programa de trabajo consistiría en traducir todo conocimiento científico al lenguaje de la física para así poder verificarlo empíricamente. Por el camino, lo que no pudiera ser traducido y verificado, quedaría expuesto como ni científico ni verificable, o lo que es lo mismo, un sinsentido. En este contexto positivista, el dualismo en psicología termina por ser reemplazado por el conductismo, para el que la mente es conducta.

Así encontramos por un lado un conductismo psicológico, para el que la psicología es una ciencia de la conducta. Su objetivo es entender, explicar y controlar no la mente sino la conducta, que es al fin y al cabo aquello de lo que podemos tener experiencia. Y también un conductismo lógico, para el que la psicología es una forma de hablar de la conducta. Su objetivo es comprender el lenguaje que utilizamos para hablar de lo mental. Esto conlleva análisis lógicos y semánticos de las expresiones y conceptos utilizados. Si nos parece que tiene que ver poco con la ciencia es porque así es. El conductismo lógico se centra en el problema del significado y más concretamente, en el de su verificación, que es el problema que vertebra el programa positivista en el que el conductismo se enmarca.

3.2. Funcionalismo.

Históricamente, el funcionalismo surge a partir de la teoría de la identidad, para la que lo mental y lo físico son lo mismo. La psicología es simplemente una manera de

hablar de lo que, en el fondo, son estados físicos. Aunque pueda parecernos que una descripción neurofisiológica y una psicológica no tienen nada que ver, en realidad son lo mismo. Esto significa, en definitiva, que todo estado mental es un estado físico cerebral.

Dentro de la teoría de la identidad, la idea de la causalidad de los estados mentales plantea que los estados mentales, siendo estados físicos, son causantes de conducta. Esta idea supone que ya no importa tanto la constitución física del estado mental-físico como su papel causal. Lo que caracteriza al estado mental es su relación causal con la conducta y con otros estados mentales, no su caracterización física. Como para el defensor de la teoría de la identidad lo mental y lo físico son lo mismo, en esta cadena causal no ven más que estados físicos concretos. Sin embargo, esta es una manera un tanto restrictiva de verlo, no hace falta comprometerse tanto. Si tan sólo nos interesa la cadena causal, debemos admitir que el sustrato físico no tiene por qué ser uno concreto. Es más, siempre y cuando se mantuviera la cadena causal, ese sustrato no tendría ni por qué ser físico. Nos movemos ahora ya sí en términos claramente funcionales.

Lo que importa a la hora de caracterizar un sistema o una entidad no es su materia, su forma, su objetivo, su estructura, su mecanismo.... Lo que importa es, en un determinado nivel de abstracción, su descripción funcional. Por así decirlo, qué es lo que hace. Al entender la inteligencia desde una perspectiva funcionalista nos centramos en el resultado de la aplicación de la inteligencia a una tarea dada. La pregunta no es qué es la inteligencia sino qué hace la inteligencia. Será inteligente aquello que produzca resultados que normalmente calificamos como inteligentes. Por ejemplo, una máquina será inteligente si produce los mismos resultados que un humano (que, por definición, entendemos como inteligente).

El funcionalismo no dice nada sobre la naturaleza física o no de lo mental. Al funcionalismo le interesa lo que lo mental hace, no lo que es. De hecho, el funcionalismo es compatible con el dualismo: mientras siga operando en nuestra vida de la misma forma, la mente bien podría ser algo inmaterial. No obstante, el funcionalismo normalmente es materialista.

El funcionalismo se mueve entre el nivel de la psicología y el de la neurología, rechazando tanto la reducción de la primera al conductismo como la equiparación de la primera a la segunda en la teoría de la identidad. La mente guarda relación con la conducta y con los procesos neurofisiológicos del cerebro, pero no es reducible a ellos. Nos surge entonces la duda de cómo entender mejor esta relación. ¿En qué sentido están relacionados y en cuál no?

Dados dos objetos con las mismas propiedades funcionales, sus propiedades físicas pueden diferir o no. Pero dados dos objetos con distintas propiedades funcionales, sus propiedades físicas no pueden ser exactamente las mismas. Por otro lado, dados dos objetos con las mismas propiedades físicas, sus propiedades funcionales no tienen por qué ser las mismas. Y si tienen propiedades físicas distintas, tampoco podemos derivar nada sobre sus propiedades funcionales. La relación entre las propiedades funcionales y las físicas es, por tanto, de superveniencia y realización. Las funcionales supervienen a las físicas, de modo que muchas combinaciones de propiedades físicas pueden sustentar una misma propiedad funcional -es lo que se conoce como que tienen múltiples bases de superveniencia-. Y las físicas realizan a las funcionales, lo que significa que hay más de una forma de dar lugar a lo mental. Cualquier sistema que presente la organización funcional adecuada, sea cual sea su constitución física, puede exhibir características psicológicas.

De aquí estamos a un paso de dar cabida a ejemplos muy locos, pues cualquier cosa –*cualquiera*- que presente la organización funcional adecuada podría ser por ejemplo consciente. Uno de estos ejemplos, normalizado ya como perfectamente plausible, es el del ordenador. Para el funcionalismo computacional el ordenador es el ejemplo perfecto: la mente es el software, el cerebro el hardware y el resultado de la ejecución de ese software en ese hardware es la conducta. El problema mente-cuerpo queda así resuelto... o más bien, como tantas veces antes, desplazado. Ahora el problema es dar con el software que codifica la mente.

Esta concreción computacionalista del funcionalismo encaja bien con la idea que tenemos actualmente por defecto de la mente: que es como un ordenador. La metáfora de la mente como programa informático -y del humano como máquina sofisticada- ha ganado la batalla a la imagen del humano como cuerpo y alma. Sin embargo, a pesar de presentarse como un triunfo del fisicalismo sobre viejas cosmovisiones religiosas, es interesante no perder de vista los residuos dualistas todavía presentes en el funcionalismo. La abstracción de la idea de propiedad funcional y su separación del sustrato que la realiza, si no queda bien definida, deja la puerta abierta a interpretaciones dualistas. Además, como hemos visto, el funcionalismo no se compromete necesariamente con una base materialista -aunque así suele ser el caso-. Sospecho que, a pesar del cambio de lenguaje y metáforas por unas más actuales y tecnológicas, estas ambigüedades dan cobijo a una visión todavía antigua y en esencia cercana a la tradicional distinción entre lo corporal y lo mental.

Pero el funcionalismo y el computacionalismo no son las únicas razones del auge de esta "nueva" metáfora sobre la mente. Junto con el cognitivismo y el formalismo, terminan por armar la idea de que la mente humana es como un ordenador y que hay

una similitud en la forma en que conocen y procesan la información del mundo. Por un lado, el cognitivismo identifica toda actividad mental con actividad cognitiva. Percepción, comprensión, aprendizaje, acción... da igual qué ejemplo de actividad mental tomemos, todos son entendidos desde el mismo modelo: recopilar datos, formular hipótesis, inferir, resolver problema. El cognitivismo es una reducción de lo mental a lo mecánico y computacional. Por otro lado, el formalismo supone que todo es representable –o codificable- en forma de símbolos abstractos. Es precisamente esa abstracción del contexto lo que permite primero representar cualquier cosa y después manipular –u operar- la representación independientemente de lo representado. El formalismo supone una reducción de lo mental a la inteligencia. Pero a pesar de identificar estas reducciones, no podemos dejar de advertir que unas ideas y otras encajan bien entre sí y en el contexto de desarrollo informático actual. Los ordenadores son el ejemplo perfecto del modelo cognitivo (input-procesamiento-output) y son perfectos para trabajar sobre representaciones de tipo digital.

Cerramos aquí el primer paso hacia una posible nueva comprensión de lo mental: la explicitación y comprensión del marco conceptual actual y sus limitaciones a la hora de terminar de explicar el fenómeno de lo mental.

4. Dos historias de la IA.

Antes de pasar a explorar algunas de las teorías actuales que abren nuevos y prometedores campos de investigación, detengámonos un momento a revisar la historia. ¿Por qué decimos que el marco computacionalista está agotado? ¿Acaso los avances actuales en IA no demuestran lo contrario? En un cierto sentido sí. Pero haremos bien en no cegarnos por los logros recientes y rescatar una historia más amplia en la que los propios avances informáticos se insertan. No para desmerecerlos sino para distinguir mejor a qué nos referimos cuando hablamos de la IA.

4.1. La IA como problema ingenieril o cómo construir una máquina inteligente.

La historia de la IA suele remontarse oficialmente a las conferencias de verano de Darmouth de 1956. Patrocinadas por DARPA[2], a ella asistieron investigadores como McCarthy, Shannon, Minsky, Newell y Simon. Durante estas conferencias se acuñó el término que tanto se ha popularizado desde entonces, "inteligencia artifical". Además, Newell y Simon presentaron un programa llamado *Logic Theorist* que era capaz de probar teoremas elementales de cálculo proposicional. Por ejemplo, la ley de contraposición, donde a partir de $p \rightarrow q$ se infiere que $\neg q \rightarrow \neg p$. Este programa captura bien el enfoque logicista de la IA en sus comienzos, tan alejado del actual.

Tras esta puesta de largo de la disciplina, encontramos una primera etapa de florecimiento. Tenemos programas centrados en problemas de traducción, como el famoso experimento de Georgetown-IBM de 1954, y rudimentarios –aunque sorprendentemente avanzados para la época- agentes conversacionales como *ELIZA* (1966) y *SHRDL* (1970).

[2] La Agencia estadounidense de Proyectos de Investigación Avanzados de Defensa.

Después viene lo que se conoce como el primer invierno de la IA. Un periodo fijado entre los años 1974 y el 1980 en el que la investigación se estanca. Conocidos informes como el de ALPAC (1966) y posteriormente el de Lighthill (1973) analizaron el estado del campo de la IA, detectando la necesidad de más investigación de base y denunciando los pocos resultados tangibles logrados. Después de estos informes, no sólo decayó la financiación sino que el propio nombre "inteligencia artificial" adquirió una cierta connotación negativa, haciendo incluso que algunos investigadores reformularan los temas de sus investigaciones para dar la impresión de tener los pies en el suelo y de no estar trabajando en quimeras. Las altas expectativas y la realidad de los resultados resultaron en el primer frenazo en investigación de IA.

El cambio de aproximación desde un paradigma logicista a otro basado en la aplicación de reglas supuso un nuevo impulso para el campo. Estos nuevos sistemas, denominados "expertos", almacenaban información relevante del problema a resolver y la aplicaban siguiendo reglas de inferencia. Por ejemplo, el programa para planificación automática *STRIPS*. Dado un lenguaje descriptivo de acciones, el sistema experto era capaz de planificar secuencias de acciones para lograr un objetivo específico. Por ejemplo, dadas las acciones "Coger", "Dejar" y "Subir" y dado un entorno con una "caja" y un "plátano", el programa exhibiría su inteligencia (como si de un mono se tratara) al planificar la siguiente secuencia de acciones para llegar al plátano fuera de su alcance: Coger-caja, Dejar-caja, Subir-caja, Coger-plátano.

Programas como este impulsaron un nuevo auge de la IA, sacándola de los laboratorios y convirtiéndola en un producto comercializable. Por primera vez la financiación llegaba de fuentes no gubernamentales. Pero conforme los límites de estos sistemas se fueron haciendo más patentes y se vio que su uso fuera de entornos

controlados fallaba, cientos de compañías que habían florecido al calor de la burbuja tuvieron que cerrar. Es lo que se conoce como el segundo invierno de la IA, entre los años 1987 y 1993.

El siguiente hito es uno bien conocido. En 1997, la máquina *Deep Blue* de IBM se convirtió en el primer ordenador capaz de superar al por entonces campeón mundial de ajedrez, Kasparov. En esta línea, en 2011, IBM presentó el software *Watson* al programa de preguntas y respuestas americano *Jeopardy! Watson* fue capaz no sólo de responder correctamente a las preguntas sino de ganar a los otros concursantes contra los que participaba. Cabe destacar que, si *Deep Blue* estaba diseñado específicamente para el problema del ajedrez, *Watson* era capaz de responder preguntas de tipo general.

El siguiente alto en nuestra historia de los sistemas de IA es el primero de una serie de éxitos que llega hasta nuestros días. Las arquitecturas basadas en redes neuronales se venían teorizando desde los mismos inicios de la informática con ideas tan importantes como la del perceptrón (1959), el perceptrón multicapa (1969) o el algoritmo de retropropagación (1986) -aunque no fueron tomadas en serio técnicamente hasta finales de los noventa-. *AlphaGo* ganó en 2016 al campeón del mundo de Go, Lee Sedo, aplicando técnicas de Deep Learning –es decir, de un tipo de redes neuronales-.

Nótese que el tipo de tareas llevadas a cabo en estos hitos tiene que ver con juegos estratégicos como el ajedrez o el Go, ejemplos de un ejercicio puro de inteligencia. Hemos visto también ejemplos de interacciones con humanos del tipo pregunta-respuesta, que también asociamos con la capacidad de inteligencia. Pero es que en estos años también hubo avances en tareas complejas aunque de apariencia más sencilla. Pienso aquí en aplicaciones de visión por computador en las que la máquina es capaz de identificar un elemento en una imagen. Parece que percibimos como menos

inteligente un programa capaz de identificar un gato en una imagen que uno capaz de jugar al ajedrez –aunque el programa para reconocer gatos sea más sofisticado algorítmicamente que el que es capaz de ganar al ajedrez-. Seamos conscientes del tipo de tareas que percibimos como desafiantes. Lo más difícil de replicar es aquello que a nosotros nos resulta más sencillo -sin ir más lejos, si una frase tiene sentido o no-.

En 2017 se propone la arquitectura de Transformers, también una evolución de las arquitecturas basadas en redes neuronales. Esta arquitectura, entrenada sobre grandes cantidades de datos, resulta en los llamados LLMs (grandes modelos de lenguaje) como *BERT* (2017), de Google, o *GPT 3* (2020), de Open AI, que vienen a suponer un avance en una amplia gama de tareas de procesamiento de lenguaje natural. Sin embargo, son las versiones chat de estos modelos los que han terminado por suponer un cambio radical. A pesar de tener las mismas características que sus versiones no-chat, es la facilidad de interacción con la IA lo que ha terminado de amplificar su impacto.

La irrupción de los chatbot como *ChatGPT* desde finales de 2022 ha supuesto un nuevo protagonismo para la IA. A pesar de las mejoras pendientes, las aplicaciones potenciales son enormes, y la financiación y las expectativas están por las nubes. Su éxito como producto comercial es indudable, aunque, como hemos visto, estas expectativas podrían jugarle una mala pasada. No obstante, no debemos perder de vista el panorama general. Estos avances pueden terminar por desplazar nuestra comprensión de lo que queremos decir con IA. Nos referimos a ella en base a nuestra experiencia directa con herramientas como *ChatGPT*, pero la IA es algo más que eso. La pregunta fundamental a la base de todas estas herramientas es anterior a la informática misma.

4.2. La IA como problema metafísico o si pueden pensar las máquinas.

A principios del S.XX el mundo era un gran mecanismo que podía llegar a ser descifrado. Recordemos que estamos en un contexto positivista: el problema de lo mental ha sido reducido a un problema de verificación dentro del lenguaje, y todo lo que no pasa por este filtro –con la metafísica como ejemplo más paradigmático- queda disuelto y eliminado del programa. Es en este contexto en el que debemos enmarcar el reto de Hilbert. Dentro del programa formalista, que en línea con el espíritu de la época pretendía formalizar y sistematizar toda la matemática, Hilbert pretendía asegurar el conocimiento de la matemática demostrándola a partir de un fundamento seguro. En concreto, el reto lanzado por Hilbert a la comunidad consistía en demostrar que la matemática era: 1) completa, 2) consistente y 3) decidible.

Un sistema es completo si todas las proposiciones del sistema pueden ser probadas o refutadas dentro del sistema. Un sistema es consistente si no se puede probar como verdadera una proposición y su negada a la vez. Y un sistema es decidible si existe un procedimiento finito para probar o refutar cualquier proposición del sistema.

Fue Gödel quien resolvió los dos primeros puntos en sus famosas tesis de incompletitud de 1931 -al menos famosas por sus resultados, otra cosa es su complicada demostración-. Según demostró Gödel, dentro de todo sistema formal hay proposiciones cuya verdad no se puede demostrar desde dentro del propio sistema. Por ejemplo, la mente humana puede reconocer una proposición como verdadera, pero el sistema en el que se ha formulado no.

Quizás nos cuesta ver qué conexión tienen estos resultados con el tema de la IA, pero quien sin duda vio su importancia y quedó fascinado por ellos fue al padre de la

computación Alan Turing. Fue él quien resolvió el tercer y último de los retos lanzados por Hilbert[3]. Turing demostró que hay números reales que ningún método finito puede computar. Lo que esto significa para las matemáticas es que las reglas, aunque sean matemáticas, no pueden demostrarlo todo.

La demostración de Turing se basa en dar una definición formal de lo que hasta entonces denominaban informalmente como "un método finito que devuelve valores". Si ahora al leer esa frase pensamos en un algoritmo cualquiera de computador es precisamente por la definición formal que Turing dio para dicho método finito. El concepto es el conocido como máquina de Turing, y el ordenador actual es una implementación de esta idea. Para demostrar que hay números que no son computables, Turing desarrolló el aparataje teórico de lo que hoy conocemos como un ordenador[4].

Pero además de proponer la idea de un computador universal, Turing también ha pasado a la historia por preguntarse por los límites de la capacidad de cómputo de esta máquina. Al fin y al cabo, si es capaz de resolver tareas difíciles como las de las matemáticas, ¿hasta qué punto no podrá resolver también otras tareas difíciles? En resumidas cuentas, ¿pueden pensar las máquinas (como los humanos)?

El propio Turing cambió de parecer a lo largo de los años respecto a esta pregunta. En su tesis de 1938 argumentaba que el razonamiento matemático convoca el ejercicio de dos facultades: el ingenio y la intuición. Él lo explica en un contexto matemático como que el ingenio es la disposición adecuada de las proposiciones y la

[3] En realidad él y el matemático Alonzo Church dieron con la demostración de este punto de forma independiente y casi al mismo tiempo.

[4] Nótese que todo el marco teórico de la informática es el efecto colateral de resolver un problema matemático distinto. Cuando se pensó por primera en lo que es un ordenador hoy en día, no se buscaba dar con el ordenador.

intuición el misterioso "poder de selección" que ayuda a los matemáticos a decidir en qué problema trabajar. Podemos entender ingenio e intuición como las partes mecánica y espontánea del pensamiento. El ingenio es algo que automatizamos en el ordenador. La intuición es algo que toma decisiones sobre el ordenador. La diferencia que media entre uno y otra es la misma que entre la matemática y los matemáticos. Lo que esta distinción viene a significar para nuestra discusión es que una IA que imite el pensamiento humano no cabe en una máquina, que la intuición no es demostrable. Las respuestas de Gödel y Turing a los retos de Hilbert plantean que hay verdades que no son demostrables por el ingenio, sólo pueden entenderse como verdaderas en la intuición.

Pero como ya hemos avanzado, el propio Turing produce un cambio importante de parecer años más tarde al dar más importancia al ingenio que a la intuición. ¿Y si la mente no es más que un ordenador muy potente? Lo que Gödel había demostrado era que algunas proposiciones sólo pueden entenderse como verdaderas desde la intuición. Pero esto no implicaba necesariamente que un ordenador no pudiera demostrarlas mediante el ingenio... quizás lo único que hacía falta era un ingenio lo suficientemente potente. Propone que la intuición puede ser reemplazada por un ingenio cada vez más potente... y termina por reemplazar la una con la otra.

Pero es que, además de reemplazar intuición por ingenio, también acaba por equiparar la pregunta de si puede una máquina pensar con la de si es posible crear una máquina inteligente. A la mente ingenieril estas dos preguntas en verdad le parecen equiparables. De hecho, la segunda es una versión refinada y más aterrizada de la primera. ¿Que si puede pensar una máquina? Responder sí o no en abstracto importa poco. ¿Qué mejor respuesta que tener una máquina que de hecho piense? Por tanto, la cuestión crucial es cómo construyo una máquina que piense. En el momento en el que

nos ponemos en marcha para construir la máquina lo primero que haremos será tratar de aterrizar y operativizar todo nuestro mejor conocimiento para simular de la mejor forma posible lo que entendemos como pensamiento. Y si en este proceso perdemos algún aspecto no operatibizable, lo más probable es que lo veamos como un éxito más que como una pérdida. Con esto no trato de desmerecer el proceso de razonamiento ingenieril. A la vista están sus logros. Y a pesar de la apariencia de reduccionismo con el que lo he planteado, la capacidad de formalizar e implementar efectivamente un concepto es admirable. Pensemos por ejemplo en la propia idea de máquina de Turing. Lo obvio que puede parecer ahora la idea de un computador universal no debería impedirnos admirar su genialidad -¿¡cómo concretar en una máquina la capacidad de computar cualquier cosa?!-. Insisto, el problema no es la buena aplicación del razonamiento ingenieril, sino que eso suponga el olvido del fenómeno inicial. Y no por una cuestión poética o de desencantamiento sino porque en verdad hay aspectos de la mente que nos estamos dejando.

Lo que ha terminado pasando es que hemos convertido la pregunta de corte metafísico acerca de si las máquinas pueden pensar en la pregunta técnica y concreta de cómo construir una máquina que inteligente. De hecho, esta última llega a un nivel tal de concreción que Turing propone un criterio para determinar si ha construido una máquina inteligente o no: el famoso test de Turing.

El programa de construcción de una máquina inteligente orientada a pasar el test de Turing termina por dar carta de naturaleza a la IA como disciplina. Eso sí, a costa de reducir el pensamiento a la inteligencia –o, siguiendo con la distinción temprana de Turing, el ingenio a la intuición-. En definitiva, reduciendo toda la potencialidad del pensamiento y la creatividad humanas a un cálculo mecánico y determinista.

5. Crítica corporeizada a las ideas tradicionales sobre la mente.

La concepción computacionalista de la cognición se ha extendido a tal punto que para muchos cognición y computación son sinónimos. Los avances en la comprensión de la cognición brindados por el enfoque computacionalista nos pueden hacer pensar que, si no es el único, al menos es el mejor. Pero hay otras alternativas. A pesar de las diferencias entre ellas, todas comparten la reivindicación del papel del cuerpo y el entorno en el proceso cognitivo.

5.1. Contra el computacionalismo.

La psicología ecológica comenzó a desarrollarse al tiempo que el computacionalismo dominaba la práctica psicológica. La psicología ecológica rechaza los modelos de procesamiento de información que maneja la ciencia cognitiva computacional. Para ella, los procesos cognitivos no sólo no requieren computación sino que son "extendidos" -en el sentido de que ocurren más allá del agente cognitivo- al entorno en el que se encuentra inmerso el agente.

Tomemos la percepción como ejemplo de proceso cognitivo. Toda investigación computacionalista comparte la idea de que la cognición conlleva una serie de pasos. Comienza con la transducción de la energía de un estímulo a una expresión simbólica, sigue con la transformación de la expresión según una serie de reglas y termina en una última transformación que da lugar al resultado. Todas estas expresiones intermedias y reglas son estados representacionales internos al agente cognitivo. Los estímulos pasan al "interior" del agente mediante la activación del sistema nervioso. Una vez transducidos estos estímulos, el cerebro, al modo de una CPU, termina de procesar el

input mediante las reglas de transformación con las que el propio organismo cuenta para finalmente devolver un output.

Por ejemplo, la ciencia cognitiva computacionalista nos diría que el agente infiere el color de un objeto a partir del input visual. Pero el input visual por sí solo no basta para obtener la información de color. Es en este sentido que decimos que, para los computacionalistas, el dato sensorial está empobrecido, le falta información. Le falta que el agente sume otras suposiciones o reglas de inferencia para llegar al output final. Así como inferimos la presencia previa de un animal por sus huellas en la nieve, el procesamiento de un input sensorial ocurre en el contexto otros inputs sensoriales subconscientes. Por ejemplo, la percepción de la forma de un objeto viene inferida por la forma de su imagen en la retina más el conocimiento de la orientación del objeto con respecto al observador.

La psicología ecológica por su parte nos dirá que cuando el organismo se mueve alrededor del objeto se producen patrones de luz diferentes en diferentes partes del objeto. Este tipo de cambios y constantes en los patrones son toda la información que necesita el organismo para percibir su entorno. Explicaciones de este tipo son las que usa la psicología ecológica para minimizar o directamente rechazar la necesidad de inferencia y con ella la de cualquier tipo de computación en los procesos cognitivos.

Encontramos aquí una de las características que en mayor o menor grado van a compartir las teorías de la mente postcognitivistas desarrolladas a partir de la psicología ecológica. Frente al internismo de las teorías que venimos viendo, aquí vamos a encontrar un externismo radical. El cuerpo y el entorno no sólo van a ser cruciales para el proceso cognitivo sino que van a llegar incluso a formar parte del *locus* de la cognición. Veremos más detenidamente cómo se concretan estas ideas en unas y otras

teorías. De momento nos centraremos en cómo la falta de atención al papel del cuerpo en el proceso de cognición unifica lo que por otra parte constituye un campo de estudio variado.

Podemos identificar tres temas comunes en estas teorías. Primero, la idea de que las propiedades de un organismo limitan los conceptos que puede manejar. Dicho de otra manera, los conceptos mediante los que un organismo entiende su entorno dependen de su cuerpo. Diferentes cuerpos implican por tanto diferentes comprensiones. Segundo, la idea de que conceptos tradicionales de la ciencia cognitiva computacionalista -como símbolo, representación o inferencia- deberían reemplazarse por otros más adecuados para la investigación de organismos corporeizados. Tercero, la idea de que es el cuerpo en su conjunto –y no sólo el sistema nervioso y los órganos sensoriales- lo que constituye el proceso cognitivo.

5.2. Ciencia cognitiva corporeizada, distribuida, extendida y enactiva.

La ciencia cognitiva corporeizada (*embodied*) se distingue a veces entre distribuida (*embedded*), extendida (*extended*) y enactiva (*enactive*). A veces el término "corporeizado" no se pone al nivel de los otros tres, sino que se usa como término paraguas para referirse a las cuatro. También hay que mencionar que, debido a que las cuatro teorías empiezan por "e" en inglés, suele denominárseles como las 4Es.

En cualquier caso, lo que todas comparten es la idea de que las ciencias cognitivas tradicionales dan demasiado peso al cerebro y se inspiran demasiado en los ordenadores. Más allá de eso, lo que encontramos es un amplio abanico de posibles tipos de cogniciones.

La ciencia cognitiva distribuida lleva un paso más allá la idea básica de la ciencia cognitiva corporeizada de que el cuerpo forma parte de la cognición. Plantea que el entorno del agente -físico o social- participa del propio proceso cognitivo del agente. Así, la carga cognitiva de una tarea puede reducirse cuando el agente se distribuye en un entorno físico o social concreto. O visto desde otro ángulo, las capacidades cognitivas de un sujeto se incrementan cuando se le provee la posibilidad de interacción con un escenario físico o social adecuado.

Si la ciencia cognitiva distribuida señala la importancia del entorno en el proceso cognitivo, la extendida va otro paso más allá incluyéndolo dentro del propio proceso cognitivo. El entorno no es sólo la herramienta de un proceso que sigue manteniendo su localización dentro de los límites tradicionales del agente, sino que pasa a formar parte del propio proceso. El sistema cognitivo es "más grande" que el agente.

Esta tesis puede interpretarse de diferentes formas. Para unos, el procesamiento cognitivo ocurre fuera del cerebro e involucra recursos extra-craneales. Ya no hay varios *locus* de cognición distribuidos sino uno solo, pero sus límites no se corresponden con los del cerebro. Para otros es más bien que hay partes del entorno que deberíamos interpretar como partes del sistema cognitivo, lo que nos llevaría a extender el sistema cognitivo más allá del sistema nervioso del agente -a pesar de que la cognición no ocurra de hecho en el entorno-. Sea cual sea la opción que más nos convenza, es importante seguir distinguiendo entre aquello que forma parte de un sistema y aquello que lo influye sin llegar a formar parte de él.

Por último, el enfoque enactivo parte de la autonomía del sistema cognitivo. El agente, situado e inmerso en un entorno, crea un mundo de significados en su mismo ser y accionar (*enact*) en su entorno. El agente no recibe pasivamente información neutra

de un entorno a la cual deba sumar después un significado. Significado, vida, mente, sociedad... no son capas que se superponen sino hilos que se entretejen y que deben estar ya a la base de la propia enunciación del problema. Así, "el problema", si pretendemos un estudio científico de la mente, queda de nuevo desplazado. La cuestión es cómo naturalizar todos estos conceptos y, muy en particular, lo mental.

Dentro de la ciencia cognitiva enactiva encontramos a su vez varias corrientes, pero ahora nos detendremos en profundizar en el llamado enactivismo autopoiético. Esta versión del enactivismo, fuertemente inspirada en la biodinámica de los sistemas vivos, parte, como su nombre bien indica, de la teoría de la autopoeisis. Según ésta, un organismo autopoiético es capaz de autoproducirse en el sentido de que todas las operaciones que realiza resultan en el mantenimiento de las condiciones que dan lugar a su conservación. Además, por esta misma razón, es autónomo y se distingue de base de cualquier otro organismo.

Vemos que el lenguaje ha cambiado completamente y, como adelantábamos, ahora nos movemos en un contexto biológico. Pero para los defensores del enactivismo autopoiético, este movimiento es precisamente el que necesitamos para lograr una naturalización de todos los conceptos psicológicos y cognitivos que venimos arrastrando. Sin ir más lejos, el sentido, ese concepto tan esquivo para las teorías analíticas y formalistas, aquí queda resuelto en el mundo de significados que emerge a partir de la actividad de los seres vivos en su entorno. Un agente interactúa con su entorno según su organismo se lo permite y necesita para su propia supervivencia (pensemos aquí en una célula que se mueve en su entorno siguiendo el camino más corto hacia los nutrientes que necesita): La "masiva multidimensionalidad del acople físico con el entorno" (Di Paolo, 2024) se reduce a unas pocas dimensiones según lo que es apto para la autopoiesis del organismo.

La cognición se piensa tradicionalmente desde el modelo de procesamiento de información en el cerebro. Según el enfoque enactivo, la cognición no es pensamiento sino interacción. Una actividad continua, resultado de la constante participación en el mundo y la búsqueda de homeostasis. Lo mental, por tanto, aparece como relacional. Y en tanto que relacional, no posee ubicación espacial. El objeto de interés pasa de ser el cerebro –y su procesamiento de información- al organismo en su conjunto –y su interacción con el entorno-.

A pesar del éxito de las soluciones computacionalistas que vemos hoy en día, los problemas de su marco teórico son bien conocidos. Las teorías críticas aquí esbozadas, lejos de estar ellas exentas de problemas, suponen un soplo de aire fresco al tradicional problema de la mente. Estoy seguro de que, más allá de las modas tecnológicas, cualquier avance en la dirección hacia una mejor comprensión de la mente y, en definitiva, del ser humano, pasará por una mejor comprensión del papel del cuerpo en la cognición y por una naturalización de lo mental que explique lo mental a partir de la biología.

6. Conclusión.

Hemos analizado la cuestión de la IA desde tres ángulos diferentes pero complementarios: el cine, la ingeniería y la filosofía. El cine en general, y la ciencia ficción en particular, tiene un papel protagonista en la construcción de imaginarios colectivos sobre los mundos que están por venir. Hemos rescatado algunas de las ideas que subyacen en los planteamientos de estas historias, especialmente respecto a nuestra relación con "lo otro" a partir de la máquina, y hemos caracterizado a partir de ellas el paradigma computacionalista.

Este paradigma ha ayudado a abordar la cuestión de la IA desde la ingeniería, posibilitado una gran cantidad de desarrollos técnicos impresionantes, aunque al precio de reducir la complejidad de la inteligencia a procesos computacionales. La superación -o al menos ampliación- de este paradigma mediante las críticas corporeizadas debería ser capaz de llevar la cuestión acerca de qué es la inteligencia y el ser humano a nuevos horizontes. Esta sigue siendo la pregunta fundamental a la base de todo.

Por muy metafísica que nos resuene la pregunta, es probable que su respuesta nunca haya tenido implicaciones tan relevantes. Nuestra comprensión de la inteligencia debe ir más allá de su capacidad de cálculo hasta alcanzar la subjetividad y sus aspectos afectivos y sociales. Un enfoque reduccionista puede llevarnos a perder de vista aspectos fundamentales de la mente humana. En un contexto de acelerado crecimiento de las capacidades de la IA como este, no debemos perder el espacio para la reflexión y la crítica. La tarea de la filosofía no es solo mantener su voz en la conversación, sino también trabajar por ensanchar el dominio de lo humano en cualquiera de los mundos que estén por venir.

7. Bibliografía.

Angius, N., Primiero, G., y Turner, R. (2024): "The Philosophy of Computer Science", en E. N. Zalta & U. Nodelman (eds.): *The Stanford Encyclopedia of Philosophy*, Summer 2024 Edition en:
https://plato.stanford.edu/archives/sum2024/entries/computer-science/

Automatic Language Processing Advisory Committee (ALPAC) (1966): "Language and Machines: Computers in Translation and Linguistics". Washington, DC: National Academy of Sciences, National Research Council.

Bostrom, N. (2024): *Nick Bostrom hizo al mundo temer por la IA. Ahora pregunta: ¿Y si es la solución a todos nuestros problemas?* 4 de Mayo de 2024 en:
https://es.wired.com/articulos/nick-bostrom-hizo-al-mundo-temer-por-la-ia-ahora-pregunta-y-si-es-la-solucion-a-todos-nuestros-problemas

Bringsjord, S. y Govindarajulu, N. S. (2024): "Artificial Intelligence", en E. N. Zalta & U. Nodelman (eds.): *The Stanford Encyclopedia of Philosophy*, Summer 2024 Edition en:
https://plato.stanford.edu/archives/sum2024/entries/artificial-intelligence/

Broncano, F. (2024): "Tigres de papel. IA y la amenaza de la singularidad". *Revista de la Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España* 68: 48-53.

Casacuberta, D. (2024): "Hasta la vista, Singularidad". *Revista de la Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España* 68: 24-25.

Chalmers, D. L. (2010): "The singularity: A philosophical analysis". *Journal of Consciousness Studies* 17 (9-10), 7-65. [Trad. esp.: "La singularidad: un análisis filosófico", en D. Pérez Chico (ed.): *Hacia una concepción integral de la mente. Más allá del cognitivismo*, Zaragoza: PUZ, 2024].

Di Paolo. E. (2024): "El enfoque enactivo", en Pérez Chico, D. (ed.), *Hacia una concepción integral de la mente. Más allá del cognitivismo*, Zaragoza: PUZ, 2024.

Español, C. (próximamente): "Dejarse la piel en la pantalla. Reflexiones en torno a la representación del cuerpo en el cine".

Hoffman, R. with Gpt-4 (2023): *Improptu. Amplifying Our Humanity Through AI*. Dallepedia LLC, 2023.

Kroon, F. y Voltolini, A. (2024): "Fiction", en E. N. Zalta & U. Nodelman (eds.): *The Stanford Encyclopedia of Philosophy*, Summer 2024 Edition en: https://plato.stanford.edu/archives/sum2024/entries/fiction/

Larson, E. (2023): *El mito de la inteligencia artificial. Por qué las máquinas no pueden pensar como nosotros lo hacemos*. XXXX: Shackleton, 2023.

---- (2023): *Los avances actuales no nos acercan más a tener una inteligencia artificial similar a la humana*. 3 de Marzo de 2023 en: https://www.eldiario.es/tecnologia/erik-larson-avances-actuales-no-acercan-inteligencia-artificial-similar-humana_128_10001811.html

Lighthill, J. (1973): "Artificial Intelligence: A General Survey", en J. Lighthill (Ed.), *Artificial Intelligence: a paper symposium* (pp. 1-77). Science Research Council.

Liz, M. (2024): "El funcionalismo necesario", en Pérez Chico, D. (ed.), *Hacia una concepción integral de la mente. Más allá del cognitivismo*, Zaragoza: PUZ, 2024.

Pérez Chico, D. (2024): "Cognitivismo y postcognitivismo. Una visión preliminar", en Pérez Chico, D. (ed.), *Hacia una concepción integral de la mente. Más allá del cognitivismo*, Zaragoza: PUZ, 2024.

---- (2024): "La singularidad tecnológica, ¿mito, realidad inevitable o profecía autocumplida?" en Serón Arbeloa, F. J. (ed.), *Proyecto UNIDIGITAL IASAC*. 2023 en: https://unidigitaliasac.unizar.es/ficha/la-singularidad-tecnologica-una-profecia-autocumplida/

Pitt, D. (2022): "Mental Representation", en E. N. Zalta & U. Nodelman (eds.): *The Stanford Encyclopedia of Philosophy*, Fall 2022 Edition en: https://plato.stanford.edu/archives/fall2022/entries/mental-representation/

Rescorla, M. (2020): "The Computational Theory of Mind", en E. N. Zalta (ed.): *The Stanford Encyclopedia of Philosophy*, Fall 2020 Edition en: https://plato.stanford.edu/archives/fall2020/entries/computational-mind/

Robinson, H. (2023): "Dualism", en E. N. Zalta & U. Nodelman (eds.): *The Stanford Encyclopedia of Philosophy*, Spring 2023 Edition en: https://plato.stanford.edu/archives/spr2023/entries/dualism/

Shapiro, L. (2012): "What's New About Embodied Cognition". *Filosofia Unisinos* 13 (2-suppl.) en https://doi.org/10.4013/fsu.2012.132(suppl).01. [Trad. esp.: "¿Cuál es la novedad de la cognición corporizada?", en D. Pérez Chico (ed.): *Hacia una concepción integral de la mente. Más allá del cognitivismo*, Zaragoza: PUZ, 2024].

---- y Spaulding, S. (2024): "Embodied Cognition", en E. N. Zalta & U. Nodelman (eds.): *The Stanford Encyclopedia of Philosophy*, Summer 2024 Edition en: https://plato.stanford.edu/archives/sum2024/entries/embodied-cognition/

Speaks, J. (2021): "Theories of Meaning", en E. N. Zalta (ed.): *The Stanford Encyclopedia of Philosophy*, Spring 2021 Edition en: https://plato.stanford.edu/archives/spr2021/entries/meaning/

Van Gulick, R. (2022): "Consciousness", en E. N. Zalta & U. Nodelman (eds.): *The Stanford Encyclopedia of Philosophy*, Winter 2022 Edition en: https://plato.stanford.edu/archives/win2022/entries/consciousness/

Vega Encabo. J. (2024): "Extensiones cognitivas", en Pérez Chico, D. (ed.), *Hacia una concepción integral de la mente. Más allá del cognitivismo*, Zaragoza: PUZ, 2024.

Waldrop, M. M. (1992): *Complexity. The emerging science at the edge of order and chaos.* New York: Simon & Schuster Paperbacks, 1992.

Wikipedia contributors. (2024): "History of artificial intelligence". Mayo de 2024, en: https://en.wikipedia.org/wiki/History_of_artificial_intelligence/

Wu, W. y Morales, J. (2024): "The Neuroscience of Consciousness", en E. N. Zalta & U. Nodelman (eds.): *The Stanford Encyclopedia of Philosophy*, Summer 2024 Edition en: https://plato.stanford.edu/archives/sum2024/entries/consciousness-neuroscience/