Capítulo **16**

Hacia la conciencia en inteligencia artificial: Un viaje por la evolución hasta las redes neuronales artificiales

Mariano Rivera

https://doi.org/10.61728/AE24001168

Inteligencia y conciencia como necesidad evolutiva

En un sentido amplio podemos entender a la inteligencia como la capacidad de aprender, adaptarse, planear, resolver y crear. Esta ha sido una herramienta crucial para la supervivencia y el éxito evolutivo de varias especies. En el caso de los humanos, la inteligencia facilitó la supervivencia al permitir mejoras en las técnicas de caza y de recolección. También, fomentó el desarrollo de habilidades sociales complejas, como el lenguaje y la cooperación. Tempranamente, los Homo sapiens (nuestra especie) usaron su inteligencia para crear herramientas, desarrollar estrategias de caza y establecer sistemas sociales complejos, lo que les proporcionó ventajas significativas sobre otras especies. Tomasello (2009) proporciona una visión de cómo y por qué las habilidades sociales y la inteligencia evolucionaron en los humanos. Como lo sugiere Dunbar (1998), el desarrollo y crecimiento del cerebro en los homínidos fue el resultado de la necesidad de procesar, manejar y recordar relaciones sociales complejas. Lo qué a su vez impulsó el desarrollo de la capacidad intelectual. Una clara analogía a la correlación histórica entre el crecimiento de la capacidad de cómputo de dispositivos informáticos y la complejidad de los sistemas de procesamiento. Con el desarrollo de la inteligencia, surgió la necesidad de evaluar y reflexionar sobre los propios procesos cognitivos (Mithen, 1997). Tal autoevaluación permitió a un individuo comprender las razones detrás del éxito o fracaso de sus acciones. Este nivel de consciencia, donde un ser es consciente de sus procesos cognitivos, puede considerarse un estado primario de conciencia. La habilidad de introspección es fundamental para el aprendizaje y la adaptación.

Para entender mejor la anterior aseveración, analizaremos dos escenarios hipotéticos que se llevan a cabo en situaciones similares.

Figura 1. Manada de lobos cazando un tigre.

En el primero, una manada de lobos trata de cazar a un tigre dientes de sable. La escena se desarrolla con los lobos rodeando al tigre en medio de una zona pantanosa, lo cual reduce la movilidad del tigre. La manada se compone de unos siete individuos, cada uno gruñendo al tigre desde diferentes direcciones y alternándose para amenazar con ataques. El tigre, en medio del círculo, voltea hacia uno y otro lado sin lograr predecir de dónde vendrá el siguiente ataque. De repente, un lobo se acerca y muerde una pata del tigre, que gira tratando de atrapar al atacante. Antes de que logre dar un zarpazo, otro lobo ataca rápidamente su cuello; es un ataque rápido pero no decisivo. El tigre gira y logra, ahora sí, golpear al contrincante con un zarpazo. Esto lo tumba, pero no logra eliminarlo, otros lobos lo atacan desde diferentes flancos antes de que el tigre dé un golpe definitivo.

La escena se repite. En un momento el tigre logra romper el cerco y avanzar algunos metros. Desafortunadamente, la zona pantanosa y fangosa en la que se desarrolla la lucha limita la habilidad y rapidez del tigre para escapar definitivamente de sus agresores. Pronto es rodeado nuevamente y se reinician los ataques desde diferentes flancos, con éxitos parciales para ambos bandos. Unas veces los lobos logran morder y lesionar al tigre. Otras tantas, el tigre logra dar un zarpazo a sus enemigos. Después de un tiempo, la resistencia del tigre es minimizada y los cazadores lo derrotan. Sin embargo, los cazadores pagan un costo: dos lobos están gravemente heridos y no pueden ponerse en pie. Algunos lobos, mostrando empatía,

se acercan y lamen a los heridos; no hay nada más que hacer. Los sobrevivientes van sobre los restos del tigre y comienzan a comer. Los lobos han mostrado una inteligencia colectiva que les permitió derrotar a un enemigo más fuerte. Alternándose de manera precisa desde diferentes lados no han dando la menor oportunidad a la presa de responder o escapar; juntos son una máquina de cazar.

Figura 2. Grupo de humanos ancestrales cazando un bisonte.

Nota: Ilustración generada con asistecia de DALL-E 2.

Ahora analicemos una situación similar. En este segundo caso, los cazadores son unos homínidos, de la especie Homo sapiens para mayor precisión. Ahora, la presa es un tipo de bisonte. Los homínidos son menos ágiles que los lobos de nuestro primer ejemplo, pero cuentan con herramientas: varas largas cuyas puntas han sido afiladas y endurecidas al fuego y, hachas hechas de rocas afiladas para cortar y arrojar. Sin intención, emulan la conducta de los lobos: se acercan desde diferentes flancos contra el bisonte que, igualmente en una zona de charcas, se revuelve para tratar de contrarrestar los ataques. La escena se desarrolla de manera muy similar al caso anterior. Los cazadores atacan desde diferentes flancos, tratan de no ofrecer un blanco fijo, se alternan en sus ataques y, algunas veces el bisonte logra en sus embates lesionar a algún atacante. Como en el caso anterior, los cazadores logran su cometido y someten a su presa. Igualmente, hay

un costo para los cazadores: de un grupo de nueve, tres han sido abatidos. La cacería ha concluido.

Aunque ambos grupos de cazadores (lobos y humanos) han actuado de forma coordinada para someter a una presa más fuerte y ágil, hay diferencias notables entre ambos casos. En el caso de los humanos la coordinación fue más compleja: se intercambiaron miradas, señas y sonidos que guiaron el ataque; esto ha sido resultado de una conducta social más sofisticada producto de su mayor capacidad cerebral. De inicio, los sapiens han tenido que elaborar herramientas, planear y coordinar su ataque en forma más compleja dada su limitación de agilidad, de fuerza, y de armas naturales. Por ello, no es de sorprender imaginarnos que lo que ocurre después también sea diferente.

Los Sapiens sobrevivientes se acercan a sus colegas caídos. Se entristecen por su pérdida. Sacan sus restos del fango y los cubren con piedras para evitar que algún depredador, tal vez lobos que se escuchan aullar, vengan por ellos. Sienten un gran pesar el que bestias devoren los restos de sus compañeros. Para aliviar la sensación de pérdida y fomentar su esperanza de que esta no sea definitiva, los sapiens realizan ritos funerarios. Luego, en nuestra escena imaginada, el líder recoge el hacha de piedra de uno de los caídos. No comen la presa directamente en el lugar, sino que la destazan con sus herramientas de piedra y se llevan las partes a una cueva donde el grupo y sus familias viven temporalmente. Ahí la asan en el fuego y la comparten con los demás miembros del grupo, generalmente mujeres y niños.

Por la noche, el líder del grupo de cazadores observa el hacha de piedra que recogió. Si pudiéramos leer sus pensamientos estos serían, en nuestras palabras, algo como: "es una herramienta formidable, de buen pedernal, con un lado romo para asirse, con filo doble y cuyo diseño permite ser reafilada". Dicha hacha es el resultado de más de un millón y medio de años de desarrollo tecnológico; desde que el Homo habilis creó las primeras hachas de piedra. Eso es mucho tiempo, como referencia, el sapiens aparece hace unos 170 000 años; 10 000 años es el lapso de tiempo que separa el hoy de la invención de la agricultura.

Regresemos a nuestro cazador que empuña el hacha de piedra recuperada. Este reflexiona sobre la pérdida de los colaboradores. Asumimos

que, al igual que los Sapiens de la tribu actual Piraha del Amazonas que no tienen concepto de número sino de relaciones muchos y pocos (Gordon, 2004), el líder nota que hay más espacio junto a la fogata. Comprende que hay ahora una menor proporción de machos adultos en el grupo. Recuerda cómo se desenvolvió la cacería, y trata de identificar errores en sus decisiones y acciones. Por necesidad, las próximas veces tendrán que buscar presas de menor tamaño, que aportarán menos comida y por lo tanto necesitarán incrementar la frecuencia de las cacerías. Y serán los débiles (ancianos, enfermos y niños) los primeros en ser dejados atrás.

La gran diferencia entre los dos ejemplos que hemos imaginado es la actitud reflexiva de los cazadores sobre los acontecimientos. La escena que involucra a individuos sapiens, tal vez ocurrió en Tanzania hará unos 100 000 años; y el de los lobos en algún lugar de Europa en fechas similares. En ambos casos, en los cazadores sobrevivientes se generaron sentimientos negativos (tristeza) por la pérdida de miembros del grupo. Sin embargo, lo relevante es que en el caso que hemos imaginado del humano, se generó un proceso mental que revisó, evaluó y corrigió los procesos mentales que condujeron su proceder. Después de ese proceso, el sujeto en cuestión modificó su razonamiento para que la próxima vez, las decisiones que habrá de tomar se adapten a las circunstancias, minimicen las pérdidas, y garanticen el éxito de la cacería. A este proceso mental que se desarrolla cuando el individuo se da cuenta de las implicaciones de sus decisiones lo llamamos conciencia primaria (o primitiva). Es un proceso mental que evaluá otros procesos mentales. Baars (1997) proporciona una explicación accesible de cómo los procesos cognitivos pueden llevar a un estado de conciencia. Este nivel de conciencia se observa cuando los individuos no solo reaccionan al mundo que les rodea, sino que también comprenden y reflexionan sobre sus acciones y pensamientos. Esta capacidad de introspección es crucial para adaptarse a nuevos entornos y situaciones, permitiendo aprender de errores y éxitos pasados.

La habilidad de autoevaluar los propios procesos cognitivos es un paso fundamental hacia el desarrollo de una autoconciencia. La autoconciencia va más allá de la simple reflexión y revisión, involucra un reconocimiento de uno mismo como entidad individual. Damasio (1999) proporciona una perspectiva neurobiológica sobre la relación entre la conciencia y el sen-

tido del yo. Esta forma de conciencia implica entender que uno existe de manera independiente dentro de un contexto más amplio. En humanos, esto se manifiesta en la capacidad de reconocerse en un espejo (aunque también elefantes y orcas, entre otras especias actuales, se pueden reconocer en el espejo), ser consciente del paso del tiempo, y reflexionar sobre el propio pensamiento y emociones. Metzinger (2004) aborda la noción de autoconciencia y cómo se relaciona con la experiencia subjetiva del mundo. La subjetividad es un aspecto clave de la experiencia consciente, permite a los individuos tener experiencias emocionales y personales únicas. Las emociones desempeñan un papel crucial en la forma en que los humanos interactúan con el mundo. Las memorias influyen en la toma de decisiones, en la formación de aspiraciones y en la definición de objetivos a largo plazo. De acuerdo con Damasio (1995), la emoción y la razón no están separadas, sino que son críticas para la cognición racional. Esta dimensión emocional de la conciencia es fundamental para entender la complejidad del comportamiento humano. Los sentimientos y las emociones (derivados de la subjetividad) colorean nuestra percepción del mundo, influyen en nuestras decisiones y acciones, y moldean nuestra individualidad.

La autoconciencia se debió desarrollar entre los cazadores Sapiens. Pues hace 100 000 años, los Sapiens ya utilizaban pigmentos basados en ocre para decorar sus cuerpos con fines ornamentales (Henshilwood et al., 2011). La necesidad de identificación personal, de reconocerse como distinto, implica autoconciencia. Más recientemente, hace unos 40 a 45 mil años nuestros antepasados ya pensaban sobre sí mismos (contaban con autoconciencia). Los Sapiens ya colgaban en su cuello las garras de presas cazadas, adornaban su cuerpo con joyas compuestas por huesos, pedazos de madera, cuentas de piedra, plumas, conchas y caracoles, añadiendo a cada individuo elementos que los hacían únicos (Henshilwood et al., 2002). Todas ellas son actividades con un sentido estrictamente estético cuyo propósito va más allá de la supervivencia. Otro signo de la autoconciencia se manifiesta en el arte rupestre, –la pintura rupestre elaborada por sapiens más antigua (descubierta) se encuentra en la isla de Cáseres, en Indonesia–. Dicha pintura data de hace unos 45 000 años y representa un jabalí. El arte rupestre refleja una conciencia del sentido del tiempo, del ímpetu por trascender a su paso, de enviar un mensaje a otros "yo".

Cuando observamos arte rupestre damos por sentado que fue realizado por individuos con un sentido de autoconciencia.

Sobre la inteligencia artificial

Si, como dijimos, la inteligencia es la capacidad de aprender, adaptarse, planear, resolver y crear. Entonces, como lo propone Gardner (2011) en su teoría de las inteligencias múltiples, podemos dar por hecho que existen diferentes tipos de inteligencia en los seres humanos. Por ejemplo, hay individuos que tienen una gran habilidad para las relaciones personales, otros para crear nuevas formas o herramientas, y algunos cuya inteligencia les permite resolver problemas matemáticos complejos. También están aquellos con habilidades de liderazgo, con una habilidad destacada para las finanzas, con una mente preclara para emprender, aquellos con un talento especial para analizar situaciones complejas y proponer soluciones novedosas.

Todas estas formas de inteligencia son complementarias y es difícil encontrar a un individuo que posea todas. Sternberg (1985) proporciona una visión amplia de la inteligencia más allá del coeficiente intelectual tradicional, abarcando aspectos creativos, prácticos y analíticos. Por lo que es difícil establecer una métrica que englobe las distintas formas de inteligencia. Así, viendo las diferentes vertientes de la inteligencia, es razonable pensar que un sistema computacional que tenga la capacidad de aprender, adaptarse, y resolver problemas, puede ser considerado un sistema inteligente. Note que no estamos hablando de un grado mayor o menor de inteligencia, sino solo de la capacidad para implementar sistemas inteligentes. Los sistemas computacionales que exhiben estas capacidades implementan estrategias denominadas de inteligencia artificial (IA) (ver Russell and Norvig, 2016).

Hoy en día, contamos con herramientas sofisticadas capaces de crear imágenes a partir de descripciones textuales, resumir libros, reconocer el habla y traducirla en texto, traducir entre distintos idiomas, analizar exámenes de laboratorio y sugerir diagnósticos, reconocer objetos complejos e inferir la estructura tridimensional de escenas a partir de imágenes bidimensionales; por mencionar solo algunas tareas. No nos queda duda

de que dichos sistemas sofisticados poseen, en gran medida, habilidades que identificamos como inteligencia. La IA es una realidad con la que ya podemos convivir y obtener ventajas de su uso; a la vez enfrentamos el reto de su regulación (Bostrom, 2014; Marcus y Davis, 2019). Como demostración de las capacidades de los sistemas de IA para crear nuevas imágenes a partir de descripciones de texto hemos incluido dos imágenes generadas con asistencia del sistema DALL-E 2, ver Figuras 1 y 2 de este capítulo. Dichas imágenes sorprenden por el nivel de detalle. Sin embargo, es importante decir que para generar dichas imágenes requerimos de al menos ocho intentos por cada una. En cada nuevo intento refinamos la descripción para mejorar la generación. Un problema que tuvimos era que no se generaban correctamente el número de cazadores solicitado (variando generalmente en uno o dos), por lo que al final ajustamos el texto a la imagen generada que nos pareció más adecuada. Además, podemos notar en la Figura 1 que el lobo en primer plano parece contar con una pata trasera extra y el tigre dientes de sable luce más como un tigre de bengala: la estructura ósea del cráneo no corresponde a los dientes de sable prehistóricos. Es decir, es notable que el sistema no incorpora adecuadamente el conocimiento general de que los mamíferos tienen cuatro extremidades y es evidente su limitación para generar imágenes de elementos poco representados en la base de datos usada en el entrenamiento de DALLE-" 2. Hace falta un sistema que cierre el lazo. Un sistema que revise las posibles inconsistencias entre la inferencia y el conocimiento a priori, que una vez detectadas dichas inconsistencias, condicione una nueva inferencia para corregir dichos errores. Esto lo hicimos nosotros mismos al refinar iterativamente la descripción para que la imagen generada se fuera ajustando a lo deseado. Este proceso extra es el que haría las veces de lo que denominamos conciencia primaria en el caso del cazador sapiens; Sección 1.

Hemos evitado introducir fórmulas matemáticas, solo hasta ahora recurriremos a su uso, no lo haríamos si no creyéramos que simplificamos la exposición al recurrir al auxilio que la abstracción matemática nos otorga. Sin embargo, mantendremos la complejidad matemática en un mínimo. Dicho esto, representamos el proceso de generación de una imagen como la tarea de obtener una muestra x de una distribución p condicionada a satisfacer una descripción t:

$$x \sim p(x \mid t).$$

En esta notación $p(x \mid t)$ es la probabilidad de una imagen x dado un texto t. Por ejemplo si Texto ="una manada de lobos cazando un tigre dientes de sable en una zona de charcos en la estepa". Entonces, la Figura 1 tendrá una mayor probabilidad que la Figura 2:

$$p(x = \text{Fig. } 1 \mid t = \text{Texto}) > p(x = \text{Fig. } 2 \mid t = \text{Texto}).$$

Ahora, $x \sim$ significa que obtenemos una imagen al azar dado un texto: escogemos una imagen al azar de entre el conjunto de todas las imágenes posibles, pero la probabilidad de seleccionar una imagen en particular depende de qué tan consistente es con la descripción en el texto.

Por otro lado, al proceso de obtener la descripción (denotada por t') a partir de una imagen x lo podemos representar como:

$$t' \sim p(t \mid x).$$

En esta notación, la generación de imágenes se representa por la primera fórmula, y el análisis de una imagen por la última fórmula. Hoy en día existen sistemas de redes neuronales artificiales muy eficientes que implementan ambos modelos. Esto es, podremos introducir un texto a un modelo que implementa la primera fórmula y el resultado (imagen) introducirlo a un segundo modelo que implementa la tercera fórmula, obteniendo una descripción de la escena. Luego, podremos comparar el texto introducido con la descripción obtenida y ver el grado de acuerdo. Tenemos ahora un par de textos (t,t'). Entonces, el grado de acuerdo (consistencia) entre ambas descripciones lo podríamos denotar por $p(t' \mid t)$; que, de nuevo, lo podemos implementar mediante un sistema de IA.

En los párrafos anteriores presentamos de manera muy general el proceso de generar de datos en una representación a partir de una pista en otra representación: imágenes a partir de texto o texto a partir de imágenes. Asi mismo hemos presentado un esquema para evaluar la congruencia de lo generado. Estos esquemas no son simples suposiciones de cómo debería funcionar la IA, sino que ya son implementados por modelos de IA so-

fisticados. Lo que observamos es que en tanto tengamos datos suficientes del tipo (x = "datos de entrada", y = "salidas esperadas") es posible entrenar redes neuronales que "aprenden" a realizar la transformación: y = f(x) –siempre y cuando f sea suave: x similares correspondan a y similares.

En la actualidad estamos ante el surgimiento de reservorios masivos de datos. De hecho, los reservorios más extensos son recolectados y administrados por las grandes compañías proveedoras de servicios en Internet: telefonía, de video streaming, email, buscadores de internet, gestores de redes sociales, comercio electrónico, etc. Por lo que no es de extrañar que dichas compañías sean las que estén a la vanguardia en el desarrollo e implementación de sistemas complejos de IA.

Estamos apenas descubriendo el potencial de dichos sistemas en nuestras vidas. Para bien o para mal, la IA formará parte medular en nuestras sociedades (Marcus y Davis, 2019). El impacto de esta revolución tecnológica no la hemos siquiera imaginado, por lo que los mecanismos de regulación van atrasados con respecto al uso de la IA (Bostrom, 2014).

Sobre la conciencia artificial

La autoconciencia nos permite estar al tanto de nosotros mismos y de nuestro entorno; es tema de fascinación y estudio en filosofía, psicología y ahora en la IA. Con la finalidad de darnos una idea del reto de crear conciencia artificial nos hacemos la siguiente pregunta: ¿Es la conciencia un fenómeno raro en la naturaleza? La respuesta simple es no de acuerdo con la Declaración de Cambridge Sobre la Conciencia (Low et al., 2012):

"De la ausencia de neocórtex no parece concluirse que un organismo no experimente estados afectivos. Las evidencias convergentes indican que los animales no humanos tienen los sustratos neuroanatómicos, neuroquímicos, y neurofisiológicos de los estados de la conciencia junto con la capacidad de exhibir conductas intencionales. Consecuentemente, el grueso de la evidencia indica que los humanos no somos los únicos en poseer la base neurológica que da lugar a la conciencia. Los animales no humanos, incluyendo a todos los mamíferos y pájaros, y otras muchas criaturas, incluyendo a los pulpos, también poseen estos sustratos neurológicos."

Esta declaración realizada por un grupo de neurocientíficos y psicólogos define la sintiencia. Para distinguirla de la conciencia primaria y auto-

conciencia que hemos esbozado previamente, y que son el tema que nos ocupa. La sintiencia es la facultad de que el individuo pueda tener experiencias subjetivas (sentimientos) o conocimiento. Sentimientos que puede experimentar como positivos o negativos; sin necesidad que reflexione sobre ellos o tenga un concepto de sí mismo (Antony, 2002). No podemos poner en duda que las mascotas desarrollan afecto (apego) por sus amos. El reconocimiento explícito de que los animales desarrollan sentimientos sustenta las acciones para señalar y detener actos de crueldad animal.

Sin embargo, la autoconciencia (capacidad de pensar y razonar) no es parte necesaria de la sintiencia. La conciencia primaria puede resultar de procesos mentales que se activan al tiempo que se llevan a cabo los procesos mentales relacionados con la actividad que se realiza. En los ejemplos de los grupos cazadores de lobos o humanos, en plena cacería se activaban sentimientos de furia o temor. Lo que indica la emergencia y ocurrencia de procesos mentales paralelos en especies con sistemas nerviosos desarrollados. Lo que haría suponer que la conciencia primaria, que implica procesos mentales no reactivos sino de evaluación de los propios procesos mentales, también es factible en animales superiores. Caben las preguntas: ¿qué tan superiores? ¿Solo los sapiens hemos desarrollado conciencia primaria y autoconciencia? ¿Somos la excepción?

En las secciones anteriores presentamos argumentos que nos permiten ver a la autoconciencia como consecuencia y necesidad en especies, como el sapiens, con cerebros complejos. Sin embargo, los sapiens no son la única especie inteligente que ha manifestado claramente capacidad de ser autoconsciente. El Homo Neanderthalensis (los Neandertales, una especie de humanos extinta hace unos 40 000 años cuya presencia se extendió por Europa y norte de Asia) también fue capaz de crear herramientas, tener una vida social compleja, crear joyería, realizar pinturas rupestres complejas y realizar ceremonias fúnebres. Como referencia, entre las pinturas rupestres más antiguas de Europa se encuentran las de la cueva de Maltravieso (Extremadura, España) atribuidas a Neandertales y datan de aproximadamente unos 67 000 años (Hoffmann et al., 2018). También en la cueva de La Ferrassie (Dordoña, Francia) se encontraron restos de un niño que evidenciaban prácticas funerarias propias de procesos cognitivos complejos realizadas por neandertales hace cerca de 41 000 años (Balzeau et al., 2020).

Los neandertales fueron una especie social, creativa, que cuidaba de sus congéneres débiles, que construía herramientas de piedra, y que creaba joyería a partir de conchas y cuentas de piedra por el puro placer estético que brindaban tales objetos. Las pinturas neandertales fueron plasmadas mucho tiempo antes de la pintura rupestre de sapiens más antigua conocida en Indonesia. El fechado de la pintura sapiens en Indonesia coincide con la llegada de los sapiens al continente Europeo; alrededor de 20 000 años después de las pinturas neandertales en la cueva de Maltravieso. El lapso de tiempo transcurrido entre las mencionadas pinturas neandertales y las sapiens es aproximadamente el doble del lapso entre el invento de la agricultura y hoy día.

El estudio de la evolución de la inteligencia y de la conciencia humana nos ha mostrado que al menos en dos ocasiones el mecanismo evolutivo ha producido especies autoconscientes. Seguramente, la conciencia de ambas especies difería. A la naturaleza le ha tomado más de dos millones evolucionar del Homo habilis al Homo sapiens: de la especie que elaboró la primera hacha de piedra, a la que acuñó el término de IA. A nosotros, los sapiens, nos ha tomado menos de un siglo pasar de la publicación por McCulloch and Pitts (1943) del modelo de la neurona hasta el desarrollo de los grandes modelos de lenguaje (ChatGPT, Llama, Claude; por mencionar algunos). Con dichos sistemas, que implementan técnicas de IA, podemos entablar conversaciones que con dificultad las distinguimos de las que entablamos con otro sapiens [prueba de Turing (1950)]. Lo que nos lleva a pensar que, como en el caso de la IA, solo es cuestión de tiempo para que seamos testigos de la aparición de modelos computacionales que manifiesten comportamiento que podamos definir como un tipo de "conciencia artificial", del tipo conciencia primaria o autoconciencia.

En el contexto de las redes neuronales y sistemas de IA, con su evidente complejidad creciente, diseñadas para procesar y evaluar grandes cantidades de datos, que aprenden y se adaptan de manera similar a un cerebro biológico. La sola idea de que estas redes puedan autoevaluar su proceso de inferencia podría considerarse un paso hacia una forma rudimentaria de autoconsciencia. De acuerdo con Kurzweil (2005), en esta etapa, la IA no solo ejecutaría tareas, sino que también desarrollaría una capacidad para monitorear y ajustar sus procesos cognitivos. Dado que la conciencia

primaria es un proceso mental y, como tal, podría ser representado por un modelo de IA. Cada vez nos acercaríamos más a implementar lo que podríamos denominar "conciencia primaria artificial".

La posibilidad de que las redes neuronales implementen una forma de autoconciencia es un área de gran interés y especulación en el campo de la IA. Queda abierta la cuestión si la conciencia, o al menos una forma primitiva de ella, podría ser una necesidad o una consecuencia inevitable de que estas redes se vuelven cada vez más sofisticadas y capaces de realizar tareas complejas: ¿podrán alcanzar un estado de autoconciencia rudimentaria, donde reconozcan y ajusten sus propios procesos de inferencia (razonamiento)? Bostrom (2014) explora la implicación de IA avanzada, incluyendo la posibilidad de una conciencia emergente.

La conciencia humana tiende a verse desde aproximación unificada, pero debe ser considerarla en sus distintas capacidades, tipos y niveles (Montemayor, 2021; Morin, 2006). Basados en este reconocimiento de niveles de conciencia, postulamos que debemos estar abiertos a flexibilizar nuestro concepto de conciencia (como lo hemos hecho respecto a la inteligencia) y considerarla como un proceso mental que evalúa de los propios procesos mentales. La idea de que una IA pueda no solo ser consciente sino también reflexiva, es decir, capaz de pensar sobre sí misma, es un tema fascinante. Esto plantearía preguntas sobre la naturaleza de la conciencia y si la conciencia artificial podría emular o incluso superar la experiencia humana.

Es decir, que pase con la conciencia artificial lo que Kurzweil (2005) pronostica al decir que estamos en los albores de la llamada singularidad: formas de IA avanzadas, que incluyen la autoconciencia y la reflexividad. Si se desarrolla un nivel de conciencia primitiva en IA, surge la cuestión de si este sistema podría, eventualmente, reflexionar sobre sí mismo y evolucionar hacia una forma de conciencia más avanzada. Esta posible conciencia artificial podría, en teoría, emular aspectos de la conciencia humana, pero también diferir de ella en formas fundamentales. La conciencia artificial podría no incluir elementos emocionales o subjetivos como los experimentamos los humanos, o podría desarrollar formas completamente nuevas de experiencia consciente, dictadas por su arquitectura y capacidades únicas.

Tendemos a colocarnos en la cima evolutiva y damos por hecho que no hay un escalón más allá. Esta visión antropocéntrica la hemos reforzado con imágenes icónicas como la llamada "Marcha del progreso"(imagen popular donde se observan una secuencia que inicia en un Mono, pasa por un Australopithecus, un Homo Erectus y concluye en un Sapiens) para asumirnos como el resultado final del proceso evolutivo; como si toda especie aspirara a ser Sapiens. Sobrevaloramos aquellas habilidades que, en nuestra opinión, nos distinguen de otras especies y por ello no concedemos que puedan ser emuladas, ya sea por sistemas con sustrato de carbono (otras especies de seres vivos) o con sustrato de silicio (sistemas computacionales actuales).

En este trabajo no pretendemos presentar una teoría sobre la construcción y aparición de la conciencia [para ello recomendamos revisar Chalmers (1997)], solo la exponemos como una consecuencia de la evolución de especies con grado cognitivo complejo. Reconocemos que la IA es un hecho hoy en día. Sabemos que modelos de conciencia artificial están siendo investigados en distintos laboratorios de computólogos por el mundo. ¿Cuándo veremos sistemas con conciencia artificial? ¿Cómo serán? ¿Cómo los distinguiremos? ¿Cuáles serán sus implicaciones en nuestras vidas y nuestro futuro? Más que con respuestas, terminamos con una frase atribuida a Yogi Berra: el futuro ya no es lo que solía ser.

Referencias

Antony, M. V. (2002). *Concepts of consciousness, kinds of consciousness, meanings ofconsciousness*. Philosophical studies, 109, 1-16.

Baars, B. J. (1997). *In the theater of consciousness: The workspace of the mind*. Oxford University Press, USA.

Balzeau, A., Turq, A., Talamo, S., Daujeard, C., Guérin, G., Welker, F., ... others (2020). *Pluridisciplinary evidence for burial for the la ferrassie 8 neandertal child*. Scientific reports, 10(1), 1–10.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press, Oxford.

Chalmers, D. J. (1997). *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks.

Damasio, A. R. (1995). *Descartes' error*. Picador.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt.

Dunbar, R. I. (1998). *The social brain hypothesis. Evolutionary Anthropology: Issues, News, and Reviews,* 6(5), 178–190.

Gardner, H. E. (2011). *Frames of mind: The theory of multiple intelligences*. Basic books.

Gordon, P. (2004). Numerical cognition without words: Evidence from amazonia. *Science*, 306(5695), 496-499.

Henshilwood, C. S., d'Errico, F., Van Niekerk, K. L., Coquinot, Y., Jacobs, Z., Lauritzen, S.-E., ... García-Moreno, R. (2011). A 100,000-year-old ochreprocessing workshop at blombos cave, south africa. *Science*, 334(6053), 219–222.

Henshilwood, C. S., d'Errico, F., Yates, R., Jacobs, Z., Tribolo, C., Duller, G. A., ... others (2002). Emergence of modern human behavior: Middle stone age engravings from south africa. *Science*, 295(5558), 1278–1280.

Hoffmann, D. L., Standish, C. D., García-Diez, M., Pettitt, P. B., Milton, J. A., Zilh˜ao, J., ... others (2018). U-th dating of carbonate crusts reveals neandertal origin of iberian cave art. *Science*, 359(6378), 912-915.

Kurzweil, R. (2005). *The singularity is near. In Ethics and emerging technologies* (pp. 393–406). Springer.

Low, P., Panksepp, J., Reiss, D., Edelman, D., Van Swinderen, B., & Koch, C. (2012). The cambridge declaration on consciousness. In *Francis crick memorial conference* (Vol. 7).

Marcus, G. y Davis, E. (2019). *Rebooting ai: Building artificial intelligence we can trust*. Vintage.

McCulloch, W. S., & Pitts, W. (1943). *A logical calculus of the ideas immanent in nervous activity*. The bulletin of mathematical biophysics, 5, 115-133.

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*. mit Press.

Mithen, S. (1997). The prehistory of the mind. *Cambridge Archaeological Journal,* 7, 269-269.

Montemayor, C. (2021). Types of consciousness: The diversity problem. *Frontiers in Systems Neuroscience*, 15, 747797.

Morin, A. (2006). Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. *Consciousness and cognition*, 15(2), 358-371.

Russell, S. J. y Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

Sternberg, R. J. (1985). *Beyond iq: A triarchic theory of human intelligence*. CUP Archive.

Tomasello, M. (2009). Why we cooperate. *MIT press*.

Turing, A. M. (1950). I.—*Computing machinery and intelligence*. Mind, LIX(236), 433-460.