Estudios Filosóficos LXII (2013) 115 ~ 130

EL FRACASO DE LA INTELIGENCIA ARTIFICIAL COMPUTACIONALISTA Y SU POSIBLE SUPERACIÓN: UNA APROXIMACIÓN METAFÍSICA

Miguel Penas López
Universidad Autónoma de Barcelona
Université de Toulouse II-Le Mirail

Resumen: *las investigaciones en Inteligencia Artificial guiadas por el paradigma computacionalista, dominante en las primeras décadas de las ciencias cognitivas, no han logrado cumplir sus optimistas previsiones. En este artículo tratamos de desvelar, desde una aproximación metafísica, las causas de dicho fracaso y de mostrar que dichas causas tratan de ser superadas por un nuevo paradigma surgido en las ciencias cognitivas que desarrolla el proyecto de la Inteligencia Artificial desde unos fundamentos radicalmente diferentes.*

Palabras clave: *Inteligencia Artificial, ciencias cognitivas, computacionalismo, metafísica subjetivista, representacionalismo.*

Abstract: *Research on Artificial Intelligence guided by the computational paradigm, dominant in the first decades of cognitive sciences, has failed to meet its optimistic predictions. In this article we try to explain, from a metaphysical approach, the causes of this failure and to show how a new paradigm, emerged in cognitive sciences and developing the project of Artificial Intelligence from radically different foundations, tries to overcome those causes.*

Keywords: *Artificial Intelligence, cognitive sciences, computationalism, subjectivist metaphysics, representationalism.*

1. Introducción histórico-crítica a la Inteligencia Artificial

En 1968, durante la presentación ante la prensa de la película *2001: A Space Odissey*, Marvin Minsky[1], colaborador de Stanley Kubrick en dicha película, proclamó: "en una generación tendremos computadoras inteligentes como HAL en la película, *2001*"[2]. Han pasado más de 40 años desde entonces y la consumación del proyecto de la IA, que según las previsiones de Minsky sería lograda por la siguiente generación de investigadores y, según la insinuación un poco más pesimista de Kubrick, en 2001, no se ha producido. La intención del presente artículo es ofrecer, desde una aproximación metafísica, una explicación del fracaso de dichas expectativas y mostrar que sus causas, que podemos resumir en una cierta concepción de lo que es el comportamiento inteligente que ha guiado la investigación en IA durante décadas, tratan de ser superadas por un nuevo paradigma nacido dentro de las ciencias cognitivas que intenta llevar a cabo el proyecto de la IA desde una aproximación bien diferente.

Para ello debemos, en primer lugar, tratar de clarificar qué es la IA y cuál ha sido su desarrollo. Podemos definir la IA como el intento de simular en algún tipo de artefacto –sea una computadora digital o análoga, un robot o los algoritmos evolucionarios propios de las redes neuronales artificiales– el comportamiento inteligente humano. En su origen poseía un doble propósito, el cual a partir de los años 70 supuso una bifurcación de caminos[3], a saber: por un lado, una utilidad teorética, i.e. una herramienta que permitiría comprender la naturaleza de la inteligencia humana y testar empíricamente las teorías disponibles acerca de ella; por otro, la creación de sistemas y herramientas que tuvieran una utilidad práctica (no en vano, la principal fuente de financiación del *AI Lab* del MIT fue, en su origen, el Ministerio de Defensa estadounidense). El hecho de que ambas finalidades se hayan bifurcado es ya un síntoma de las dificultades a las que se enfrenta la IA, pues si bien no poseemos todavía una explicación suficientemente integral de lo que es la inteligencia, el desarrollo de utilidades prácticas en dominios concretos ha tenido un relativo éxito. Puesto que el interés aquí es ofrecer una comprensión filosófica de la IA, nos centraremos en la primera finalidad.

El primer trabajo dentro del campo de la IA, aunque entonces todavía no existía la disciplina con ese nombre, se remonta a 1943 con la creación de un modelo de neuronas artificiales por parte de Warren McCulloch y Walter Pitts.

[1] Minsky es considerado como una de las principales figuras de la investigación en Inteligencia Artificial (IA). Ha sido co-director del Laboratorio de IA (*AI Lab*) del MIT (Instituto Tecnológico de Massachusetts), cuna de la IA, entre 1959 y 1974.

[2] "Within a generation we will have intelligent computers like HAL in the film, 2001". Citado en Hubert L. Dreyfus, "Why Heideggerian AI failed and how fixing it would require making it more Heideggerian", en Philip HUSBANDS y otros (eds.), *The Mechanical Mind in History*, Cambridge, The MIT Press, 2008, p. 331.

[3] *Cf.* Fernando Flores & Terry Winograd, *Understanding computers and cognition. A new foundation for design*, Norwood, Ablex Corporation, 1990, cap. 10.

En 1950, Alan Turing planteó explícitamente la posibilidad de la IA al preguntarse si las máquinas pueden pensar. Para ello, ideó la famosa prueba de Turing[4], un sencillo experimento que serviría para determinar si la actuación de una máquina es indistinguible de la de un ser humano[5]. Desde los años 50 hasta mediados de los 70 la investigación en IA estuvo guiada por el paradigma triunfante en las ciencias cognitivas, el computacionalismo, el cual está basado en la idea de que el funcionamiento de nuestra mente es análogo al de las computadoras.

El computacionalismo concibe la mente como un procesador de información que efectúa representaciones simbólicas de los *inputs* recibidos por ella y que maneja esas representaciones de acuerdo a unas reglas sintácticas, lo que le permite generar los *outputs* deseados. La analogía con las computadoras es clara, aunque es necesario distinguir entre los dos tipos de computadoras existentes: las análogas, cuyas unidades de información radican en un soporte físico y son continuas, y las digitales que, tal como su nombre indica, formalizan la información en unidades numéricas que sólo poseen dos posiciones, uno y cero, y son, por tanto, discretas –cada unidad mínima es una entidad aislada con una posición fija–. El modelo que ha triunfado históricamente es el de las computadoras digitales y esto, como veremos, tendrá serias consecuencias para la IA, pues implicará que la simulación de la inteligencia sea modelada en función de unas determinadas posibilidades abiertas por la ingeniería computacional. La inteligencia es concebida de una manera intelectualista como el manejo, por medio de reglas sintácticas, de símbolos que representan o están-por rasgos aislados y bien definidos de la realidad que se pretende formalizar. El uso de computadoras digitales en IA implica una concepción de la realidad como algo que se puede analizar o descomponer en unidades aisladas e independientes (objetos y propiedades) que pueden ser expresadas en un lenguaje formal. La cognición es concebida como un proceso abstracto, simbólico, representacional, cuyas unidades de información son independientes del contexto: se sueña con un lenguaje formal neutral, libre de toda ambigüedad (y, por ende, de toda interpretación), que pueda representar la realidad. Se sueña con la posibilidad de la reducción del comportamiento inteligente humano, esto es, su saber-hacer (*know-how*) a un mero saber-qué (*know-what*), un conjunto de hechos y reglas que pueda ser aplicado en todos los casos.

No entraremos aquí a una exposición detallada del desarrollo histórico de la IA. Desde los años 50 hasta los 70, las investigaciones, guiadas por el

[4] Cf. Alan M. Turing, "Computer Machinery and Intelligence", en *Mind* 49 (1950) 433-460.

[5] En la prueba, se mantiene un diálogo en el que se puede mentir entre un juez en una habitación, y una persona y una máquina que están en otra habitación. Si el juez es incapaz de distinguir quién es la persona y quién la máquina, entonces ésta habría pasado el test de Turing y sería considerada inteligente. John Searle contestó a esta versión computacionalista de la inteligencia con un contra-experimento denominado la habitación china en el que trata de mostrar que una máquina puede pasar el test de Turing sin tener ninguna comprensión de lo que ha hecho y sin que pueda, por tanto, ser considerada inteligente.

computacionalismo, se dirigieron principalmente a los siguientes ámbitos: la resolución de problemas (*General Problem Solving* de Newell y Simon), el reconocimiento de formas (*pattern recognition*), la traducción de lenguas (p. ej. el programa STUDENT de Bobrow), la creación de micro-mundos en los que los programas se puedan desenvolver (el más exitoso fue el SHRDLU de Winograd, capaz de entender órdenes y de responder adecuadamente a ellas en inglés dentro de un micro-mundo de bloques) y la creación de programas para jugar a determinados juegos, principalmente el ajedrez. En general, la metodología seguida fue la de tratar de avanzar en dominios restringidos con la esperanza de que las técnicas alcanzadas pudieran ser aplicadas en ámbitos cada vez más generales. Efectivamente, se alcanzaron unos tempranos éxitos, pero el funcionamiento de las computadoras plantea problemas a la hora de tratar de generalizar esos éxitos. La implementación del conocimiento sobre un determinado ámbito es realizada en las computadoras por medio de la programación, consistente en la acumulación de datos libres de contexto relativos a objetos, propiedades y acciones sobre esos objetos. El acceso a esa masa de datos es guiado por unas reglas heurísticas que exigen un mayor tiempo de búsqueda a medida que la cantidad de datos es mayor. Y aquí es donde parece surgir una disparidad entre el funcionamiento de las computadoras y el comportamiento inteligente humano, pues en los humanos ocurre a la inversa: a medida que dominamos más un ámbito y nos hacemos expertos, nos podemos mover con una mayor facilidad en él, mientras que a las computadoras les cuesta cada vez más trabajo. ¿Por qué? Aquí es donde entra en juego un concepto central, la relevancia, el cual es dependiente del contexto. Puesto que los programas están formados por datos independientes del contexto, la máquina no puede tener un sentido de qué datos son relevantes para abordar una determinada tarea, por lo que tiene que emplear mucho tiempo en tantear heurísticamente dentro de su base de datos. Así planteado, esto puede parecer un mero problema tecnológico que podría ser superado a medida que las técnicas de búsqueda fueran mejoradas. Antes bien, lo que muestra es que el hecho de *estar situados* en un determinado contexto nos permite a los seres humanos el ahorro de un enorme esfuerzo computacional: no necesitamos pensar ni tantear heurísticamente en nuestra mente para recuperar los datos relevantes necesarios para realizar una tarea. Éstos son mostrados por el contexto[6].

Minsky captó la seriedad de estas dificultades en lo que se ha denominado el problema del sentido común. En nuestro trato con el mundo, los seres humanos manejamos una gran cantidad de conocimientos implícitos que llamamos el sentido común (p. ej. conocimientos tan simples como: "si Ramón

[6] En Martin Heidegger, *Ser y tiempo*, Madrid, Trotta, 2009 (Primera Sección, cap. tercero) encontramos un intento de explicación de este proceso. En lugar de vivir en un mundo compuesto de hechos desprovistos de significado (así es como "vive" una computadora) a los que posteriormente les asignamos un valor, el ser humano está constitutivamente abierto a (está-en) un mundo de significaciones en virtud de su estructura fundamental que es el estar-en-el-mundo. La significación, por tanto, es previa a todo discurso.

está en la habitación de al lado, entonces la nariz de Ramón está en la habitación de al lado"). Las computadoras no pueden "saber" cosas tan sencillas como esa a menos que el diseñador las escriba en su programa (imaginemos ahora el problema: ¿qué tamaño podría alcanzar la base de datos del sentido común, y de qué manera podemos acceder a los datos que nos son relevantes en cada situación?). El ser humano ya sabe implícitamente todas esas cosas porque está *situado en* o está *abierto a* un contexto en el que hay seres humanos, cada uno de los cuales posee una nariz –salvo raras excepciones–, y dichos seres humanos se encuentran a veces en una habitación, con su nariz incluida. Minsky denominó a la cuestión de cómo las máquinas podrían acceder a los datos relevantes para cada tarea el problema de los marcos (*frame problem*), el cual ha sido ampliamente abordado por la literatura sobre IA. Una manera de encarar el problema consistió en intentar dotar a los programas con unas meta-reglas de segundo orden (recordemos que un programa consiste en unas reglas sintácticas –que serían entonces las de primer orden– para manejar símbolos) que permitirían circunscribir los contextos en los que un conjunto de datos son relevantes. El problema es que esas meta-reglas serían más hechos (acciones sobre símbolos) desprovistos de significado, por lo que entraríamos en una regresión infinita de reglas para manejar reglas[7].

En general, podemos resumir los problemas aparecidos en la IA computacionalista en que el comportamiento de los artefactos producidos es totalmente dependiente de lo que el diseñador pueda escribir en su programa y el tipo de escritura de la programación está basado en esa acumulación de datos libres de contexto, esto es, universalmente definidos, de la que hablábamos. La independencia del contexto de los datos de los programas es lo que ha generado grandes problemas en el avance de las computadoras, sean éstas diseñadas para reconocer caras, traducir idiomas o jugar al ajedrez[8].

Desde mediados de los años 70, los diseñadores han tratado de idear artefactos intentando que el papel jugado por el diseñador en su comportamiento sea cada vez menor. En ese sentido han tratado de aprovechar la aparición, a principios de los años 80, de un nuevo paradigma dentro de las ciencias cognitivas, el conexionismo, con el fin de crear redes neuronales artificiales. Éstas permitirían que la intervención del diseñador fuera menor. Se han creado, por ejemplo, algoritmos matemáticos que simulan redes neuronales y que pueden aprender y evolucionar –los antes mencionados algoritmos evolucionarios–.

[7] Dreyfus (Hubert L. Dreyfus, *What Computers* Still *Can't Do: A Critique of Artificial Reason*, Cambridge, The MIT Press, 1992, cap. 5) expone las dos posibilidades abiertas por esta solución: o bien caemos en el peligro de la regresión infinita, o bien tendríamos que demostrar la existencia de unos elementos mínimos libres de toda ambigüedad –libres de contexto– que permitirían al programa funcionar a partir de ellos sin apelar a unas meta-reglas por encima de ellos. Esto no ha sido mostrado.

[8] Este último caso, el de las computadoras para jugar al ajedrez, ha sido uno de los grandes campos de batalla entre los creadores de IA y sus críticos. No debemos dejar de mencionar la histórica derrota sufrida por el campeón mundial Garry Kasparov ante la computadora Deep Blue, la cual no estuvo exenta de polémica. De todos modos, tal vez sea el terreno en el que la IA computacionalista ha llegado más lejos.

Esto ha posibilitado avances en campos como el *pattern recognition*. Sin embargo, este paradigma continúa siendo dependiente de un importante concepto que analizaremos más tarde, el de representación. Dichos artefactos siguen operando por medio de una representación simbólica del mundo, por lo que su desarrollo no está basado en un trato con el mundo mismo. No están, por tanto, *situados en un contexto*.

Esto ha dado origen al último paradigma surgido en las ciencias cognitivas, la *embodied cognition*, centrado en la idea de que la mente, la conciencia o la inteligencia no pueden ser concebidos como algo que opera independientemente por medio de representaciones simbólicas, sino que hay que concebirlos como algo unido tanto a nuestros cuerpos físicos como a nuestros entornos. Abordaremos esto al final del artículo.

2. IA y metafísica: en torno al concepto de representación

¿Por qué puede resultar interesante la IA para la filosofía? Los diferentes proyectos de realización de la IA, tanto si fracasan como si son exitosos, pueden arrojar luz a cuestiones tratadas por la filosofía desde siempre. En el caso de que fracasen, mostrarían, en principio, que los principios desde los que han sido llevados a cabo los proyectos, especialmente su comprensión de lo que es el comportamiento inteligente humano, son inadecuados. En caso de que triunfen, mostrarían que es posible simular la inteligencia humana en un artefacto, y ello revertiría en una comprensión de la inteligencia misma, aunque no hay que descartar la posibilidad de que una máquina pueda exhibir comportamiento inteligente siguiendo mecanismos diferentes a los seguidos por el ser humano. Por otra parte, y puesto que la IA está siempre sometida a consideraciones de ingeniería, podría llegarse a la conclusión, en caso de un estancamiento severo de la IA, de que la inteligencia no es reproducible por las técnicas de diseño actuales o –lo que sería más definitivo y pondría de manifiesto unos límites de la tecnología– que la inteligencia es un atributo exclusivamente humano que no es posible simular en ningún tipo de artefacto –y en este caso también obtendríamos una compresión de las causas que hacen que la inteligencia sea un atributo exclusivamente nuestro–. Esta interacción entre las teorías sobre la inteligencia y la práctica de la IA se muestra claramente en la íntima relación existente entre las ciencias cognitivas y la IA. La investigación en IA ha estado claramente dirigida por los diferentes paradigmas surgidos en las ciencias cognitivas y, asimismo, los resultados de la IA influyen en el seguimiento o el abandono de un paradigma dentro de ella[9], pues constituyen una útil manera de testar empíricamente las teorías, aunque,

[9] Un ejemplo de esto lo encontramos en Tom Froese, "On the role of AI in the ongoing paradigm shift within the cognitive sciences", en Max Lungarella y otros. (eds.), *50 Years of AI*, Berlin, Springer-Verlag, 2007, pp. 63-75, donde se expone la influencia que ha tenido la IA en el surgimiento del paradigma de la *embodied* y *enactive cognition* dentro de las ciencias cognitivas.

como apuntábamos antes, esta evaluación empírica es problemática, pues las posibilidades son diversas: puede ocurrir que una teoría plausible sobre la inteligencia fracase a la hora de implementarse en una máquina porque las técnicas disponibles no sean apropiadas, lo cual no podría tomarse como una invalidación definitiva de la teoría (este argumento ha sido esgrimido, por ejemplo, por los defensores del paradigma computacional); también podría ocurrir que una determinada implementación en una máquina obtenga éxitos cuya extrapolación con el fin de comprender la inteligencia humana no sea legítima (siguiendo la idea expuesta un poco más arriba de que no hay nada que descarte la posibilidad de exhibir un mismo comportamiento inteligente siguiendo mecanismos diferentes), lo cual pondría en duda lo que anunciamos al comienzo como la finalidad teorética de la IA: la comprensión de la inteligencia. En cualquier caso, pese a que es necesario tener en cuenta estos matices, consideramos que la IA siempre puede proporcionarnos algo de luz en el camino de la comprensión.

Las relaciones entre los investigadores de IA y la filosofía[10] resultan esclarecedoras. Cuando Terry Winograd y Hurbert Dreyfus comenzaron a introducir a Heidegger en el MIT con el fin de mostrar un nuevo camino a los investigadores, algunos alumnos mostraron reacciones del tipo: "vosotros los filósofos habéis estado reflexionando en vuestras butacas durantes más de dos mil años y todavía no entendéis cómo funciona la mente. Nosotros en el *AI Lab* hemos tomado el mando y estamos triunfando donde vosotros los filósofos habéis fracasado. Ahora estamos programando computadoras para exhibir inteligencia humana: para resolver problemas, entender el lenguaje natural, percibir, y aprender"[11].

Apreciamos aquí la presuposición de que la práctica de la IA puede ser entendida al margen de la historia del pensamiento filosófico y, lo que es más, en oposición y superación respecto a ella. Precisamente pretendemos defender lo contrario: consideramos que los presupuestos que guían el proyecto computacionalista de la IA son dependientes de manera implícita de la conformación subjetivista de la metafísica moderna. ¿Qué es la inteligencia? A pesar de que las computadoras son un invento bastante nuevo, las ideas que guían al paradigma computacionalista poseen una gran antigüedad.

A la base de la constitución del proyecto metafísico se encuentra la doctrina sustancialista según la cual todo lo que es puede ser reducido a entes o sustancias y sus propiedades. Como sabemos, la programación de una

[10] En Philip E. Agre, "Constructions of the Mind: Artificial Intelligence and the Humanities", en *Stanford Humanities Review* 4, n.2 (1995) 1-19, se realiza un interesante análisis de algunos aspectos de esta relación.

[11] *"You philosophers have been reflecting in your armchairs for over 2000 years and you still don´t understand how the mind works. We in the AI Lab have taken over and are succeding where you philosophers have failed. We are now programming computers to exhibit human intelligence: to solve problems, to understand natural language, to perceive, and to learn"*. Citado en Hubert L. Dreyfus, "Why Heideggerian AI failed and how fixing it would require making it more Heideggerian", p. 331.

computadora digital sigue este esquema para poder definir su dominio de actuación. La metafísica clásica se remite en última instancia a un principio racional que es tomado como garantía ontológica y epistemológica. En su primitiva constitución onto-teológica, este principio racional es lo máximamente ente, el ente que posee todas las actualizaciones posibles por lo que es acto puro: el Ente Supremo o Dios. En la Modernidad se opera un importante cambio en dicho esquema por medio de su constitución subjetivista. El principio racional último sobre el que descansa el edificio metafísico ya no es Dios, sino el sujeto. Descartes traza una frontera divisoria en la totalidad de lo ente entre el mundo físico o *res extensa* y el mundo psíquico o *res cogitans*. El criterio del conocimiento emana del sujeto, del yo que piensa, el cual es entendido de una manera formal y universal y cuya expresión más perfecta son las matemáticas. El conocimiento es la representación por parte del sujeto de un objeto independiente de él de acuerdo a esa legalidad universal expresada en las matemáticas. Nos preguntábamos antes qué es la inteligencia. Pues bien, el computacionalismo es la idea de que la inteligencia humana se basa en el manejo de una representación simbólica de lo real expresada en un lenguaje formal –universal, libre de contexto– cuyos elementos son objetos y propiedades sobre los que se opera de acuerdo a unas reglas sintácticas. Dicho en lenguaje metafísico: es la representación del objeto siguiendo la legalidad universal en que consiste el sujeto. Si las previsiones de Minsky hubieran sido acertadas, si a día de hoy estuviéramos familiarizados con máquinas como HAL-9000 que dirigieran las empresas o impartieran clases de filosofía en las universidades, las aspiraciones teoréticas de la metafísica moderna se habrían visto cumplidas. Pero parece que no ha sido así.

¿A qué se debe la importancia del concepto de representación? A nivel teórico, la idea de una representación del mundo se hace necesaria desde el momento en que imponemos una división entre el sujeto y el objeto, y concedemos una primacía a la relación teorética entre ambas esferas escindidas. No podemos negar que la capacidad humana de realizar representaciones simbólicas del mundo ha dado sus frutos y funciona con éxito en determinados ámbitos. Ahora bien, debemos preguntarnos lo siguiente: ¿es posible reducir, de una manera extremadamente intelectualista, el comportamiento inteligente humano a meras operaciones sintácticas sobre representaciones simbólicas de la realidad?, ¿nuestra actividad consiste en un trato con modelos simbólicos del mundo, o con el mundo mismo? Dicha pregunta está inspirada en las posturas antirrepresentacionalistas que han mantenido, por motivos diferentes, gente como Dreyfus o Brooks[12]. Aunque tal vez no sea positivo plantear una división tan tajante. La cuestión empírica acerca de si la cognición humana

[12] En Rodney A. Brooks, "Intelligence without representation", en *Artificial Intelligence* 47 (1991) 139-159, el autor nos aclara que su postura se debe a cuestiones de ingeniería exclusivamente, y no a una toma de partido filosófica. Considera que seguir el esquema "percepción (*input*)-representación mental-acción (*output*)" no es adecuado para la creación de robots pues concibe la percepción como algo directamente ligado a la acción sin que medie un centro de control ejecutivo.

utiliza representaciones del mundo es más compleja. Clark[13], aún tratando de superar la división entre la mente, el cuerpo y su entorno, considera, siguiendo un punto de vista adaptativo sobre el comportamiento humano, que el uso de representaciones es en ciertos casos útil y necesario para la economía cognitiva humana. Pero lo que queremos poner en duda aquí es si podemos basar la cognición en la representación[14]. Las aspiraciones de la IA computacionalista dependen de una respuesta afirmativa a dicha duda, es decir, de la posibilidad de concebir la inteligencia como el manejo de representaciones simbólicas del mundo que descansan, en última instancia, en un conjunto de hechos y reglas expresados en un lenguaje formal y universal independiente del contexto, libre de toda ambigüedad y, por ende, de toda interpretación. ¿Por qué consideramos que estas aspiraciones no se pueden ver cumplidas?

La posibilidad de la comprensión del sentido de una frase, de la función de un martillo, o del diagnóstico de una enfermedad a partir de unos síntomas, descansa en una pre-comprensión implícita de esos fenómenos que no es verbalizable en su totalidad. Dicha pre-comprensión constituye un *background* de asunciones implícitas que no podemos reducir a un conjunto de hechos y reglas independientes del contexto. Dicho de una manera más simple: la destreza humana, su saber-hacer, no puede ser recogida en un manual. No puede existir el manual del buen-filósofo, del buen-médico o del buen-músico. Por eso constituye una pretensión fabulosa el intento de implementar la inteligencia humana en una computadora digital. Podemos comprender el mundo porque estamos situados en él –i.e. porque estamos existencialmente abiertos a sus significaciones–, y no porque podamos hacernos representaciones de él. Por tanto, las características de la actividad inteligente, que presupone el poder-estar situados en un contexto equipados con una sabiduría implícita que no es –al menos en su totalidad– verbalizable y la cual nos permite *interpretar* los fenómenos, parecen abocar a la superación de la creación de IA basada en presupuestos computacionalistas o exclusivamente representacionalistas. Estos intentos de superación son los que expondremos a continuación.

3. Una nueva metafísica, un nuevo paradigma: *embodied* y *enactive cognition*

Según la perspectiva que hemos seguido, la historia de la metafísica constituye un lugar de explicación adecuado de cara a la comprensión de los fundamentos y las limitaciones de los diversos paradigmas desde los cuales se ha tratado de llevar a cabo el proyecto de construcción de la IA. Como hemos

[13] Andy Clark, *Being There: Putting brain, body, and world together again*, Cambridge, The MIT Press, 1997.

[14] Winograd y Flores defienden abiertamente que no es así: "La cognición humana incluye el uso de representaciones, pero no está basada en la representación" ("*Human cognition includes the use of representations, but is not based on representation*". Fernando Flores & Terry Winograd, *op. cit.*, p. 99). En este libro, atento tanto a la tradición filosófica como a las peculiaridades de la ingeniería computacional, se realiza una crítica de la tradición racionalista e intelectualista que alimenta los postulados del computacionalismo.

visto, consideramos que los presupuestos que alimentan el primer paradigma dominante en las ciencias cognitivas, el computacionalismo –y asimismo, en cierto sentido, el segundo paradigma: el conexionismo–, encajan con las aspiraciones subjetivistas propias de la metafísica moderna. Desde el siglo XIX, esta metafísica, asentada en la división entre sujeto y objeto y en una comprensión intelectualista del sujeto por la cual el conocimiento que éste pueda tener del objeto es entendido en términos de representaciones simbólicas, formales y universales cuya forma más perfecta es la representación matemática, entra, como es bien sabido, en crisis. Para poner a prueba nuestra hipótesis, según la cual es posible comprender la problemática suscitada por la IA atendiendo a los presupuestos metafísicos presentes en ella, podemos comprobar en qué medida las nuevas propuestas metafísicas surgidas a raíz de dicha crisis se encuentran presentes o resuenan en los intentos de creación de IA que someten a crítica las aspiraciones del computacionalismo, ya que estos intentos, como veremos, no son asimilables a una metafísica de la subjetividad. Pues bien, lo que queremos defender es que el surgimiento en las últimas dos décadas de este nuevo paradigma dentro de las ciencias cognitivas, denominado *embodied cognition* o poscognitivismo, el cual trata de hacer frente, tal como lo expone Froese[15], al estancamiento de la IA basada en el paradigma computacionalista[16], parece confirmar nuestra hipótesis.

De entrada, debemos aclarar que no pretendemos realizar una exposición lineal en la que una determinada perspectiva o paradigma metafísico supuestamente erróneo es sustituido por otro paradigma que finalmente es considerado el correcto, pues consideramos que la corrección no es una categoría útil para orientarnos en cuestiones metafísicas. La cuestión es apreciar rigurosamente el poder explicativo que poseen las creaciones especulativas. Si hemos sometido a crítica la IA computacionalista, y los presupuestos metafísicos que según nuestra perspectiva subyacen en ella, no es tanto para defender que sus propuestas son incorrectas como para mostrar que se trata de un poder que ha tratado de exceder sus límites. Esto explica, por una parte, los éxitos que ha logrado el computacionalismo en determinados ámbitos y, por otra, el fracaso de sus desmesuradas aspiraciones. Es esta extralimitación la que asimismo explica y justifica el surgimiento del poscognitivismo. Nuestra tesis es que la inteligencia no puede ser reducida a operaciones sintácticas sobre representaciones simbólicas de lo real; si bien ciertos procesos cognitivos pueden ser comprendidos siguiendo este esquema (y de ahí que la IA computacionalista posea un ámbito en el que efectivamente funciona, véase como ejemplo lo comentado en la nota 8), consideramos que lo que se deja aquí de lado impide alcanzar una comprensión integral de la inteligencia. Y el paradigma *embodied* nos parece interesante porque pone de relevancia qué es lo que se ha dejado de lado.

15 Tom Froese, *op. cit.*

16 Este paradigma, que como hemos visto ha sido el dominante en la IA desde sus orígenes, también recibe el nombre de cognitivismo, de ahí que el nuevo paradigma, el cual trata de ser una auténtica superación de él (a diferencia del conexionismo), sea denominado también como poscognitivismo.

3.1. Sujetos con cuerpos: el rechazo del intelectualismo

Una constante a lo largo de la historia de la metafísica ha consistido en el rechazo del cuerpo y, en general, de la realidad material. Tanto en su primitiva constitución onto-teológica como en su moderno proyecto subjetivista, el principio ontológico último es considerado como una realidad intelectual, abstracta, despojada de toda realidad material. El sujeto de la metafísica moderna no consiste en los diversos sujetos empíricos, sino que es un sujeto formal y universal: la razón. Es, por tanto, un sujeto *sin cuerpo*; el pensamiento es un proceso puramente intelectual que está desligado de nuestra constitución como seres biológicos. No hay continuidad entre biología y psiquismo, y esto explica el antropocentrismo presente en esta metafísica: los seres vivos son meros cuerpos; el ser humano, sin embargo, es un ser dotado de razón.

Nietzsche introduce una quiebra en la historia de la metafísica al poner en el centro de su filosofía el concepto de cuerpo[17]. El pensamiento ya no puede ser entendido al margen de nuestra constitución como seres biológicos con unas necesidades y unos objetivos. Esto explica la constante aparición, en la obra de Nietzsche, de la cuestión de los instintos; la repetida afirmación, a modo de recordatorio, de que el ser humano es un *animal* y, en general, la articulación que trata de establecer entre procesos intelectuales y procesos fisiológicos. A pesar de que la filosofía de Nietzsche no es reconocida como una influencia directa por el paradigma *embodied*[18], podemos afirmar que estas preocupaciones están a la base de su constitución, tal como su nombre indica. Frente a la concepción de la percepción como representación y de la cognición como computación propia del computacionalismo, esta denominación[19] apunta a la importancia que posee, en los procesos cognitivos, el hecho de estar situados corporalmente en un entorno con el que interactuamos continuamente, por lo que la férrea división entre el cuerpo, la mente, y el entorno, deja de ser válida. La cognición se concibe como un fenómeno biológico en el cual el organismo es visto como un sistema dinámico que se auto-organiza en virtud de su indisoluble unión con el medio. Esta apreciación de una continuidad entre vida, mente y cognición está fundamentada en la teoría biológica de la autopoiesis de Maturana y Varela, cuyo origen se encuentra, entre otros elementos, en las investigaciones del propio Maturana sobre la percepción de las ranas. Maturana se encuentra con que no podemos seguir hablando de sistemas que perciben una realidad externa objetiva e independiente del

17 De cara a una comprensión de las implicaciones de esta quiebra, véase Barbara Stiegler, *Nietzsche et la biologie*, Paris, PUF, 2001.

18 Para una exposición de estas influencias, véase el Prefacio de Francisco Varela a la 2ª edición de la siguiente obra: Humberto Maturana & Francisco Varela, *De máquinas y seres vivos. Autopoiesis: la organización de lo vivo*, Santiago de Chile, Editorial Universitaria, 1998 5ª ed., pp. 34-61.

19 Resulta difícil hallar una traducción al castellano del término *embodied* que recoja la idea que pretende transmitir. Posibles soluciones son "encarnado", "incorporado" o "corporeizado", pero ninguna de ellas nos parece satisfactoria; por ello preferimos dejarlo en inglés. *Enactive*, por su parte, es perfectamente traducible como "enactivo".

observador. No hay una conexión directa entre las longitudes de onda que afectan a la retina de la rana y las percepciones generadas en su sistema nervioso. Antes bien, éstas son el resultado de las relaciones entre los sistemas neuronales concebidos como un todo. Las perturbaciones provocadas por el exterior alteran las neuronas, pero es la estructura de éstas la que genera las percepciones: "las perturbaciones no determinan lo que ocurre en el sistema nervioso, sino que simplemente provocan cambios de estado. Es la estructura del sistema perturbado la que determina o, mejor dicho, *especifica* qué configuración estructural del medio puede perturbarlo"[20].

Esta última afirmación apunta a que la relación del organismo con el entorno no es una relación estructural que pueda ser explicada como una relación de referencia entre las estructuras neuronales y un mundo independiente, sino que es una relación histórica. Es la historia de la unión estructural (*structural coupling*) del organismo con su entorno la que especifica qué estímulos pueden afectarle y la percepción no es la representación de dicho estímulo, sino que es el producto de la auto-organización neuronal del organismo.

A partir de estas ideas, Maturana y Varela desarrollan la teoría biológica de la autopoiesis. Se conciben los organismos vivos como entidades que alcanzan su autonomía por medio de la auto-constitución, bajo unas condiciones precarias, de sus estructuras, caracterizadas por una clausura organizacional. Las estructuras de los seres vivos generan y destruyen sus componentes para poder mantener un equilibrio (homeostasis). Hay una unión estructural con el entorno en la cual el sistema persigue su fin de no desintegrarse (de ahí que se hable de esa clausura organizacional) por medio de reajustes estructurales que no pueden ser concebidos como meras representaciones de un mundo externo. La continua auto-constitución del sistema, que es lo que asegura su pervivencia, es provocada por su unión estructural con el entorno. Estas ideas constituyen un rechazo del behaviorismo que reduce la conducta a la reacción ante unos estímulos externos. Se le da la primacía a las estructuras internas del sistema, que son las que, como dijimos, especifican qué estímulos pueden afectarle y se auto-estructuran en base a esos estímulos recibidos, por lo que la división interior-exterior, organismo-entorno, se difumina, aunque eso no impide que se le de una gran importancia a dicha auto-estructuración (la cual sería la parte interna del proceso, pero en este caso sólo cobra sentido por su unión con el exterior). La cognición ya no puede ser considerada como el manejo por parte de una mente independiente del cuerpo y del entorno de unas representaciones simbólicas de una realidad externa, independiente y objetiva, sino como la auto-constitución de estructuras que están unidas a su entorno.

Esto nos lleva a la apreciación de una de las grandes influencias presentes en el paradigma que nos ocupa, la cual es en este caso reconocida explícitamente.

[20] "*The perturbations do not determine what happens in the nervous system, but merely trigger changes of state. It is the structure of the perturbed system that determines or, better,* specifies *what structural configuration of the medium can perturb it*". Fernando Flores & Terry Winograd, *op. cit.*, p. 43.

La filosofía fenomenológica, y en especial el pensamiento de Heidegger, constituyen su principal referencia metafísica, y por ello Dreyfus llega a hablar de una IA heideggeriana. A pesar de que se podría señalar la presencia del antropocentrismo propio de la metafísica subjetivista en el privilegio ontológico que Heidegger concede al *Dasein*, el análisis de éste realizado en *Ser y Tiempo* es aprovechado como punto de partida para una explicación no intelectualista de la cognición. La estructura fundamental del *Dasein* como estar-en-el-mundo aspira al quebrantamiento de la división sujeto-objeto de la metafísica moderna. Esta estructura nos muestra que nuestra relación primaria con el entorno no es una relación teorética en la que operaríamos por medio de un cálculo sobre representaciones simbólicas, sino que es una relación práctica en la cual actividad del sujeto y significación del objeto están intrínsecamente unidos: estamos existencialmente abiertos a un mundo de significaciones (véase nota 6). Únicamente cuando se produce una avería, nos dice Heidegger, un fallo en esta relación práctica, es cuando el objeto se tematiza expresamente, marcando una distancia teorética con él. El saber-habérselas, saber-hacer o saber-cómo, antecede ontológicamente al saber-qué. El proceso de la actividad inteligente no consiste en una descomposición de la realidad en unidades a las que les atribuimos una representación simbólica para poder realizar cálculos por medio de los cuales obtendríamos la respuesta deseada, esto es, la actividad buscada. Para ser más rigurosos, la actividad inteligente no puede ser reducida por completo a ello. Los procesos intelectuales abstractos, desligados de la interacción física con el entorno, juegan un cierto papel en las tareas cognitivas y pueden ser más o menos prominentes según la naturaleza de la tarea. Pero dichas tareas cognitivas no son desligables de nuestra constitución como seres biológicos con unas necesidades y objetivos que dotan de significación a nuestro entorno ni pueden ser comprendidos al margen de la interacción con este último; se ha tratado de mostrar, en autores como Lakoff, que hasta los pensamientos más abstractos están relacionados con metáforas espaciales basadas en dicha interacción.

Podemos acudir a un ejemplo para explicar esto. En la película *Matrix* hay una famosa escena en la que se plasma la esencia del computacionalismo. En ella, la protagonista Trinity necesita aprender a pilotar un helicóptero y solicita que carguen, debemos suponer que en su cerebro, el programa informático que le permita hacerlo, es decir, se presupone que el saber-hacer, la actividad inteligente en que consiste pilotar un helicóptero es reductible a un conjunto de reglas abstractas aplicadas a un conjunto de símbolos. Esto es exactamente lo que consideramos que no es posible. Para aprender a pilotar un helicóptero, *hay que pilotarlo*: pilotar incluye la interacción física con el aparato, el manejo de sus elementos, y asimismo obliga a tener en cuenta la interacción del aparato con su entorno en diferentes condiciones (viento, visibilidad etc.), lo cual nos obliga a desarrollar una destreza práctica. Ciertos conocimientos abstractos son, por supuesto, necesarios, pero en ningún caso pueden ser considerados como suficientes.

Los intentos de construcción de IA a partir de las premisas *embodied* revisten un carácter bien diferente. Como hemos visto, la corriente *embodied* ha puesto el acento en la necesidad de abordar el problema de la situacionalidad (*situatedness*), es decir, considerar como requisito de la inteligencia, a diferencia del cognitivismo, el *estar situados en el mundo*; es este estar ya siempre situados lo que confiere una relevancia, una significatividad a los elementos del entorno. Por ello la IA no puede limitarse a la implementación de programas informáticos en una máquina; en lugar de que el diseñador determine por completo y de manera externa el comportamiento de la máquina, ésta ha de poseer un cuerpo que perciba e interactúe con el entorno a fin de desarrollar, por medio del aprendizaje, las destrezas prácticas que le permitan alcanzar los objetivos buscados. Así, se ha creado *embodied* IA incluyendo circuitos sensomotores cerrados que son sensibles al contexto combinados con algoritmos evolucionarios. Sin embargo, la creación de IA basada en las ideas *embodied* también se ha topado con dificultades, tal como es expuesto por Dreyfus[21]. Estos problemas tratan de ser abordados en el desarrollo de la corriente *embodied* realizado por las ideas enactivas.

3.2. Cognición y emergencia: una alternativa al subjetivismo representacionalista

La imposición de una férrea división entre sujeto y objeto implica unas graves consecuencias que hemos tratado de desvelar a lo largo del artículo. A nivel ontológico, se establece una distancia insalvable entre pensamiento y materia, así como entre constitución biológica y realidad psíquica. La imposibilidad para apreciar una continuidad entre las diversas esferas de lo real deja como única vía posible la reducción de unas esferas a otras; irónicamente, esta reducción se realiza en ambas direcciones: el objeto es reducido a las categorías intelectuales del sujeto y, a su vez, el pensamiento del sujeto, su realidad psíquica, es reducida al correlato material en el que se supone que tiene asiento: el cerebro.

El concepto de emergencia ofrece una vía de escape en la que la reducción es sustituida por la articulación. El necesario reconocimiento de la diferencia, de la pluralidad de las diversas esferas (materia-vida, vida-pensamiento, organismo-entorno etc.) no implica, en este caso, el establecimiento de una división entre ellas ni la reducción de unas a otras: la apreciación de una continuidad no está reñida con el respeto a la pluralidad. Las propiedades emergentes de un sistema son propiedades que no son reductibles a las propiedades de los componentes del sistema, es decir, las interacciones entre los componentes del sistema dan lugar a la aparición de propiedades nuevas. Por ello el concepto de emergencia permite dar cuenta de la articulación entre diversos elementos sin reducir unos elementos a otros, ya que aparecen nuevos niveles de organización que engloban a los anteriores.

La cognición puede ser entendida como uno de esos niveles de organización que emergen a partir de la interacción entre diversas esferas. Esto se

[21] Hubert L. Dreyfus, "Why Heideggerian AI failed and how fixing it would require making it more Heideggerian".

aprecia claramente en otro de los elementos recogidos por el paradigma *embodied*: la teoría de la mente extendida de Andy Clark y David Chalmers[22]. En ella la mente no se entiende como algo limitado por "la carne y el cráneo", sino como el resultado de la interacción presente en el *continuum* ontológico mente-cuerpo-entorno. Cobra fuerza la idea de un sistema que engloba a dicho *continuum* en el que las propiedades emergentes tienen una gran importancia y la idea de que una de esas esferas del sistema –la mente– representa la otra –el mundo– pierde fuerza. Las destrezas humanas son concebidas como la aparición, sin que haya un plan racional previo por parte de un centro organizador, de propiedades emergentes en el sistema mente-cuerpo-entorno. Y la emergencia no permite ser entendida en base a la descomposición del sistema en sus elementos constituyentes, lo que explica que la división entre estos elementos, propia de la metafísica subjetivista, no sea útil. Antes bien, consiste en la aparición de variables incontrolables que surgen o emergen a partir de la interacción entre los elementos del sistema[23].

En este punto debemos detenernos y observar el peligro que supone, tal como nos alerta Stengers[24], convertir la emergencia en un concepto todoterreno, pues podríamos caer en la misma extralimitación de poder que hemos atribuido al computacionalismo. A pesar de la revalorización del concepto de emergencia presente en los estudios sobre sistemas complejos, es un concepto que permanece en cierta medida en estado especulativo y el intento de transformarlo en un programa de investigación científico es algo que está teniendo lugar de manera muy reciente[25]. Aquí apreciamos que la IA puede jugar un papel especial, ya que, en consonancia con lo que venimos defendiendo, consideramos que la IA es una *práctica* (es decir, una efectiva construcción empírica que puede resultar exitosa o no) que está, dada la naturaleza de sus aspiraciones, intrínsecamente conectada con la especulación metafísica. Lo cual no quiere decir, en ningún caso, que los constructores de IA tengan que compartir estas preocupaciones metafísicas (tal como declara Brooks) o, al contrario, que el éxito de un determinado programa de IA pueda ser tomado como validación empírica de cierta perspectiva metafísica (nos remitimos, aquí, a los matices introducidos al comienzo del punto 2). Ninguna especulación metafísica puede ser validada empíricamente; al contrario, son los programas de investigación empíricos los que asumen ciertos presupuestos metafísicos que no pueden ser demostrados.

22 Véase Andy Clark y David J. Chalmers, "The Extended Mind", en *Analysis* 58 (1998) 10-23; para un desarrollo más extenso de estas ideas, véase Andy Clark, *Being There: Putting brain, body, and world together again*.

23 Debemos matizar que las propiedades emergentes se pueden dar tanto en un sistema homogéneo, sometido a los efectos de un parámetro externo, como en un sistema heterogéneo, en virtud de la interacción de sus componentes. Aquí se trata, evidentemente, de un ejemplo del segundo tipo. Las formas más fuertes de emergencia se dan cuando las propiedades emergentes son provocadas por relaciones no lineales entre los elementos del sistema.

24 Véase *Cosmopolitiques VI: La vie et l'artifice: visages de l'émergence*, en Isabelle Stengers, *Cosmopolitiques, 2 vols*., Paris, La Découverte, 2003.

25 Para un desarrollo de esta tesis, véase Claus Emmeche et al., "Explaining emergence: towards an ontology of levels", en *Journal for General Philosophy of Science* 28 (1997) 83-119.

La IA ocupa un lugar especial porque ofrece un terreno de experimentación en el que se pone a prueba el alcance del poder de los conceptos, realizando un movimiento en el que se abandona el terreno especulativo y se aborda la tarea de construir artefactos que pretenden realizar actividades propias de un comportamiento inteligente. En este sentido, podemos observar la importancia que adquiere el concepto de emergencia en las ideas enactivas. El enactivismo radicaliza la idea, propia de los intentos evolucionarios y dinámicos de la *embodied* IA, de que el diseñador debe intervenir lo menos posible en el artefacto. Considera que no es suficiente con dotar al agente cognitivo de un cuerpo que interactúe con el entorno y de unas herramientas que le permitan aprender y evolucionar para que podamos hablar de agentes con una perspectiva significativa. La relación de asignar relevancias no debe ser implementada en el agente artificial, sino que debe surgir espontáneamente de éste y, para ello, debe estar concernido con su existencia, es decir, debe tener un sí-mismo que actúe con una finalidad propia.

Esto nos lleva a una idea central de la corriente *enactiva*: la autonomía constitutiva. Siguiendo la idea kantiana de propósito natural, se considera que los seres vivos poseen una teleología inmanente en virtud de la cual se autoproducen o se auto-constituyen. Los seres vivos, desarrollados en un ambiente de condiciones precarias que amenazan su existencia, se dotan a sí mismos de una finalidad. La autonomía constitutiva, esto es, la capacidad por parte del agente de auto-constituir sus estructuras dotándolo de una identidad (diferenciándose así del entorno y, por tanto, no desintegrándose[26]), es lo que permite el mantenimiento de una finalidad *propia*.

En lugar de dotar al agente de un sistema de valores para dar significado al mundo, el enfoque enactivo trata de crear las condiciones de emergencia para que el agente se auto-constituya, dotándose a sí mismo de dicho sistema de valores. Las diferencias con la IA computacionalista son notables. No se crea un mundo, en forma de representación simbólica, para implementarlo en el agente por medio de un programa informático. Al contrario, se considera que es el agente quien debe enactuar su mundo, esto es, crear una red de significaciones en base a lo que es relevante para su propia existencia. Para poseer un mundo, hay que crearlo; y únicamente creamos si estamos concernidos con algo. No deberíamos pensar, sin embargo, que esta creación de mundo es la actividad unilateral de un sujeto escindido de su entorno; antes bien, es algo que emerge fruto de la interacción entre un agente cuyo interés propio es el auto-sostenimiento y un entorno que a la vez posibilita y pone límites a su existencia.

[26] Esta idea, en plena consonancia con la teoría de la autopoiesis, puede parecer confusa pues, por un lado, se afirma la indisociable unión del organismo con su entorno y, por otro, se expone la auto-constitución del organismo como una necesaria diferenciación respecto al mismo, la cual le confiere su identidad, evitando así su desintegración. Podría parecer que caemos en esa división metafísica entre el yo y el mundo que pretendíamos evitar. Sin embargo, debemos tener en cuenta que precisamente es la unión del organismo con su entorno la que lo obliga a auto-constituirse y, por tanto, a diferenciarse de él.