PENSAMIENTO, CREATIVIDAD Y MÁQUINAS

Carlos Blanco Pérez
Universidad Pontificia Comillas, Madrid

Resumen: ¿Puede una máquina pensar creativamente? Este artículo explora el significado de los términos "pensamiento" y "creatividad" para concluir que, desde una acepción lo más parsimoniosa posible de estos conceptos, no existen razones *de iure* contra la posibilidad de que una hipotética inteligencia artificial fuerte logre pensar de manera creativa.
Palabras clave: Pensamiento, conciencia, creatividad, inteligencia artificial.
Thinking, Creativity, and Machines
Abstract: Can machines think creatively? This paper explores the meaning of "thinking" and "creativity" in their most parsimonious forms. The main conclusion points to the absence of fundamental reasons against the possibility that a hypothetical strong artificial intelligence may think in creative ways.
Keywords: Thinking, rationality, consciousness, artificial intelligence.
Recibido: 02/01/2019 **Aprobado**: 20/02/2019

1. El pensamiento ¿Qué significa "pensar"?

Antes de discutir la posibilidad de que una máquina piense resulta imprescindible consensuar una definición de pensamiento aceptable desde un prisma lógico y empírico. Pues, en efecto, sólo si esclarecemos el concepto más general y parsimonioso de pensamiento podremos adentrarnos en el debate actual sobre el pensamiento de las máquinas.

Lo cierto es que los importantes desarrollos producidos en la investigación sobre inteligencia artificial exigen fomentar un diálogo profundo entre la filosofía, la ciencia y la tecnología. De hecho, creo que entender la naturaleza y los límites del pensamiento humano representa la gran frontera de la razón, pues inevitablemente nos plantea las siguientes preguntas: a partir de circuitos electrónicos, ¿podremos crear una conciencia, una mente capaz de abstraer hasta llegar a darse cuenta de su propia existencia? ¿Hasta qué punto es el pensamiento abstracto una propiedad exclusiva de la especie

humana? ¿Qué hay más allá de nosotros mismos, o acaso somos la estación final de la evolución en sus dimensiones biológicas y culturales?

En definitiva, lo que está en juego es el significado de "pensar" y, más aún, el significado de lo humano.

Pocas nociones nos parecen tan intuitivamente válidas como la de "pensar". Todos creemos saber qué significa pensar. Sin embargo, reconocer la ignorancia e intentar subsanarla es el camino preeminente hacia la más genuina reflexión filosófica. Y basta con reparar en el significado de "pensar" para advertir fácilmente la insuficiencia de muchos de nuestros conceptos.

La tesis principal que trataré de defender en este artículo es sencilla: si renunciamos al dualismo ontológico (no necesariamente al epistemológico, ni al de propiedades, sino a la hipótesis ontológica de que existen dos clases de realidad: la materia y el espíritu), no veo argumentos *de iure* contra la posibilidad de que las máquinas ejecuten funciones cognitivas de orden superior, tradicionalmente atribuidas en exclusiva a los humanos. Mantendré, eso sí, que considero sumamente complicado que lo logren *de facto*.

Ya nuestro insigne ingeniero e inventor Leonardo Torres Quevedo presagió el inmenso poder de la inteligencia artificial cuando escribió:

> Intentaré demostrar en esta nota —desde un punto de vista puramente teórico— que siempre es posible construir un autómata cuyos actos, todos, dependan de ciertas circunstancias más o menos numerosas, obedeciendo a reglas que se pueden imponer arbitrariamente en el momento de la construcción. Evidentemente, estas reglas deberán ser tales que basten para determinar en cualquier momento, sin ninguna incertidumbre, la conducta del autómata. No hay entre los dos casos la diferencia que veía Descartes. Pensó sin duda que el autómata, para responder razonablemente, tendría necesidad de hacer él mismo un razonamiento, mientras que en este caso, como en todos los otros, sería su constructor quien pensara por él de antemano. Creo haber mostrado, con todo lo que precede, que se puede concebir fácilmente para un autómata la posibilidad teórica de determinar su acción en un momento dado, pesando todas las circunstancias que debe tomar en consideración para realizar el trabajo que se le ha encomendado (Torres Quevedo, 2003: 11-13).

No cabe duda de que Torres Quevedo se mostraba profundamente optimista sobre la posibilidad de construir una máquina pensante. En la estela de muchos de los grandes pioneros de la investigación en ciencias computacionales, creía que un concepto tan insondable y filosóficamente amplio como el de pensamiento era plenamente susceptible de un esclarecimiento científico, capaz de romper el hechizo de impenetrabilidad aparente y misticismo irreductible obrado por siglos de planteamientos dualistas (tácitos o explícitos). Así, las propiedades más genuinas de la mente serían reproducibles mecánicamente, lo que propiciaría el surgimiento de una verdadera inteligencia artificial.

Parece por tanto perentorio clarificar filosóficamente las notas básicas del acto de pensamiento en cuanto tal, antes de abordar el interrogante sobre la viabilidad de un pensamiento artificial.

Sin embargo, es también necesario tener en cuenta que un exceso de rigor analítico puede llevarnos a excluir o desdeñar determinados aspectos del pensamiento, por haber podado en demasía las ramas del árbol y habernos afanado sobremanera en descomponer sin luego recomponer el rompecabezas.

Aunque en el libro *La integración del conocimiento* (Blanco Pérez, 2018: 23ss.) he discutido las posibles notas esenciales del pensamiento en su sentido más abstracto y universal, en aras de la completitud conceptual considero oportuno reiterar aquí las tesis más importantes de ese escrito, para más tarde atacar el problema del pensamiento de las máquinas.

De manera laxa, el pensamiento puede definirse como una asociación de contenidos mentales. Esta característica parece constituir, de hecho, el elemento fundamental de todo tipo de pensamiento, en cualquier especie biológica capaz de ser sujeto de pensamiento. Ciertamente, sería preciso afinar más esta acepción. Habría que indicar, por ejemplo, qué es un contenido mental y cómo se asocian exactamente. En cualquier caso, resulta incontestable que toda forma de análisis y selección de opciones exige una asociación previa de contenidos mentales. Pensar implica entonces ponderar esos contenidos, supervisarlos, discriminarlos de acuerdo con algún

criterio. Esta relación de ideas, por vaga que se nos antoje, ha de expresarse en un lenguaje, contemplado como un sistema de signos útil para quien sabe usarlo. Por ello, en el acto de pensar no hacemos sino establecer un conjunto de relaciones entre contenidos mentales a través de constantes lógicas articuladas en un lenguaje específico (que puede interpretarse como una representación interna inteligible para el agente cognitivo que la elabora y emplea). En esta manipulación de contenidos mentales el agente se vale de expresiones simbólicas referidas a objetos reales o posibles.

De este modo, puede decirse que pensar implica diseñar un sistema de monitorización, encargado de supervisar el contenido sobre el que versa ese pensamiento concreto. Al pensar, el sujeto se apodera de los elementos que constituyen dicho contenido: asimila una secuencia lógica mediante la elaboración de una función que comprende sus elementos. El diseño de esta función puede interpretarse como un proceso de *categorización*, en el cual se selecciona una de las posibles configuraciones de la relación entre los elementos de esa asociación mental, para así seguir un itinerario inferencial específico.

Desde esta perspectiva, pensar consistirá primordialmente en la capacidad de aprehender y asociar contenidos mentales a través de un sistema formal en el que sea posible articularlos (un "lenguaje"). Así, el acto de pensar comportaría la creación y selección de correlaciones entre un contenido mental y un símbolo. Dicha relación entre contenidos mentales mediante constantes lógicas permite anticipar un resultado (una consecuencia, una inferencia lógica que puede ser correcta, incorrecta o indeterminada). Al pensar es entonces posible elaborar un marco "meta", un espacio virtual de correlaciones más amplio que los propios contenidos mentales que lo integran: un sistema de ideas. Este sistema puede concebirse como esencialmente equivalente al diseño de una función que abarca contenidos mentales ("objetos"), conectados en virtud de una serie de factores lógicos. Por supuesto, estas correlaciones pueden ser unívocas (si cabe diseñar una correspondencia biyectiva entre cada contenido mental y un

referente particular —real o posible—) o plurales, si diferentes objetos pueden corresponder a cada contenido mental, o si diferentes contenidos mentales pueden corresponder a cada objeto. Un pensamiento más riguroso y significativo tenderá a establecer relaciones biyectivas entre los contenidos mentales y los objetos, mientras que uno polisémico ganará en extensión y elasticidad, pero a costa de perder "intensidad conceptual" y exactitud. Por su parte, un pensamiento más profundo logrará discernir conexiones más fundamentales entre los contenidos mentales que están siendo manipulados.

Si el proceso de pensamiento aspira a ser estrictamente racional, habrá de satisfacer ciertas reglas de consistencia y deberá ser consciente de sus premisas iniciales, a fin de trazar un camino lógico que conduzca desde ellas hasta unas conclusiones. La racionalidad (que es, al fin y al cabo, una de las clases posibles de pensamiento) la interpreto así como la capacidad de organizar la información sobre la base de unas premisas y unas reglas de inferencia. El pensamiento más racional será aquél que con un menor número de premisas consiga integrar una mayor cantidad de información. Por tanto, la racionalidad implica eficiencia cognitiva a la hora de fundamentar un desarrollo lógico (aquí, la lógica se refiere al proceso de inferencia en cuanto tal, a la forma del pensamiento).

De forma sintética, el pensamiento se nos presenta entonces como un vínculo entre contenidos mentales a través de constantes lógicas y gramaticales. Fieles al esquema que acabamos de enunciar, semejante hecho puede también interpretarse como el diseño de una función con un dominio de aplicación: el de los objetos sobre los que versa ese pensamiento. Sin embargo, este análisis de las características fundamentales del pensamiento resultaría incompleto sin distinguir entre la perspectiva de las reglas y la de las intuiciones.

En su faceta más algorítmica o reglada, el pensamiento racional se estructura mediante reglas que garantizan la posibilidad de alcanzar conclusiones consistentes, y el sujeto pensante ha de mostrar competencia

en el uso de esas reglas. No obstante, este pensamiento debe estar supervisado por un sujeto que asimile los contenidos y se muestre capaz de aprehender un significado, un sentido unitario del conjunto de los contenidos mentales que componen ese pensamiento específico. Esta asimilación del objeto como tal (ya sea un concepto, un principio de la razón o la integración de ambos en el seno de una proposición), evaluada en su dimensión unitaria y no sólo en la de los elementos individuales que constituyen el objeto, parece señalar una faceta genuinamente "intuitiva" de la mente, donde la descomposición analítica de las partes que entran en juego en el contenido del pensamiento cede el testigo a la elaboración de una síntesis unitaria. No basta, por tanto, el seguimiento ciego de unas reglas sintácticas: es necesario hacerse cargo de ellas, aun precariamente.

Estas consideraciones no implican, sin embargo, aceptar una especie de primacía unilateral y despótica de lo intuitivo sobre lo racional, como si en el pensamiento comparecieran elementos inasequibles a una comprensión lógica y científica. De hecho, los contenidos de nuestras intuiciones han de subsumirse, de una u otra forma, en los mecanismos generales del pensamiento racional, en las reglas que guían nuestros procesos intelectuales. No sería exagerado sostener que las intuiciones obedecen, en efecto, a reglas internas, por lo que el sueño de conquistar una comprensión más netamente explicativa del pensamiento humano no tendría por qué alzarse como una meta vedada a nuestros esfuerzos científicos más sólidos.

En suma, en todo pensamiento es posible identificar al menos dos clases de objetos: contenidos ("imágenes mentales") y relaciones lógicas (unas reglas articuladas en torno a constantes lógicas que permiten establecer relaciones entre esos contenidos). Al pensar aplicamos un conjunto de categorías lógicas a nuestras representaciones de la realidad externa e interna a nuestra mente.

Se trata, ciertamente, de una división conceptual, que no tiene por qué implicar una separación temporal auténtica entre ambos procesos, el de elaboración de representaciones mentales y el de asociación de sus contenidos.

2. La creatividad

Al pensar, al relacionar contenidos mentales, la mente humana no se limita a asimilar información, sino que también la organiza activamente (creo que es aquí donde estriba la aportación más perdurable de Kant a la teoría del conocimiento). Contemplamos, así, cierto grado de creatividad en este proceso de gestación de nuevas ideas y de nuevos marcos de referencia. Nunca asimilamos la realidad a escala 1:1. "El mapa no es el territorio", pues existe siempre una asimetría entre la realidad y su representación, que tiene que ver con la organización creativa de la información mediante conceptos y sistemas conceptuales. Nuestras mentes elaboran representaciones de la realidad que condensa en grupos de símbolos, y en este proceso de asimilación del mundo externo es inevitable que incorporemos esquemas internos cuyas estructuras introducen un cierto grado de subjetividad, de "originalidad", de tratamiento y filtrado de la información.

De hecho, en la somera descripción del acto de pensar que he planteado en los párrafos anteriores me he visto obligado a invocar elementos no puramente algorítmicos (es decir, no subsumibles en un programa de instrucciones bien definido), sino intuitivos y, en cierta manera, creativos. Sin embargo, no deberían considerarse creativos como si en ellos se hubiera propiciado el surgimiento de realidades *ex nihilo*, el nacimiento de un *novum* auténtico, sino en relación a los contenidos presentes; creativos por cuanto existe una ruptura, un salto que, a mi juicio, se recapitula muy bien en la noción de "analogía".

Resulta entonces pertinente plantearse la pregunta por la naturaleza de la creatividad.

No pretendo ahora resumir los importantes trabajos de investigación neurocientífica que se han llevado a cabo en los últimos años, sino esbozar

algunas sugerencias filosóficas que pueden ayudarnos a esclarecer este interrogante tan hondo e ineludible, dado que afecta a una de las dimensiones más deslumbrantes de la mente humana: la capacidad de gestar lo nuevo[1].

Entre una visión mística e idealizadora de la creatividad y una perspectiva netamente racionalista cabe, a mi juicio, una posición intermedia, no ecléctica, sino integradora. Desentrañar los procesos neurobiológicos que subyacen a la génesis de una idea nueva es perfectamente compatible con apreciar el valor filosófico del poder que ostenta la mente humana para crear, es decir, para abrir nuevos horizontes de reflexión, expandir nuestros marcos conceptuales e imaginar conexiones imprevistas entre los fenómenos del mundo y del pensamiento. De hecho, estoy convencido de que sólo una comprensión más profunda de los mecanismos cerebrales precisos implicados en esta habilidad tan extraordinaria nos permitirá descubrir cómo se entrelazan los procesos neurobiológicos y los contextos históricos y culturales que moldean las distintas manifestaciones de la creatividad humana. También nos ayudará a interpretar adecuadamente sus expresiones en otras especies, pues aunque parece innegable que la expansión de las cortezas prefrontales ha incrementado inconmensurablemente las capacidades creativas del *Homo sapiens*, es posible discernir signos incuestionables de creatividad en numerosos animales, cuya comprensión exige trascender las rígidas explicaciones en términos de reacción a estímulos y de adaptación frente a presiones ambientales para reconocer genuinas habilidades creativas.

Ciertamente, es en esta convergencia insoslayable de causalidades "de abajo arriba" y "de arriba abajo" donde se pone de relieve la insuficiencia tanto de un entendimiento puramente cerebral y neurofisiológico de la creatividad como de un estudio meramente social, basado en factores extrínsecos a las estructuras y funciones de la propia mente humana. Más aún, las investigaciones sobre los mecanismos que regulan la plasticidad del cerebro humano parecen llamadas a proporcionar un vínculo cada vez más

[1] He abordado este problema en otros lugares (cf. Blanco Pérez, 2014; Blanco Pérez, 2018c). En el presente texto recapitulo las ideas principales del libro y añado algunas nuevas.

robusto entre los procesos *bottom-up* y *top-down*. Estos trabajos no hacen sino mostrar que no estamos determinados a pensar de una manera concreta e irrevocable por poseer tal o cual dotación de conexiones sinápticas, sino que junto a un programa relativamente rígido de instrucciones genéticas es posible modificar las conexiones neuronales en interacción con el ambiente, con lo externo, con el entorno social y cultural en el que navegamos (Merzenich *et alii*, 1988). De esta feliz indeterminación de la arquitectura de nuestras conexiones neuronales dimana un notable grado de flexibilidad organizativa, una riqueza configurativa sin la cual resultaría imposible explicar cómo la mente aprende y eventualmente crea lo nuevo.

Desde un punto de vista fenomenológico, en todo acto creativo comparecen elementos de continuidad y de divergencia. Ninguna creación rompe radicalmente con lo anterior, pues siempre es posible identificar nexos con elementos lógicos y materiales precedentes. Crear implica, en cierto modo, "recordar", o al menos buscar nuevas conexiones entre objetos con los que seguramente ya estábamos familiarizados (Benedek *et alii*, 2014). A pesar de ello, es indudable que se produce también una divergencia con respecto a lo anterior, una innovación genuina que, aun sin escindirse súbitamente de un itinerario lógico, no se limita a evocar ideas precursoras, sino que aporta una nueva configuración de los objetos mentales.

En esta especie de aleatoriedad estructurada, la mente creativa construye y reconstruye, innova y reordena, dando saltos en el vasto espacio de posibilidades de la imaginación. Lo hace, sin embargo, con elementos materiales preexistentes. Incluso las creaciones más pintorescas de la imaginación suelen partir (salvo en escasísimas excepciones) de objetos ya conocidos, de experiencias ya acumuladas, de reflexiones ya esbozadas. Pues, en efecto, junto a rutas lineales, itinerarios lógicos secuenciales perfectamente susceptibles de elucidación que permiten transitar de un antecedente a un consecuente, en el proceso creativo es inevitable distinguir una cierta "ruptura de la simetría lógica", una serie de saltos conceptuales y figurativos que remiten a la noción de "analogía" como relación no necesaria, aunque legítima, entre lo distinto mediante la

identificación de mínimos comunes denominadores. Es precisamente en la elaboración de homologías, o semejanzas analógicas entre distintos elementos, donde reside una de las herramientas más fecundas de la creatividad. Estas homologías podrán ser directas, si la conexión entre antecedentes y consecuentes salta a la vista, o indirectas, si el vínculo es más remoto.

En cualquier caso, el problema lógico y filosófico más profundo se refiere no tanto a la habilidad de encontrar combinaciones inusitadas de ciertos elementos, que en muchos casos puede obedecer a la mera fuerza bruta de cómputo (o, más bien, a una estrategia de expansión/contracción o de excitación/inhibición, cuya dualidad, en términos biológicos, guardaría una interesante analogía con el mecanismo de variación/selección que impulsa la evolución de las especies), sino a la posibilidad de que tenga lugar un hiato genuino, la eclosión de un auténtico *novum* en el seno de la mente humana. En otras palabras: más allá de la concurrencia de factores racionales, emocionales, espontáneos o deliberados, ¿cuál es la causa neurobiológica de la disrupción creadora? ¿Existe un hiato absoluto entre el momento infinitesimal previo a alumbrar una idea y el acto de alumbrarla, un verdadero "cisma" en el colosal entrelazamiento de causas neurobiológicas y ambientales que allí confluyen? Me cuesta pensar que no haya continuidad microscópica a escala cognitiva e incluso neurobiológica; el proceso creativo lo interpretaría más bien como una reorganización de contenidos mentales comparable a las transiciones de fase que estudia la física, donde no tiene por qué acontecer una ruptura completa y tajante entre antecedentes y consecuentes, sino una reconfiguración de los elementos presentes. Así, si "*Natura non facit saltus*", cabría decir que "*Intellectus non facit saltus*".

Es, no obstante, en las homologías indirectas, en la proposición de conexiones insospechadas o inverosímiles, donde podemos admirar una de las fuentes más bellas e inescrutables de las grandes creaciones intelectuales y artísticas de la humanidad. Por qué tuvo Newton la idea de que existía una conexión razonable entre la fuerza que mantiene unidas la Tierra y la Luna y la que hace caer la manzana del árbol es uno de los misterios más hermosos

de la mente humana. ¿Se trata de algo inexplicable, de un milagro sobrenatural? No lo creo, pues por improbable que se les antojase este vínculo entre lo supralunar y lo sublunar a los contemporáneos de Newton la legitimidad de buscar un enlace conceptual entre ambos mundos era irreprochable, aunque contradijese siglos de reflexión filosófica, científica y teológica. Era, de hecho, esperable que tarde o temprano se le ocurriera a alguno de sus contemporáneos, al igual que Leibniz descubrió el cálculo infinitesimal de manera prácticamente simultánea a Newton, o Wallace la selección natural al mismo tiempo que Darwin, o Poincaré muchos de los principios de la relatividad especial con independencia de Einstein. No podemos olvidar, sin embargo, que en el análisis de esta cuestión parece inexorable apelar a las capacidades intelectuales únicas de las grandes mentes de la historia, hijas de su tiempo, ciertamente, pero poseedoras de unas habilidades cognitivas sobresalientes, que les permitieron vislumbrar y justificar relaciones más profundas y trascendentales entre fenómenos aparentemente inconexos.

En la analogía podemos entonces discernir un poderoso principio heurístico, un instrumento facilitador de la génesis de ideas nuevas, de la creatividad en su acepción más límpida y genuina. La analogía carece del valor demostrativo de la inferencia lógica, pero exhibe un inmenso potencial inspirador para explorar nuevas conexiones entre los fenómenos y nuevas aproximaciones a los problemas vigentes. Se alza así como una estructura general de la imaginación que nos permite encontrar homologías entre objetos y conceptos cuyas propiedades guardan algún tipo de relación, por débil y remota que parezca. Amparada en la comprensión de las similitudes existentes entre la constitución y las propiedades de ciertos objetos y categorías, nos capacita para aventurarnos por nuevos escenarios y trascender los rígidos requisitos de la estricta inferencia lógica. Resplandece, por tanto, como un proceso eminentemente constructivo, que no se limita a desplegar el contenido de las premisas, sino que se atreve a proponer elementos innovadores más allá de los resultados esperables de una conexión lógica.

Además, en algunos casos esa creación se canaliza a través de la destrucción de formas previas y de su sustitución por otras.

De modo casi inexorable, en toda analogía osada y potencialmente rompedora brilla la luz de la intuición, de una percepción subjetiva intensa y difícilmente transferible que se halla imbuida de una fructífera incertidumbre; ajena a los cánones de una demostración racional rigurosa e inequívocamente válida, pero dotada de una desbordante fuerza creativa. ¿Cómo y por qué se producen estas intuiciones? ¿Por qué sólo bendicen a algunas mentes? ¿Son la consecuencia lógica del entrenamiento, de la familiaridad con un dominio concreto del saber y de la acción? ¿Un ordenador soportaría esa incertidumbre lógica que envuelve la intuición humana, ventana a la creatividad?

De hecho, la creatividad sólo parece despuntar con todo su fulgor allí donde hay posibilidad de ambigüedad, indefinición e incompletitud; allí donde cabe cuestionarse fundamentos, desarrollos y consecuencias. Un razonamiento lógico bien formulado es apodíctico, necesario. En él no hay atisbos de libertad creadora. Sin embargo, para crear es preciso franquear las barreras de la necesidad lógica particular y ampliar el horizonte de reflexión, a fin de identificar nuevos elementos de juicio y nuevas conexiones. Los grandes pensamientos asumen la paradoja y la resuelven en una nueva síntesis, en un nuevo marco, en una especie de *Aufhebung* recapituladora. Lo contemplamos en Einstein, quien reconcilió la mecánica newtoniana con el electromagnetismo de Maxwell al percatarse de que eran compatibles dentro de un modelo físico más amplio y profundo.

3. Las máquinas

¿Podrá entones pensar una máquina de forma consciente, es decir, relacionar contenidos y referirlos a ella misma, a la percepción de su propia identidad? ¿Podrá ordenar pensamientos de manera razonada? ¿Aprenderá a tolerar incertidumbres y ambigüedades, antesalas de la creatividad?

Una respuesta fácil argüirá que todo depende de cómo definamos el concepto de "pensamiento consciente" (si es que este binomio no es tautológico). No obstante, parece claro que si establecemos una idea demasiado exigente de conciencia, ¿cómo podemos saber que nosotros, los seres humanos, la poseemos en realidad, en vez de limitarnos a ejecutar actos puramente mecánicos y automatizados, frutos irreflexivos de un proceso algorítmico ajeno a un hipotético e inescrutable *yo*? ¿Cómo saber que la conciencia es propiedad de alguien además del sujeto consciente que la percibe de una manera tan vívida, tan íntima y casi incomunicable, como una característica concomitante a sus propios actos cognitivos?

Por ello, es imprescindible consensuar una definición provisional de conciencia, que deberá conjugarse con la de pensamiento para dilucidar mínimamente la idea de "pensamiento consciente", es decir, de aquel tipo de pensamiento del que el sujeto puede hacerse cargo. Pues, en efecto, de la discusión sobre el concepto de pensamiento que he esbozado en la primera sección es fácil colegir que, a mi juicio, en su acepción más profunda éste se revela como un fenómeno antitético a lo inconsciente, o a una mera concatenación de estímulos y respuestas sobre la que el sujeto pensante carecería de control alguno. Desde esta perspectiva filosófica, el pensamiento implica necesariamente un sujeto que supervise el proceso de asociación de ideas, para apropiarse de él. Esta posibilidad resultaría inviable sin disponer de cierto grado de conciencia del acto pensante en cuanto tal, algo que no tiene por qué circunscribirse a la mente del *Homo sapiens*, sino que probablemente abarque a otros vertebrados dotados de sistemas cognitivos de orden superior.

Así, lo que en primera aproximación consistía en una mera asociación de ideas se muestra, tras un análisis más profundo, como una aprehensión de esas conexiones entre contenidos mentales, acto que implica una monitorización subjetiva más allá de un encadenamiento de imágenes. Aun vagamente, esta noción apunta al problema de la conciencia y de cómo ésta se relaciona con el pensamiento.

Reconozco que late aquí la profunda e ineludible dificultad de la autoconciencia, de la apercepción trascendental kantiana, de la acción que *a priori* trasciende todos sus contenidos posibles: de la autorreflexión. Pese a la gravedad del interrogante, la cuestión es si dicha instancia subjetiva puede concebirse en continuidad orgánica con las estructuras cerebrales más básicas y con las formas biológicas menos complejas, en cuyo caso la hipótesis dualista, que postula una escisión ontológica entre el sustrato material y la subjetividad mental, perdería vigencia. No sería entonces ilusorio afanarse en construir una conciencia artificial en máquinas pensantes, dado que conoceríamos de modo razonable la estructura y las propiedades de un fenómeno tan esquivo. Parece claro, eso sí, que si la conciencia ha surgido de manera evolutiva (tal y como nos enseña la aplicación —inobjetable— de la teoría de la evolución al desarrollo de las facultades mentales del *Homo sapiens*), es preciso que exista un mecanismo biológico subyacente, una explicación científica de cómo han aparecido filogenéticamente nuestras habilidades cognitivas. Por supuesto, esta pregunta podría reproducirse también a escala ontogenética.

Como nos enfrentamos a uno de los misterios más hondos y fascinantes de la investigación filosófica y neurocientífica, expondré únicamente algunas ideas básicas que pueden iluminar aspectos fundamentales de esta cuestión, tan apremiante como inveterada.

Los avances en la comprensión de los mecanismos neurobiológicos de la conciencia han sido notables. Aunque importantes aspectos sigan hoy cubiertos de un denso velo de misterio, no encuentro razones para creer que esos enigmas hayan de perdurar por siempre: "*ignoramus, sed non ignorabimus*". De hecho, al menos desde Baars (cf. Baars, 1988; Baars, 2005) y, más recientemente, Dehaene y Changeux (cf. Dehaene et alii, 2017; Dehaene et alii, 2006), se han propuesto modelos teóricos susceptibles de contraste experimental. Destinadas a validar o refutar hipótesis sobre la naturaleza de la conciencia, muchas de estas pruebas experimentales versan, por ejemplo, sobre la relación entre lo inconsciente y lo consciente en la percepción (rivalidad binocular, estímulos enmascarados, competición entre

imágenes…). Sabemos también que la percepción consciente involucra elementos inconscientes, que responden, *grosso modo*, a un procesamiento en paralelo, mientras que los aspectos estrictamente conscientes se canalizarían a través de procesos secuenciales[2].

De manera simplificada, la conciencia puede contemplarse como un sistema de monitorización de la información procesada por el cerebro. No es de extrañar, de hecho, que la evolución haya propiciado el nacimiento de un sistema cognitivo semejante. Su objetivo no sería otro que el de filtrar y controlar una información tan heterogénea como la que constantemente bombardea nuestros sentidos.

Cabe argüir, no obstante, que esta concepción de la conciencia como sistema de monitorización, como elemento *extra* que añade una nueva dimensión al análisis de la información disponible (una instancia externa, o "metainstancia"), peca de suma ingenuidad, pues no justifica la verdadera necesidad de la conciencia. En efecto, podría alegarse que ese dispositivo de monitorización funcionaría perfectamente aunque careciera de conciencia y fuera un mero zombi. Sin embargo, me parece razonable pensar que la conciencia añade una ventaja evolutiva insoslayable, no incluida en la monitorización puramente mecánica e inconsciente de la información tratada por el sujeto. En mi opinión, la perspectiva de la primera persona incorpora un desdoblamiento entre el sujeto y el objeto que ofrece grandes y profundos beneficios a quien es capaz de aplicarla: confiere un mayor grado de independencia con respecto al entorno, e incluso con respecto a nuestros propios estados mentales, pues llega a "juzgarlos" desde fuera. ¿Cuál es su

[2] Esta interesante dualidad ha sido ampliamente estudiada en el marco del sistema visual. Los elementos inconscientes pueden representarse bastante bien mediante ciertos algoritmos e inferencias estadísticas, como las inferencias bayesianas. Lo cierto es que el grado de división de tareas (lo que Zeki y Schipp han descrito como la "lógica funcional de las conexiones corticales" en Zeki; Schipp, 1988) existente en el sistema visual es excepcional. Unos sistemas se encargan, por ejemplo, de procesar información relativa a la inclinación de la figura percibida, mientras que otros se han especializado en la detección del color o de la intensidad de la luz. El cerebro, como un todo, parece entonces responder a dos grandes principios de organización funcional: los de especialización e integración (cf. Friston, 2005).

mecanismo último? ¿Basta con invocar una dualidad entre percepción y asociación, posibilitada por la organización neuroanatómica del cerebro humano en distintas áreas, unas encargadas de percibir y otras de asociar la información percibida? Creo que sí, al menos desde un punto de vista conceptual. Con todo, resulta innegable que todavía falta mucho para esclarecer completamente los procesos neurobiológicos y cognitivos concretos, si es que un objetivo tan ambicioso podrá algún día llevarse a término, lo que culminaría una de las búsquedas intelectuales más apasionantes emprendidas por la especie humana: la que nos ha impulsado a preguntarnos por nuestro verdadero ser y por la génesis de nuestras habilidades más distintivas.

Lo cierto es que la mente es, ante todo, una fuerza de unificación, generadora de percepciones unitarias. Aun invadida por información heterogénea, es capaz de seleccionar unos elementos y de desechar otros. Esta operación constituye ya un atisbo de conciencia, interpretada como control o monitorización de la información disponible. La conciencia se presenta así como una especie de filtro, apto para atender selectivamente a unos contenidos y excluir otros. El nivel superior y más complejo de la conciencia implica, empero, referir esa información seleccionada a uno mismo, a una instancia subjetiva, a un elusivo *yo*. Es el acto de "saber que uno sabe", la posibilidad de que el sujeto sepa que sabe. Por supuesto, esta introspección representa uno de los aspectos más intrincados a la hora de abordar la naturaleza de la conciencia, pues su estudio depende en gran medida del grado de veracidad que atribuyamos al testimonio subjetivo sobre lo que se percibe (la denominada "*reportability*" en la literatura anglosajona)[3].

[3] Pese a esta dificultad, pueden diseñarse experimentos precisos que midan niveles de reportabilidad (Dehaene *et alii,* 2017). Además, es evidente que a veces somos conscientes y otras no, luego debe haber una función de continuidad entre ambos estados, aunque el cambio resulte prácticamente imperceptible. La conciencia, por tanto, no tendría por qué verse como un proceso de "todo o nada", sino que existirían escalas de conciencia. No obstante, es también probable que haya alguna clase de *umbral crítico*, noción que no presentaría mayores problemas para la ciencia, acostumbrada en multitud de ramas del conocimiento a la idea de puntos críticos y de rupturas de simetría.

¿Es plausible que un ordenador alcance esta enigmática autorreferencialidad, un nivel tan profundo de posesión de la información, una autoconciencia?

Lógicamente, si consideramos que las máquinas se limitan a obedecer programas de instrucciones previamente diseñados por humanos, a recibir *inputs* que, mediante unas reglas de inferencia, generan *outputs*, es difícil que les otorguemos la posibilidad de pensar creativamente como los humanos. Esas máquinas serían simples autómatas, rígidos ejecutores de tareas asignadas que procesan información, pero no la asimilarían, no la ponderarían, no la referirían a una instancia subjetiva, a una conciencia individual. Privadas de la capacidad de establecer una frontera nítida entre el estímulo (el *input* y el diseño del sistema, que vienen dados por la mente humana) y la respuesta (el comportamiento inexorable de una máquina sometida a reglas de diseño), parece inconcebible atribuir a una máquina la habilidad de reflexionar conscientemente, de separarse provisionalmente de la vasta cadena de causas y efectos para valorarla como un juez externo y aventurarse a proponer caminos alternativos, posibilidades inéditas.

Sin embargo, dicha concepción de las máquinas no tiene por qué agotar todos los modelos posibles de diseño de una inteligencia artificial. En la creación de máquinas más avanzadas, que no se limitan a aprender un programa de instrucciones, sino que aprenden ellas mismas a aprender y llegan a elaborar sus propias instrucciones, radica una de las innovaciones tecnológicas más fecundas de las últimas décadas[4].

Además, la ingeniería del aprendizaje computacional puede beneficiarse enormemente de nuestra comprensión de los fenómenos biológicos, del fructífero entrecruzamiento de variación y selección que define los procesos evolutivos, de la síntesis de programas elásticos de instrucciones y adaptabilidad flexible al medio. Torres Quevedo ya se percató de ello, pues "es necesario que los autómatas imiten a los seres vivos, ejecutando sus actos

[4] Para una panorámica sobre el denominado *machine learning*, Alpaydin, 2009; sobre el *unsupervised learning*, Hastie *et alii,* 2009.

con arreglo a las impresiones que reciban y adaptando su conducta a las circunstancias" (Torres Quevedo, 2003: 10). Semejante aptitud para aprender a aprender propicia un refuerzo positivo, un mayor grado de adaptabilidad a un entorno cambiante, una destreza más diáfana a la hora de cribar paulatinamente las inferencias iniciales y someterlas al luminoso contraste de los factores externos. Es precisamente aquí, en este relativo indeterminismo de unas entidades que se ven obligadas a adaptarse a entornos mutables, donde estriba una de las características más sobresalientes de los fenómenos biológicos y una de las virtudes explicativas más importantes de la teoría de Darwin.

Sería, no obstante, sumamente iluso pensar que reproducir las notas esenciales de los fenómenos biológicos a escala computacional es tarea sencilla. De hecho, en el estudio de los procesos biológicos parece inevitable apelar a la existencia de una unidad de diseño, de un carácter orgánico que se sobrepone a la diversidad de los miembros y confiere al todo una articulación, una consistencia y un alto grado de unificación entre las partes. El viviente (sobre todo aquellos que poseen un sistema nervioso más evolucionado) actúa como un todo, ejerce un control sobre sus partes y se propone objetivos. Cómo imitar esta extraordinaria teleonomía, esta intrigante capacidad de autoconfiguración que atesoran los seres vivos, representa uno de los desafíos más profundos para la inteligencia artificial, pues en ella reside el auténtico salto cualitativo entre el automatismo y la espontaneidad. No es ingenuo, eso sí, creer que las visiones excesivamente idealizadas de la subjetividad animal (la humana incluida), prestas a entronizarla en un arcano sitial metafísico desde donde reinaría como conjeturado límite del mundo, son susceptibles de disolverse mediante conceptos que, inspirados en la neurociencia, la biología evolutiva y el estudio de la historia natural de la vida en la Tierra, muestren cómo puede haber surgido gradualmente esa capacidad de autopercepción que se alza como uno de los mayores enigmas de la ciencia.

Por tanto, diseñar máquinas similares a los organismos vivos, capaces no sólo de procesar información, sino también de adquirir conciencia de su propio existir y de dotarse de fines, de autodeterminarse (e incluso de autonegarse y autodestruirse), no tiene por qué erigirse en una meta utópica. Pero para ello será necesario aprender más sobre la "inteligencia" (biológica) que sobre lo "artificial": comprender bien cuáles son los rasgos más distintivos de la inteligencia animal y entender sus raíces biológicas para imitar o superar fehacientemente esta habilidad tan notable, que corona su cúspide en la especie humana.

No veo, en definitiva, una imposibilidad intrínseca de reproducir los procesos computaciones simples y complejos que ocurren a nivel cerebral. cuestión distinta es la envergadura de un proyecto que implicaría condensar e incluso trascender tecnológicamente millones de años de evolución biológica, de combinaciones de variaciones genéticas, selección natural y aprendizaje transmitido mediante la cultura, donde el método del ensayo y el error ha tenido que desempeñar un papel fundamental. Sin embargo, esta dificultad no tiene por qué ser absoluta, pues aún hoy desconocemos los límites del ingenio, la imaginación y la creatividad del ser humano para solucionar problemas y, más aún, para inventar otros nuevos y expandir el radio de lo posible.

4. Conclusiones

¿Podrá una máquina pensar, esto es, relacionar contenidos mentales y referirlos a ella misma? Más aún, ¿podrá pensar racionalmente?

Sí, pero todavía no. No veo una prohibición *de iure* a la posibilidad de una inteligencia artificial fuerte, aunque sí creo necesario advertir las enormes dificultades *de facto* que se interpondrán en el camino.

Por fortuna, no podemos determinar *a priori* el horizonte de lo posible y sancionar el de lo imposible. No sabemos, en realidad, qué es imposible para la mente humana, más allá de inconsistencias lógicas insuperables (no puedo pensar lo contradictorio, por ejemplo).

Así, una suma de prudencia y confianza en la capacidad del ser humano para trascender límites que parecían infranqueables debería orientar nuestras investigaciones en estos campos.

Bibliografía empleada

E. Alpaydin, *Introduction to machine learning*, Cambridge, MIT press, 2009.

B.J. Baars, *A cognitive theory of consciousness*, Nueva York, Cambridge University Press, 1988.

J.B. Baars, "Global workspace theory of consciousness: toward a cognitive neuroscience of human experience", en: *Progress in brain research*, 1988 (150): 45-53.

M. Benedek; E. Jauk; A. Fink; K. Koschutnig; G. Reishofer; F. Ebner; A.C. Neubauer, "To create or to recall? Neural mechanisms underlying the generation of creative new ideas", en: *Neuroimage*, 2014 (88): 125-133.

C. Blanco, *Lógica, ciencia y creatividad*, Madrid, Dykinson, 2015.

C. Blanco, *La integración del conocimiento*, Madrid, Evohé, 2018.

C. Blanco-Pérez, C., "Competition, cooperation, and the mechanisms of mental activity", en: *Frontiers in psychology*, 2018 (9): 1352.

C. Blanco-Pérez, "The logic of creativity", en: *The Heythrop Journal*, 2018 (59): 1-19.

S. Dehaene; J.P. Changeux; L. Naccache; J. Sackur; C. Sergent, "Conscious, preconscious, and subliminal processing: a testable taxonomy", en: *Trends in cognitive sciences*, 2006 (10): 204-211.

S. Dehaene - H. Lau - S. Kouider, S. "What is consciousness, and could machines have it?", en: *Science*, 2017 (358), 486-492.

K. Friston, "A theory of cortical responses", en: Philosophical Transactions of the Royal Society of London B: Biological Sciences, 2018 (360): 815-836.

J.M. Fuster, The neuroscience of freedom and creativity: Our predictive brain, Nueva York, Cambridge University Press, 2013.

T. Hastie; R. Tibshirani; J. Friedman, "Unsupervised learning", en: *The elements of statistical learning*, 2009: 485-585.

M.M. Merzenich; G. Recanzone; W.M. Jenkins; T.T. Allard; R.J. Nudo, "Cortical representational plasticity", en: *Neurobiology of neocortex*, 1988: 41-67.

L. Torres Quevedo, "Ensayos sobre automática: su definición: extensión teórica de sus aplicaciones", en: *Limbo: boletín de estudios sobre Santayana*, 2003 (17): 11-13.

S. Zeki - S. Shipp, "The functional logic of cortical connections", en: *Nature*, 1988 (335): 311.

Carlos Blanco Pérez
cbperez@comillas.edu