CUADERNOS SALMANTINOS DE FILOSOFÍA
Vol. 48, 2021, 105-139, ISSN: 0210-4857
E-ISSN: 2660-9509

DE LAS EMOCIONES NATURALES A LA EMOCIONALIDAD ARTIFICIAL

FROM NATURAL EMOTIONS TO ARTIFICIAL EMOTIONALITY

José Miguel Biscaia Fernández
Graduado y Doctorando en Filosofía
Licenciado en Biología, Máster en Biotecnología y Doctor en Neurociencia
Profesor Titular de Fisiología
Facultad de Medicina
Universidad Europea de Madrid
Madrid/España
josemiguel.biscaia@universidadeuropea.es
ORCID: 0000-0002-3496-5527

Recibido: 30/04/2021
Revisado: 24/07/2021
Aceptado: 6/09/2021

Resumen: La innovación en inteligencia artificial se encuentra en la vanguardia científico-tecnológica y muchas de sus aplicaciones presentan ya una alta interactividad humana. Dado que somos seres emocionales, y que esta cualidad psicobiológica es clave para nuestro posicionamiento en el mundo, parece necesario realizar una profunda reflexión sobre el espectro afectivo de la díada humano-máquina. El presente ensayo utiliza el término "emocionalidad artificial" como concepto holístico capaz de abordar este análisis desde una triple perspectiva: la de las ciencias cognitivas y de la computación, la de la neurociencia y la psicología y la de la filosofía teórica y práctica. Partiendo de la descripción de las emociones naturales, en este estudio se discute sobre la posibilidad técnica y conceptual y sobre las consecuencias bio-psico-sociales de una inteligencia artificial con capacidad de reconocimiento, simulación, manipulación y vivencia emocional. Tras dicho análisis se concluye que las dos primeras capacidades son ya en cierto modo posibles y deseables, mientras que las dos últimas se enfrentan a dificultades tecno-científicas, ontológicas, gnoseológicas y neuroéticas discutidas ampliamente por el transhumanismo.

Palabras clave: Emoción, inteligencia artificial, filosofía de la mente, transhumanismo, neuroética.

Abstract: Innovation in artificial intelligence is at the forefront of science and technology, and many of its applications already have high human interactivity. Given that we are emotional beings, and that this psychobiological quality is key to our positioning in

the world, it seems necessary to make a deep reflection on the affective spectrum of the human-machine dyad. This essay proposes the term of "artificial emotionality" as a holistic concept capable of approaching this analysis from a triple perspective: that of cognitive and computer sciences, that of neuroscience and psychology, and that of theoretical and practical philosophy. Starting from the description of natural emotions, this study discusses the technical and conceptual possibility and the bio-psycho-social consequences of an artificial intelligence capable of recognition, simulation, manipulation and emotional experience. After said analysis it is concluded that the first two capabilities are already in a certain way possible and desirable, while the last two face techno-scientific, ontological, epistemological and neuroethical difficulties widely discussed by transhumanism.

Keywords: Emotion, artificial intelligence, philosophy of mind, transhumanism, neuroethic.

INTRODUCCIÓN

Desde su nacimiento a mediados del siglo XX, a partir de las ideas de Alan Turing, la inteligencia artificial (IA) no ha hecho más que progresar. Si bien no se ha cumplido la "Ley de Moore" (1965), que proponía un crecimiento exponencial de la tecnología cibernética, nadie negará que sus aplicaciones actuales y sus futuros desarrollos se encuentran en la avanzadilla de la ciencia y la tecnología. Entre todas estas aplicaciones, las más destacadas son las que se relacionan con el lenguaje natural y el reconocimiento del habla, la visión artificial, la robótica o la conducción autónoma.

Una de las principales características de estos desarrollos en IA es su alto grado de interacción con nosotros. Convivir con máquinas inteligentes supondrá, como mínimo, una adaptación a los nuevos usos tecnológicos y, en su límite, un auténtico cambio de paradigma relacional ético-social. Teniendo en cuenta que somos seres emocionales, y que dicha cualidad psicobiológica es clave para comprender nuestra relación con el mundo, se antoja imprescindible que reflexionemos de un modo bidireccional al respecto del continuo afectivo humano-máquina.

Así pues, este artículo propone una visión integrada en el estudio de las emociones y de la inteligencia artificial, que no se limite sólo a lo técnico o sólo a lo filosófico (como si fueran ámbitos estancos); que ofrezca, de este modo, una aproximación holística al tema de estudio:

— Utilizando el concepto de "emocionalidad artificial" (EA) como alternativa a otros similares (aunque pretendidamente más inclusivo), tales como el de "emoción sintética" (Casacuberta y Vallverdú, 2010), "emoción artificial" (Spinola y Queiroz, 2007), "inteligencia artificial emocional" (Schuller y Schuller, 2018), "arquitectura emocional" (González y García, 2006), "computación afectiva" (Picard, 1997) o "inteligencia artificial de

la emoción" (Chakriswaran et al., 2019). Con dicho término se pretende trascender el mero análisis técnico de las emociones artificiales como subdisciplina de la IA, ampliando sus márgenes para albergar de este modo a la tríada máquina-humano-mundo.

— Planteando un abordaje multidisciplinar, conexionista y funcional de las materias de las que se nutre esta área de conocimiento. De este modo, cualquier análisis profundo del tema exigirá tener una aproximación plural:
  — Desde las ciencias cognitivas y de la computación: describiendo el estado tecnológico de la cuestión, desde lo ya realizable hasta la vanguardia y la especulación tecno-científica.
  — Desde la neurociencia y la psicología: usando las bases neuroquímicas, neurofisiológicas, neuroanatómicas y psicológicas del cerebro humano como modelo comparado para el desarrollo de la IA.
  — Desde la filosofía y las humanidades: contemplando, desde una perspectiva teórica, las dificultades y límites ontológicos y epistémicos de los futuros desarrollos en IA y, desde una perspectiva práctica, la necesidad, utilidad y alcance de la IA en la praxis humana (ética, sociedad, cultura, derecho, economía o política).

De este modo, en relación con la emocionalidad artificial son varias las preguntas generales que pueden discernirse en el horizonte: por un lado, desde la perspectiva de la IA, ¿es tecnológica y conceptualmente posible desarrollar capacidades emocionales en la IA? Si la respuesta es afirmativa, ¿qué aspectos contempla dicha emocionalidad y qué utilidad y sentido podría tener el dotar a una máquina con capacidad de reconocimiento, imitación, manipulación o vivencia afectiva? Por otro lado, desde la perspectiva humana, ¿es posible acceder (y comprender) a los posibles estados emocionales de una IA, incluso manifestar sentimientos por una máquina? Y, a modo de síntesis de todas ellas, en virtud del modo pretendidamente integrador del término "emocionalidad artificial", ¿cómo será un mundo en el que humanos y máquinas interaccionen afectivamente, qué consecuencias éticas y sociales tendrá?

En el presente estudio, más allá de introducir el debate general de discusión en torno al concepto de emocionalidad artificial, se reflexionará sobre algunas de las cuestiones arriba planteadas, en particular en relación a los aspectos más epistémico-ontológicos, si bien se seguirá el espíritu integrador propuesto y de forma paralela se discutirán aspectos técnicos y filosófico-pragmáticos. No pretende este ensayo, por tanto, agotar ni mucho menos las posibilidades de análisis al respecto de la emocionalidad artificial. En consecuencia, con el objetivo de dar

respuesta a algunas de las preguntas que se han propuesto, se seguirá el siguiente orden argumentativo:

1. En primer lugar, dado que de emociones artificiales trata este ensayo, se partirá de la descripción de las emociones naturales desde una perspectiva filosófica y psicológica.
2. A continuación, una vez explicadas, servirán para el análisis comparado de las emociones artificiales. En el marco teórico de la filosofía de la mente y de las ciencias cognitivas y de la computación, con el objetivo de describir el estado tecnológico de la cuestión y abundar en los límites ontológicos, gnoseológicos y pragmáticos de la emocionalidad artificial, haré una diferenciación interesada entre:
   2.1. IA que reconoce emociones.
   2.2. IA que simula emociones.
   2.3. IA que manipula emociones.
   2.4. IA que siente emociones.

1. EMOCIONES NATURALES

Parece lógico que para discutir sobre emociones artificiales partamos antes de sus predecesoras, a saber, las emociones naturales de los seres biológicos. Definir qué son y explicar cuál es su función resultará útil para, mediante un análisis comparado, determinar si es posible y deseable que una IA igualmente las manifieste o presente alguna capacidad en relación con ellas.

No obstante, si lo que se pretende es dar una definición canónica y unívoca del concepto "emoción" nos encontramos enseguida en un callejón sin salida, pues pocos términos son tan inasibles como el que nos ocupa. De este modo tan intuitivo lo expresaban tempranamente Wenger, Jones y Jones (1962: 3): "casi todo el mundo piensa que sabe lo que es una emoción, hasta que intenta definirla. En ese momento prácticamente nadie afirma poder entenderla". Y es que es tal la complejidad de este fenómeno psicológico que falta una teoría integrada (única y concluyente) que pueda dar cuenta de dicho concepto desde una perspectiva descriptiva, explicativa y predictiva (Cano-Vindel, 1995; Reisenzein, 2019; Tantam, 2003). Una explicación adicional de esta dificultad cognoscitiva probablemente tenga que ver con la variedad de paradigmas epistemológicos desde los que la ciencia y la filosofía se ha aproximado a este nebuloso fenómeno (donde, además, en muchas ocasiones la evidencia empírica ha ido por detrás de los modelos teóricos). Sin embargo, a pesar de las citadas dificultades conceptuales del propio evento emocional y del modo de acercarse hasta él, la ciencia empírica ha hecho

extraordinarios progresos en los últimos años, partiendo de diferentes paradigmas científico-experimentales y de distintas orientaciones metodológicas como la conductual (basada en la psicología del aprendizaje), la (neuro)biológica o la cognitiva (Fernández-Abascal et al., 2010; Reisenzein, 2019).

Tradicionalmente, las diferentes escuelas del pensamiento occidental han mantenido posturas diversas con respecto a las emociones, que van desde el desprecio y ostracismo de las mismas hasta la consideración de lo emocional como fenómeno digno de consideración y estudio filosófico[1]. Con respecto al citado descrédito, tal y como reconoce Broncano (2001), son muchos los autores que niegan a las emociones un valor intrínseco y una significación propia, incluso que consideran que no son una clase natural, reduciéndolas con frecuencia a otras instancias como disposiciones, intenciones, conductas, perturbaciones fisiológicas, actitudes proposicionales o juicios evaluativos. Entre sus más destacados detractores contemporáneos encontramos a Ryle (1967), Griffiths (1997) o Elster (1999).

Frente a dichas críticas, el filósofo español señala que las emociones son fenómenos complejos y difusos, aunque de gran interés para la filosofía, pues si bien están en un territorio entre lo fisiológico y lo conceptual, comparten (aunque

1 En la historia de la filosofía se puede sondear el interés de los autores clásicos por el tema de las emociones (o pasiones, del griego *pathos,* como habitualmente eran nombradas). De un lado está la postura general de quienes no se las han tomado en serio, por estimarlas como algo menor; de quienes, incluso, las han considerado como impulsos irracionales internos que debilitan y ensombrecen la razón e intelección humana. Platón, por ejemplo, distinguía en el alma tres dominios separados: el racional, el irascible (sentimientos nobles) y el concupiscible (sentimientos inferiores). Su mito del carro alado, presente en el diálogo del *Fedro*, ejemplifica el dominio de lo cognitivo sobre las bajas pasiones para alcanzar así la virtud y la verdad. En esta línea despreciativa, la escuela estoica tampoco se tomó demasiado en serio a las pasiones, considerándolas en todo caso como elementos perturbadores. Aristóteles, sin embargo, quien las menciona en obras como *Ética eudemia, Ética nicomaquea o Retórica*, les otorgó más valor, y creía que la parte racional e irracional (lo emocional-pasional) formaban una unidad que guía al hombre. Pasados los siglos, el dualismo ontológico de René Descartes diferencia entre *res cogitans* y *res extensa*, dando mayor importancia a la razón y señalando un tanto despreciativamente que lo emocional es conducido por "espíritus animales". Como reconoce en *Las pasiones del alma* de 1649, el *cogito* humano puede llegar a dominar lo concupiscente; los animales, por el contrario, al ser considerados como seres sin alma (meros autómatas), dependen enteramente de fuerzas externas e internas. Con el asociacionismo, mecanicismo y empirismo defendido por autores como John Locke o David Hume, se empieza a considerar la relevancia de las emociones en los procesos mentales superiores. Gracias a las ideas sobre el origen biológico de la conducta (y de las emociones) de Charles Darwin (mencionadas en su obra de 1872 *La expresión de las emociones en el hombre y en los animales*) y con el nacimiento de la psicología y la fisiología modernas a manos de autores como Walter Cannon, William James o Carl Lange, se sentarán las bases del estudio científico-experimental de las emociones (para una aproximación más profunda a la historia del estudio de las emociones consultar el capítulo 1 de *Psicología de la emoción* de Fernández-Abascal et al. [2010: 18-37] y las revisiones de Casado y Colomo [2006] y de Pinedo y Yáñez [2018]).

sea parcialmente) varias características con los estados mentales tradicionalmente estudiados por la filosofía de la mente; más en concreto (Broncano, 2001):

— Tienen intencionalidad, pues las emociones presentan contenido que se dirige hacia el objeto que despierta el estado emocional (por ejemplo, se desea o se teme a algo o a alguien en particular). Aunque hay que diferenciar a las emociones de otro fenómeno similar, como son los estados de ánimo, pues estos últimos, que se caracterizan por una mayor vigencia temporal frente al "instante" de la emoción, no muestran un objeto definido. Reconoce, no obstante, que dicho contenido emocional podría ser no-conceptual:

> Las emociones conforman un sistema de señales o mensajes que porta y procesa contenido, aunque quizás este contenido es no-conceptual […] Las emociones, o lo que sea el correlato neuronal del sistema emotivo […] pueden portar información, pero solamente cuando es descriptible conceptualmente se convierte en mental (Broncano, 2001: 12).

— Contribuyen a la individuación, ya que las emociones anclan los estados mentales en un contexto de evaluaciones, en un eje espacio-temporal clave para la configuración del individuo y para la valoración del mundo.

— Comunican estados mentales, gracias a que la expresión de las emociones es una herramienta vehicular de ciertos estados internos de otro modo inobservables. La "Teoría de la mente" (capacidad de atribuir estados mentales en el otro) da buena cuenta de esta observación (Tirapu-Ustárruz et al., 2007).

— Ayudan en la evaluación de situaciones, especialmente en contextos de urgencia cuando ha de hacerse una valoración rápida que posibilite igualmente una respuesta ajustada a las circunstancias. Son, por tanto, un soporte de la maquinaria evaluadora cognitiva.

— Participan en la creación de constructos socio-culturales, en la medida en que, según algunos autores, las normas sociales y morales derivan de nuestra capacidad emocional (Gibbard, 1990). En este sentido, la mayoría de neurocientíficos y filósofos coinciden en que los juicios morales son intuitivos y están muy vinculados a la emocionalidad, y sólo alcanzamos (a veces) a dar razones de por qué hicimos tal o cual cosa a posteriori, si se nos pregunta o reflexionamos conscientemente sobre ello (Cortina, 2011). Así pues, las emociones tendrían un rol causal en el nivel lingüístico-social, creando un sistema de compromisos que soporta

algunas de las más altas instancias creadas por el aparato cognitivo del hombre, como son las normas o las instituciones sociales[2].

La complejidad psicobiológica y la multidimensionalidad del fenómeno emocional puede apreciarse claramente si se observan las más de cien definiciones analizadas por Kleinginna y Kleinginna (1981); a partir de ellas, los autores propusieron once categorías descriptivo-explicativas que bien pueden servir para agrupar y detallar de una forma científica, rigurosa y sistemática todas estas formas de conceptualizar la emoción, a saber:

— Categoría afectiva: tiene que ver con la dimensión subjetiva y experiencial de las emociones. Al amparo de esta conceptualización se encontrarían los *qualia* emocionales (cómo es sentir la vivencia de una determinada emoción o estado afectivo). En psicología suele utilizarse el término "sentimiento" para referirnos a esta experiencia subjetiva emocional.

— Categoría cognitiva: se relaciona con aspectos valorativos, informacionales, clasificatorios y perceptivos de las emociones. Esta conceptualización es la que más asemeja los estados emocionales a otros estados mentales.

— Categoría basada en estímulos elicitadores: tiene que ver con causas desencadenantes externas, es decir, con eventos disparadores de las emociones procedentes del entorno (como una amenaza o un objeto de deseo).

— Categoría fisiológica: hace mención a las modificaciones fisiológicas, a los mecanismos neurobiológicos subyacentes y al sustrato neuroanatómico de las emociones (donde destaca, por encima de todos, el denominado como sistema límbico emocional)[3].

2 Emociones autoconscientes como el orgullo, la culpa o la vergüenza son fundamentales como elementos motivadores y controladores de la conducta moral (Fernández-Abascal et al., 2010: 431).

3 El estudio neurobiológico de las emociones (lo que se conoce como "neurociencia afectiva") ha tenido un crecimiento exponencial en las últimas décadas (Esperidiao-Antonio et al., 2017). La mayoría de investigaciones se han realizado a partir de pacientes con lesiones cerebrales que suponían algún déficit cognitivo-emocional, utilizando modelos animales en los que se alteraba quirúrgica o farmacológicamente alguna región neuroanatómica emocional o empleando sofisticadas técnicas de neuroimagen que correlacionan anatomía y función como la resonancia magnética funcional (fMRI). Gracias a estas aproximaciones metodológicas se han encontrado estructuras claramente implicadas en el procesamiento emocional como la amígdala cerebral, los núcleos septales, la formación hipocampal, diferentes núcleos diencefálicos o varias regiones corticales como la corteza orbitofrontal o la corteza cingulada. Éstas y otras estructuras adicionales forman parte del circuito emocional de nuestro encéfalo que se conoce con el nombre de sistema límbico. No obstante, según Ledoux (2000), dicho sistema no es un concepto claramente definido, ni anatómica ni funcionalmente hablando, por lo que en palabras de Fernández-Abascal et al. (2010: 50), "actualmente no se mantiene la existencia de un

— Conceptualización emocional/expresiva: se refere a la expresión externa de las emociones experimentadas, es decir, a la conducta observable que se desencadena. Es la que permite dotar a las emociones de carácter comunicativo.
— Categoría disruptiva: hace hincapié en aspectos disfuncionales y desorganizadores de las emociones, como fenómenos perturbadores que alejan al sintiente de la normalidad funcional. Esta perspectiva es la que desde la filosofía clásica ha denostado tanto a las emociones y, desde la perspectiva actual, justifica los trastornos psiquiátrico-emocionales.
— Categoría adaptativa: contraria a la anterior, pone el acento en el papel funcional y evolutivo de las emociones; en su alta conservación filogenética como garantía de su eficacia para responder a las demandas del entorno y alcanzar la supervivencia individual.
— Categoría multifactorial: considera la emoción como un conjunto amplio de fenómenos, algunos de ellos ya descritos, con diferentes dimensiones: afectiva, motivacional, cognitiva, fisiológica y conductual.
— Conceptualización restrictiva: define la emoción por contraste con otros procesos psicológicos diferentes, como puede ser la percepción o la cognición.
— Categoría motivacional: plantea una vinculación entre emoción y motivación, siendo lo primero elicitador de lo segundo.
— Categoría escéptica: dada la dificultad definitoria y explicativa, resta importancia a este esfuerzo y asume que es una tarea inabarcable e, incluso, epistémicamente cuestionable.

Como se puede deducir, resulta muy difícil determinar cuáles de los elementos expuestos son causa suficiente y necesaria para que un fenómeno sea categorizado como emocional; muy arduo, además, porque son múltiples los contraejemplos que podrían ofrecerse sobre la mencionada categorización. No obstante, Fernández-Abascal et al. (2010: 19-20) considera que a partir de estas conceptualizaciones hay cierto consenso científico sobre cuáles son los cuatro elementos esenciales para entender qué son las emociones: (1) en las emociones hay cambios fisiológicos[4]; (2) en las emociones hay una "tendencia a la acción" o

circuito único y general que explique el procesamiento emocional", afirmación avalada por la descripción reciente de múltiples subsistemas emocionales íntimamente conectados (Vogt, 2019).

4 Muchos autores defienden que la activación fisiológica es condición necesaria, aunque no suficiente, para que se desencadene una emoción. Lo "necesario" hace mención a que el sujeto debe evaluar y valorar tanto el estado de activación como la situación contextual en que se produce. Según las primeras teorías cognitivo-emocionales desarrolladas en los 60-70 del pasado siglo, como la "Teo-

afrontamiento que ayuda a responder frente a las demandas del entorno; (3) en las emociones hay una experimentación subjetiva o sentimiento; (4) en las emociones se produce un procesamiento de información. Así pues, en un intento de concretar y asir el nebuloso fenómeno de las emociones, una definición con la que trabajaré de ahora en adelante será la siguiente:

> Las emociones son un proceso que implica una serie de condiciones desencadenantes (estímulos relevantes), la existencia de experiencias subjetivas o sentimientos (interpretación subjetiva), diversos niveles de procesamiento cognitivo (procesos valorativos), cambios fisiológicos (activación), patrones expresivos y de comunicación (expresión emocional), que tiene unos efectos motivadores (movilización para la acción) y una finalidad: que es la adaptación a un entorno en continuo cambio (Fernández-Abascal et al., 2010: 40-41).

Una teoría explicativa surgida hace algunas décadas, aunque de gran vigencia aun en la actualidad, basada en el rol funcional de las emociones, que acumula gran cantidad de evidencia científica y que presenta interesantes implicaciones en otros procesos psicológicos cognitivos como la memoria o la toma de decisiones, es la "Teoría funcional del sistema emotivo" (Oakley y Jonhson-Laird, 1987: 13). Según los autores:

> Las emociones son sistemas de señales que indican a la mente transiciones de planes. Su función es detectar posibles objetivos en el medio externo o interno que son relevantes (positiva o negativamente) a un plan en marcha y disponer al organismo a una biblioteca de planes potenciales de acción que está almacenada en la memoria a largo plazo, pero que gracias a las conexiones rápidas del sistema emotivo (mediante un sistema de marcadores emotivos de la información) permite una activación mucho más rápida que a través de los medios habituales de recuperación de la memoria. El sistema emotivo parte el mundo (externo o interno) en categorías mucho más amplias que el lenguaje y el sistema conceptual, porque son categorías de control rápido de planes.

Como se deduce de este texto, resulta inaceptable explicar las características y función de las emociones sin atender a otros procesos mentales. Es cierto que el inicio del procesamiento emocional (en regiones como la amígdala cerebral)

---

ría bifactorial de la emoción" o la "Teoría de la evaluación-discrepancia", la activación indiferenciada correlacionaría con la intensidad emocional, mientras que la cualidad afectiva estaría determinada por la interpretación emocional (lo evaluativo-cognitivo). Los conceptos de "evaluación" (basado en la novedad y el agrado situacional) y "valoración" (basado en el significado, la causalidad y la normatividad social de la situación) son ampliamente reconocidos en las teorías actuales como elementos clave en el procesamiento emocional: los cambios fisiológicos que se experimentan en una emoción serían fruto de la evaluación y valoración situacional y serían fundamentales para preparar al organismo en la eventual ejecución de una acción de aproximación o evitación, según el caso (Fernández-Abascal et al., 2010).

puede ser automático e inconsciente (Janak y Tye, 2015), una especie, concluyen muchos psicólogos, de "vía rápida" de procesamiento. Sin embargo, las emociones más complejas son fruto de acciones deliberadas que exigen consciencia (sería la "vía precisa"), como la evaluación situacional, el conocimiento adquirido previamente o nuestras expectativas, capacidad planificadora y toma de decisiones, aspectos todos ellos relacionados claramente con nuestra cognición (donde están implicadas áreas cerebrales como la corteza prefrontal [Dixon et al., 2017]), lo cual nos transporta a las ideas de Broncano (2001) expuestas más arriba sobre la complejidad de los estados emocionales y su similitud con otros estados mentales tradicionalmente estudiados por la filosofía de la mente. Así pues, podemos concluir que lo cognitivo-emocional forma un continuo explicativo bidireccional en gran parte del procesamiento mental[5]. Dicho de otro modo, las emociones conforman una especie de sistema multinivel de procesamiento informacional que es clave para la supervivencia. Lo emocional tamiza el mundo (interior y exterior), lo categoriza, evalúa y valora de forma sencilla (aunque también rápida y eficazmente), para de este modo dar una respuesta adaptativa que favorezca la supervivencia del individuo en un entorno cambiante. Podemos concluir, con Reeve (1994), que las emociones tienen tres grandes funciones: adaptativas (porque contribuyen a la supervivencia y al bienestar individual), sociales (porque intervienen en las relaciones entre individuos) y motivacionales (porque movilizan recursos para la acción, favoreciendo o perjudicando el acercamiento o la evitación).

Otro aspecto a destacar en el análisis de las emociones, de interés para las intenciones de este artículo, es el intento clasificatorio que durante años lleva realizando la psicología. La dificultad deriva de los problemas descriptivos y explicativos ya comentados, además del criterio agrupador que se considere. Sin embargo, los esfuerzos en esta tarea no han decaído, no sólo por abundar en el conocimiento del fenómeno emocional sino, también, por la aplicabilidad terapéutica en el tratamiento de desórdenes psiquiátrico-emocionales. Así pues, hay dos grandes criterios agrupadores: (1) el dimensional, que define el mapa emocional basándose en tres dimensiones generales (valencia afectiva, activación y control), y (2) el discreto, que reconoce características específicas y distintivas

5 Prueba adicional de esta relación entre lo emocional y otros fenómenos cognitivos la encontramos sobradamente en la obra *Inteligencia emocional* de Daniel Goleman (1995). Como resumen de esta vinculación, Casacuberta y Vallverdú (2010: 120) señalan que "nuestras emociones están en el núcleo mismo de la racionalidad". En esta misma línea, el célebre investigador Antonio Damasio (1994) dejó a las claras en su obra *El error de Descartes* la evidente conexión entre emoción y razón. En cualquier caso, "si el término cognitivo alude a procesamiento activo de la información, la mayor parte, sino todas las emociones, requieren algún tipo de procesamiento cognitivo" (Fernández-Abascal, 2010: 61), por ejemplo, como mínimo, el implicado en los patrones evaluativos (sean o no conscientes).

en cada emoción (Fernández-Abascal et al., 2010). Siguiendo este último criterio, psicólogos como Ekman (1992) definieron seis emociones básicas o primarias (sorpresa, asco, miedo, alegría, tristeza e ira), las cuales son consideradas como innatas y universales, y surgen antes en el desarrollo individual. Adicionalmente, la psicología afirma que también poseemos emociones secundarias o complejas, también denominadas "autoconscientes" o "autoevaluativas" (como la vergüenza, la culpa, el orgullo, los celos, etc), fruto de nuestra compleja interactividad social (Lewis, 2000).

2. EMOCIONALIDAD ARTIFICIAL

El debate filosófico en torno a la emocionalidad artificial (EA) puede desarrollarse a lo largo de un continuo en cuyos extremos aparecen dos polos diferenciados: el de la filosofía teórica y el de la filosofía práctica. Desde el primer núcleo de reflexión se puede dar respuesta a cuestiones ontológicas y epistemológicas en relación a las capacidades de la emoción artificial; desde la filosofía práctica, el análisis se construye sobre aspectos relacionados con la acción humana, como la ética, la estética o la política. Dicho de modo más claro, la filosofía teórica tratará de responder a la pregunta de si es conceptualmente posible una IA con capacidad emocional; desde la filosofía práctica, las preguntas girarán en torno a la necesidad, utilidad y consecuencias de la EA en la praxis humana. Pues bien, en el presente ensayo abordo cuestiones relacionadas con la filosofía teórica; cuestiones óntico-gnoseológicas que caen bajo el amparo de la filosofía de la mente y de las ciencias cognitivas. Aunque en ulteriores análisis sería conveniente tamizar aspectos más pragmáticos, aquí sólo se mencionarán de pasada aquellas tesis generales que sean ineludibles y, por supuesto, aquellos desarrollos tecnológicos que sirvan de soporte para mi argumentación. Con esta forma de tratar el tema respeto el espíritu integrador y multidisciplinar de este ensayo, tal y como se defendió en la introducción.

Una forma adecuada de iniciar la reflexión es distinguir las cuatro grandes formas a través de las cuales la IA puede relacionarse con las emociones: reconociéndolas, simulándolas, manipulándolas o sintiéndolas. Detrás de cada uno de estos cuatro términos se esconden problemas filosóficos de honda raigambre y compleja fundamentación, dificultad potenciada en la medida en que estamos haciendo una traslación desde el ámbito biológico (descrito en el apartado 1) al sintético.

Sobre el reconocimiento y la simulación de emociones descansa un problema tratado por la gnoseología, que está en el origen mismo del debate filosófico sobre las capacidades cognitivas de la IA, y que tiene que ver con lo que se conoce y

comprende al respecto de lo emocional. "Conocimiento" o –hago un giro terminológico que matizaré en breve– "procesamiento informacional" para que la IA pueda computar la información recibida y, así, llegado el caso, ofrecer después la respuesta programada e, incluso, imitar un patrón emocional. Aquí se encuentra, pues, el primer nudo gordiano, clásico ya entre los filósofos analíticos y del lenguaje, que tiene que ver con la diferencia entre procesar información manipulando símbolos mediante una mera computación instalada en el algoritmo cibernético (sintaxis) y comprender verdaderamente el significado simbólico (semántica).

Por su parte, la explicación de qué es sentir una emoción nos conduce inexorablemente hacia un nudo gordiano adicional, que tiene que ver con la emergencia de emociones genuinas (más a allá de su programación y simulación) y la posibilidad de vivencia subjetiva. Esto nos transporta a cuestiones ónticas relacionadas con la naturaleza de la mente y la consciencia (ampliamente debatido por el fisicalismo y el funcionalismo) y nos traslada también hacia una reflexión fenomenológica relacionada con los *qualia* emocionales.

El control de emociones humanas por parte de una IA quizá no planteé tanta dificultad teórica (sí desde luego técnica), aunque se abre un abismo bioético (más bien neuroético) si nos asomamos a las posibles consecuencias bio-psico-sociales de una manipulación de estas características.

Como paso previo a este abordaje filosófico, por el momento meramente apuntado, parece de necesidad el realizar antes una sucinta –aunque suficiente– revisión del concepto tecnológico de IA; igualmente, de los diferentes tipos y principales sistemas operativos de la IA actual y de la que está por venir. Esto ayudará a comprender mucho mejor las dificultades y límites filosóficos que se pretenden explorar y, además, de este modo ratifico el ejercicio de coherencia expuesto en la introducción en relación al espíritu multidisciplinar e integrador del concepto de emocionalidad artificial.

Según la Real Academia Española (RAE), la inteligencia artificial es la "disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico"[6]. Un criterio útil para iniciar la distinción entre los diferentes tipos de IA puede ser el establecido por el filósofo John Searle, que diferencia la IA débil de la IA fuerte (Searle, 1980):

— IA débil: es la que tenemos en la actualidad y que, pese a sus grandes capacidades computacionales, nunca desarrollará una inteligencia general como la humana.

6 Definición de IA según la web de la RAE: https://dle.rae.es/inteligencia (consultado el 21/3/21).

— IA fuerte: es la que critica el propio Searle, por irrealizable y conceptualmente imposible, aunque también es la defendida por los transhumanistas más voluntariosos; es decir, la que tendría capacidades cognitivas iguales a las humanas (lo que muchos consideran como una IA general) o incluso superiores (lo cual supondría el advenimiento de una "singularidad tecnológica" en forma de super-IA).

Con respecto a la IA débil, es decir, aquella de la que ya disponemos, podemos utilizar dos sencillas clasificaciones adicionales, muy útiles para diferenciar el estado actual de la cuestión tecnológica al respecto de la IA. Una sería la propuesta por Rusell y Norvig (2019) y la otra la planteada por López de Mántaras y Meseguer (2017). Según los primeros autores, la IA se clasifica en:

— Sistemas que piensan como humanos: son los que automatizan actividades como la toma de decisiones, el aprendizaje o la resolución de problemas. Un ejemplo serían las "redes neuronales artificiales", capaces de realizar una tarea de forma óptima, como por ejemplo ganar en diferentes juegos como el ajedrez (Deep Blue), las damas (CHINOOK) o el go (AlphaGo)[7].

— Sistemas que actúan como humanos: son los que imitan tareas básicas propiamente humanas. A esta categoría pertenecen las primeras máquinas de la robótica clásica. La empresa Boston Dynamics[8] ofrece un excelente muestrario de algunos de los desarrollos robóticos más impresionantes en la actualidad; los robots espaciales diseñados por la NASA (*National Aeronautics and Space Administration*) para la exploración de Marte (como Sojourner, Spirit, Opportunity o Curiosity) son también un buen ejemplo.

— Sistemas que piensan racionalmente: a esta categoría pertenecen las IA que imitan el pensamiento lógico racional del ser humano. Aquí encontramos a los "sistemas expertos", que tienen cualidades de simulación,

7 Deep Blue es la inteligencia artificial diseñada por IBM (*International Business Machine*) que entre 1996 y 1997 derrotó al ajedrecista ruso Gary Kasparov. CHINOOK es un programa de ordenador basado en IA desarrollado en la Universidad de Alberta. AlphaGo es un programa informático desarrollado por Google DeepMind, basado en el aprendizaje profundo de redes neuronales artificiales, que ganó a varios campeones mundiales al enfrentarse al célebre juego de estrategia de origen chino go. El documental *AlphaGo* (2017) recoge de forma excelente la historia de este hito informático. Para ampliar información al respecto de los grandes éxitos de la IA recomiendo consultar el capítulo 5 del libro *Inteligencia artificial* (López de Mántaras y Meseguer, 2017: 118-144).

8 Web de la empresa donde se muestran sus desarrollos en IA robótica: https://www.bostondynamics.com/ (consultado el 19/3/21).

planificación, control o monitorización. Dendral o Mycin serían dos buenos ejemplos.

— Sistemas que actúan racionalmente: son los sistemas que emulan de forma racional la conducta humana. Aquí estarían los "agentes inteligentes" capaces de interactuar con su entorno. Un ejemplo destacado sería la sonda espacial Deep Space 1.

Por su parte, siguiendo a los científicos españoles, la IA también se puede clasificar atendiendo a diferentes modelos explicativos: el "simbólico", el "conexionista", el de la "computación evolutiva" y el de la "robótica del desarrollo" (López de Mántaras y Meseguer, 2017). Según estos autores, el modelo "simbólico":

> Se basa en el razonamiento lógico y la búsqueda heurística como pilares para la resolución de problemas, sin que el sistema inteligente necesite formar parte de un cuerpo ni estar situado en un entorno real. Es decir, la IA simbólica opera con representaciones abstractas del mundo real que se modelan mediante lenguajes de representación basados principalmente en la lógica matemática y sus extensiones (p. 11).

Esto no quiere decir que lo sistemas de IA simbólicos actuales no hayan evolucionado, pues los incorporados en robots sí que pueden interactuar con el medio para alcanzar objetivos. Por otro lado, estaría la IA "conexionista", inspirada en el funcionamiento del cerebro, la cual:

> Se basa en la hipótesis de que la inteligencia emerge a partir de la actividad distribuida de un gran número de unidades interconectadas que procesan información paralelamente. En la IA conexionista estas unidades son modelos aproximados de la actividad eléctrica de las neuronas biológicas (p. 12).

Este modelo conexionista, cuya principal novedad es la operatividad paralela en vez de secuencial, ha servido para la creación de "redes neuronales artificiales" (como las antes citadas de AlphaGo), mucho más eficaces para facilitar el aprendizaje de los sistemas de IA que la propuesta del modelo simbólico. Los antecedentes de esta aproximación se encuentran en 1943, en un primer estudio en el que se simuló el funcionamiento de una sola neurona artificial (McCullch y Pitts, 1943); o en los descubrimientos de los premiados al Nobel en 1959, David Hubel y Torsten Wiesel, que demostraron el modo jerárquico en el que trabaja el cerebro, lo que facilitó la posterior creación de "redes neuronales convolucionales y profundas", es decir, las que trabajan en capas que procesan diferentes aspectos de la información[9] (citado en Latorre, 2019: 115).

9 Como ejemplo práctico de esto último tenemos el sistema de morfo-reconocimiento, muy utilizado en la actualidad para la identificación de rostros de personas.

Otra aproximación novedosa es la de la "computación evolutiva", la cual imita los procesos de evolución biológica mediante una especie de mecanismo de selección natural (conocido como "algoritmo genético") en la que hay una competencia entre diferentes algoritmos. De una forma similar, las "redes neuronales adversarias" participan en una auténtica carrera de armamentos mediante complejas retroalimentaciones, lo cual ha servido, por ejemplo, para la creación de rostros humanos virtuales (Latorre, 2019).

Por último, tenemos el modelo de la "robótica del desarrollo", el único de los tres que exige necesariamente una corporeidad robótica para su implementación. De hecho, se basa en la interacción de la máquina con el entorno, gracias a distintos sensores capaces de imitar el funcionamiento de los sentidos especiales (como la vista o la audición) y somatosensoriales (como el tacto o la propiocepción) del ser humano, para que reciba inputs externos y construya después una representación interna del mundo. Así, la IA estaría "situada" en su entorno, lo cual facilita su desarrollo cognitivo puesto que, como algunos autores sostienen, "el cuerpo da forma a la inteligencia" (a esto podríamos llamarlo "cognición situada") (López de Mántarás y Meseguer, 2017: 14-15). Tras la captación de estímulos sensoriales, la IA debe ser capaz de procesarlos para, con ello, producir después los *outputs* correspondientes, por ejemplo, en forma de movimientos corporales. A este modelo se le pone el apellido "del desarrollo" porque la mejora cognitiva de este sistema se da cronológicamente, en la medida en que el sistema de IA va interaccionando de forma paulatina con su entorno y aprendiendo de (y con) él. Incluso, puede crearse un algoritmo que seleccione el mejor método de aprendizaje, siendo esto último una especie de, en palabras de Latorre (2019: 113), "meta-aprendizaje". Algo similar a esto se hizo con AlphaGo, programa que se entrenaba con versiones cada vez más mejoradas de sí mismo. En relación a este modelo procesual, otra estrategia similar consiste en el "aprendizaje por refuerzo", que premia a la máquina cuando realiza lo deseado (Latorre, 2019).

Tras esta breve revisión del estado tecnológico de la cuestión[10], en lo que viene a continuación sobre reconocimiento, simulación, manipulación y vivencia de emociones por parte de la IA, diferenciaré siempre entre lo ya realizable tecnológicamente (IA débil) y lo que la especulación científica y el transhumanismo más optimista propone (IA fuerte).

10 Para una mayor profundización de los aspectos técnicos de la IA, aunque con una clara orientación práctica (y muy orientada también hacia temas tratados en este ensayo, como la filosofía de la IA), se recomienda consultar la obra *Inteligencia Artificial* (Boden, 2017).

2.1. IA QUE RECONOCE EMOCIONES

Haciendo una aclaración terminológica, por "reconocer" me refiero a lo que actualmente hace la IA cuando recibe información, es decir, procesar *inputs* gracias a un algoritmo diseñado por un programador. Parto, pues, de los postulados searleanos de la IA débil y de los sistemas capaces de hacer hoy día esta operación, como por ejemplo los sistemas simbólicos. En el extremo opuesto de un eje gnoseológico imaginario estaría el término "conocimiento", o "comprensión", que es lo que hace una inteligencia humana o hará la pretendida IA fuerte[11], bien como IA general (con una inteligencia equivalente a la humana), bien como super-IA (con una inteligencia sobre-humana).

Las críticas de Searle a la IA fuerte, en su celebrado experimento mental de "la habitación china", puede ser aplicado también aquí, al diferenciar entre computar información (en este caso, emocional) y conocer o comprender verdaderamente el significado de las emociones. Lo primero es lo que hace una IA actual (IA débil); lo segundo, considera el pensador americano, nunca será posible. La razón, según él, es que el ordenador manipula símbolos formales, pero no información, al menos en el sentido de información con contenido semántico y significado (Searle, 1980, 1987). O, dicho de otro modo, la sintaxis del algoritmo que ha programado el informático no es suficiente para lograr una verdadera comprensión semántica e intencionalidad que lleve a la máquina a entender las emociones humanas como las entendemos nosotros.

Ha habido autores que han criticado las posturas de Searle. Por ejemplo, Rapaport (1988), con su experimento mental de "la habitación coreana", donde plantea la posibilidad de una sintaxis con cierto valor semántico. Para otros detractores, como Block y Fodor (1972), la clave de una IA fuerte podría estar en la imitación conexionista de nuestro cerebro (que funciona de manera paralela), en vez de utilizar un sistema simbólico. Mejorando esta forma de operar se alcanzarían una serie de ventajas, como la capacidad de almacenar información de forma distribuida o una mayor velocidad de procesamiento y mayor tolerancia a los fallos.

En la actualidad existen sistemas operativos capaces de reconocer matices prosódicos y emotivos del lenguaje hablado (como el algoritmo japonés Empath), o de identificar expresiones corporales vinculadas a la comunicación no verbal (como el *software* Emotient de Apple). Ambos métodos están siendo explotados

11 Newell y Simon (1976: 113-26) defendieron la posibilidad de una verdadera IA fuerte en base a su hipótesis de los Sistemas de Símbolos Físicos (SSF) al reconocer que un SSF "tiene los medios necesarios y suficientes para la acción general inteligente". El cerebro humano y un sistema artificial que lo pueda imitar serían ejemplos de SSF.

por empresas en campos tan diversos como la domótica, los videojuegos, la atención al cliente o la robótica. En este sentido, los algoritmos de la IA son (o podrían ser) capaces de reconocer determinados elementos expresivos externos de las emociones, que encontramos en el lenguaje hablado (determinadas palabras "clave" que pueden funcionar como señales, o ciertas frecuencias sonoras que correlacionan con la prosodia emocional) o en la posición corporal y en la gestualidad facial (las diferentes expresiones gestuales, controladas por los músculos de la expresión facial, que son un correlato externo de ciertos estados emocionales)[12].

Hay, adicionalmente, otra forma potencial de reconocer emociones, virtualmente posible aunque con considerables limitaciones en la actualidad (desde luego técnicas, aunque también conceptuales y por supuesto éticas). Me refiero al empleo de la neurotecnología (en forma de neuroescaneo, electroencefalografía o a través del uso de interfaces cerebro-computador) la cual, operada por una IA, sería capaz de establecer correlatos neurofisiológicos de las emociones (Torres et al., 2020). Sin embargo, para la filósofa Kathinka Evers (2011: 51), "leer" la mente (o las emociones) de este modo no será nunca totalmente eficaz, pues considera que "sería ilegítimo inferir que la cartografía contextual de un proceso neuronal consciente puede o debe suministrar informaciones sobre el contenido de este proceso". La autora sueca considera que un mapa del cerebro no es en ningún caso un mapa del pensamiento (o de un estado mental o emocional particular), debido a razones como la gran variabilidad inter e intraindividual. En este sentido, los correlatos entre circuitos neuronales y pensamientos (incluyo también las emociones y sentimientos) particulares no son iguales en todos nosotros, ni siquiera en un mismo individuo en momentos distintos (pues cada vez que sentimos, pensamos, imaginamos o recordamos algo cambia la circuitería neuronal). Dicho lo cual, aunque hoy día no exista un dispositivo capaz de leer la mente (y las emociones), para los transhumanistas y funcionalistas más transgresores ninguna ley

12 Los sistemas de codificación facial FACS (*Facial Action Coding System*), de Ekman y Friesen, y MAX (*Maximally Discriminative Facial Movement Coding System*), de Izard, son ampliamente utilizados para la caracterización gestual de las emociones, dada la cierta universalidad en la expresión facial de las emociones básicas o primarias. El reconocimiento de cambios fisiológicos periféricos podría ser otra forma a través de la cual un sistema de IA podría ser capaz de obtener información emocional de un humano (aunque esto sería más invasivo) dado que hay ciertos patrones de cambio visceral asociados con estados emocionales susceptibles de ser medidos. El registro de la actividad cardiovascular y respiratoria, de la actividad electrodermal o de la concentración de determinadas hormonas en sangre (como el cortisol, que correlaciona con el estrés), sería una forma de obtener información fisiológico-emocional (Fernández-Abascal et al., 2010). Lo interesante de todo esto es que, aunque una máquina no pueda "comprender" el significado más profundo de las emociones humanas, desde luego podría tener acceso a una ingente cantidad de información emocional, incluso mayor de la que un humano puede extraer de un congénere (sin ayudas artificiales). En todo caso, aunque posibles, todas estas intervenciones carecen de una aplicación directa, clara y útil en la actualidad.

física impediría registrar la conexión de cada neurona de nuestro cerebro (a pesar de la extrema dificultad), pues, al fin y al cabo, siempre que pensamos o sentimos algo se activa un determinado proceso funcional con un sustrato neurobiológico susceptible de ser comprendido (y registrado). Y, si llegásemos a emular artificialmente un cerebro con la ayuda de la IA (a esas copias Hanson [2017: 162] las llama "ems"), nada impediría acceder a la información que contenga.

Para lo que hoy día puede hacer (o presumiblemente hará) una IA con respecto al reconocimiento de emociones, lo único que interesa, desde la categorización emocional antes citada de Kleinginna y Kleinginna (1981), es que las emociones se puedan conceptualizar como entidades fisiológicas y conductuales (en la medida en que la IA reconoce correlatos morfo-fisiológicos externos, como la gestualidad, la dilatación pupilar, la sudoración o el temblor de voz; o internos, como los cambios en el flujo sanguíneo en regiones neuroanatómicas del sistema límbico emocional o en los patrones electrofisiológicos del cerebro) o como entidades expresivas (desde el momento en que dicho correlato morfo-fisiológico comunica y transmite información). Ambas son las condiciones suficientes y necesarias para que una IA actual pueda reconocer de forma básica las emociones (al menos las primarias). Y digo "básica" porque los datos empíricos obtenidos de esta forma no supondrían en ningún caso una información inequívoca del estado emocional que experimenta un sujeto particular. Esto es así debido a varias razones (y limitaciones de la IA): (1) que puede haber convergencia fisiológica y conductual entre varias emociones; (2) que la socialización y la cultura contribuyen a determinar el modo en el que las emociones son expresadas; (3) que otras capacidades mentales y cognitivas, como los deseos, las creencias, los recuerdos o las expectativas podrían influir en el patrón emocional que se evalúa; (4) que puede haber fingimiento emocional; y (5) que una IA pudiera reconocer el "qué" no es lo mismo que reconocer el "cómo". Es decir, comprender verdaderamente el estado afectivo del interlocutor es algo fenoménicamente imposible (en realidad lo es en cada uno de nosotros, cada vez que intentamos comprender los afectos de los demás).

Pese a las dificultades, tal y como se dijo más arriba, el reconocimiento de emociones por parte de la IA puede tener varias aplicaciones. La evaluación afectiva o la detección de mentiras serían algunas de las más controvertidas. En relación a lo primero, podría ser de utilidad para empresas de publicidad[13] o recursos humanos, interesadas en conocer el estado emocional de un cliente o potencial trabajador, respectivamente. Con respecto a lo segundo, su justificación

13 El neuromarketing es una disciplina que utiliza las aportaciones de la neurociencia para la dirección de las organizaciones, el estudio de las necesidades y comportamiento del cliente, el *targeting* y el posicionamiento, los negocios, las estrategias de producto, marca y precios, los canales de ventas, las finanzas o las comunicaciones (Braidot, 2005).

técnica se basa en que es sabido que al mentir se producen perturbaciones emocionales vehiculadas fisiológicamente (Gazzaniga, 2005). Su utilización en juicios o en el reconocimiento de sospechosos es una potencial herramienta no exenta de limitaciones técnico-conceptuales (como ya se ha dicho, correlacionar estados emocionales con otros estados mentales actuales o disposicionales), bioéticas (la violación de la "privacidad mental" y la "libertad cognitiva" [Farah, 2005, Sententia, 2004]) y legales (el quebranto de los "neuroderechos [Ienca y Andorno, 2017]).

2.2. IA QUE SIMULA EMOCIONES

La expresión de emociones se puede programar (Latorre, 2019), bien en forma de gestos o acciones locomotoras, bien a través de la simulación del lenguaje natural. Con respecto a lo primero destaca Kismet, robot pionero desarrollado en el Instituto Tecnológico de Massachusettes (MIT), o la empresa Hanson Robotics[14], que ha diseñado varios androides que simulan expresiones faciales de alegría, tristeza o enfado. En referencia al lenguaje natural, es posible que una IA, como la aplicada a los *chatbots* o a los asistentes inteligentes como Siri o Alexa, sea capaz de mantener una sencilla conversación de contenido aparentemente emocional[15]. Incluso, el programa MINDER, que imita a una enfermera especializada en el cuidado de bebés, es capaz de simular los aspectos funcionales de la ansiedad al interaccionar con los pequeños (Boden, 2017).

Las dificultades searleanas discutidas más arriba vuelven aquí a hacer acto de presencia. Según los postulados de la IA débil, el robot emocional no sabría lo que está simulando, tan sólo se limitaría a ejecutar un programa siguiendo las prescripciones de su algoritmo. Con la denominada "robótica del desarrollo", la IA situada en el mundo (gracias a sensores que le permitan interaccionar con su interlocutor humano), podría aprender a mejorar su simulación emocional, pero nunca llegaría a comprender verdaderamente el objeto de su acción.

Pero, ¿por qué habríamos de querer que una IA virtual o robótica simule emociones? Muchos potenciales clientes de sistemas de IA domésticos se sentirán

14 Web de la empresa: https://www.hansonrobotics.com/ (consultado el 12/3/21).

15 El programa ELIZA, un *bot* conversacional diseñado en el MIT en la década de los sesenta, fue pionero en este sentido. Dicho sistema de IA era capaz de imitar la conversación de un psiquiatra interaccionando con sus pacientes. Su objetivo era demostrar que podía pasar la prueba del test de Turing.

más cómodos y empáticos[16] con una IA, robot o androide humanizado, cuyo aspecto, comportamiento y voz sea casi indistinguible de la de un humano. A dicha mejora empática está contribuyendo enormemente la antropomorfización de la robótica, por ejemplo, gracias a la mayor coordinación motora de los androides (con el bipedismo como uno de sus máximos exponentes) o con la utilización de materiales sintéticos que imitan la piel, las uñas, el cabello o los ojos humanos. Para la mejora de las habilidades locomotoras, la ingeniería robótica ha centrado sus esfuerzos en la imitación anatomo-funcional de nuestros sistemas sensoriales y motores. Los robots más avanzados disponen de todo tipo de sensores, como los inerciales (que imitan nuestra cinestesia y propiocepción) en forma de giroscopios o acelerómetros, o los posicionales, visuales y auditivos, en forma de sistemas láser y radar, cámaras y micrófonos que recogen distancias, imágenes y sonidos. Un excelente ejemplo de la combinación de algunas de estas tecnologías lo encontramos en Atlas, el robot de Boston Dynamics que tiene la increíble habilidad de practicar *parkour*.

Sin embargo, en el polo opuesto estarían los detractores de la humanización robótica (incluidas las emociones); personas que prefieren que la IA tenga un aspecto artificial (por su voz, morfología o movimientos), que en poco recuerde a un ser humano. Esta discrepancia es explicada por la "Teoría del valle inquietante" de Masahiro Mori (1970), la cual señala que nos resulta desagradable que un robot sea demasiado antropomorfo (citato en Latorre, 2019: 154).

En la actualidad existen iniciativas robóticas que participan en importantes relaciones interactivas, como es el cuidado de las personas mayores: tal es el caso del robot Zora[17]. Estas máquinas pueden resultar muy útiles para ofrecer cuidados y paliar la soledad, y en no demasiado tiempo tendrán un papel destacado en la terapia médica. Conforme avancen las habilidades motoras y las capacidades cognitivas de estos dispositivos, no será de extrañar que sus dueños lleguen a tenerles un afecto creciente y verdadero. Un riesgo asociado será el abandono de los mayores por parte de sus familiares bajo la justificación de que el cuidado robótico es más que suficiente. Como aventura el experto en IA José Ignacio Latorre (2019: 155), "una persona mayor morirá dando la mano a un robot".

16 Ha sido ampliamente demostrado que las neuronas espejo son clave en la imitación del otro y en los procesos de empatía. En este sentido, se ha sugerido que son el sistema neurobiológico responsable de, en cierto modo, salvar el "problema de las otras mentes" (cómo acceder y comprender la mente de los otros), facilitando que la intersubjetividad sea posible (Iacoboni, 2009).

17 Web de la empresa: https://www.zorarobotics.be/ (consultado el 2/4/21).

2.3. IA QUE MANIPULA EMOCIONES

La manipulación de emociones por parte de la IA es una posibilidad técnicamente viable (aunque con muchas limitaciones) que podría aparecer en el futuro. Los interfaces cerebro-computador tal vez sean la forma de materializarlo.

El control cibernético-emocional se explicaría por el hecho de que las emociones tienen una base neurobiológica cada vez mejor conocida: desde la neuroquímica a la neuroanatomía, la evidencia científica que correlaciona la vivencia subjetiva emocional con su base cerebral es cada vez más fuerte (Esperidiao-Antonio et al., 2017; Fernández-Abascal et al., 2010). De hecho, en la actualidad existen múltiples aproximaciones en la terapia neuro-psiquiátrica para el tratamiento de patologías que cursan con desórdenes emocionales, como por ejemplo la ansiedad, la depresión o la esquizofrenia. A los tratamientos farmacológicos habría que incluir otras herramientas intervencionistas como la neurocirugía o el uso de dispositivos neuromoduladores (como la estimulación magnética transcraneal [EMT], la estimulación eléctrica transcraneal [EET], la terapia electroconvulsiva [TEC] o la estimulación cerebral profunda [ECP]).

Desde el punto de vista filosófico hay dos grandes núcleos de reflexión: partiendo de la filosofía de la mente, el debate sobre la posibilidad de manipular el contenido, la intensidad o la valencia (positiva-negativa) emocional. Desde la bioética (más en concreto, desde la neuroética aplicada), la preocupación por los límites de una intervención terapéutica o de neuromejora que podría interferir en aspectos clave de la condición humana como la identidad o el libre albedrío (Biscaia, 2021).

Con respecto a lo primero recordemos que, en el primer apartado dedicado a las emociones biológicas, se indicó que éstas compartían algunas cualidades con los estados mentales, como por ejemplo la intencionalidad, en el sentido de que las emociones instancian representaciones internas (pues se relacionan con estados mentales), frecuentemente se dirigen hacia un objeto concreto (se ama algo o a alguien en particular) y presentan contenido (aunque, según Broncano [2001], no conceptual). A la pregunta de si se podría manipular la aparición, contenido, direccionalidad y grado de dicha emoción, por supuesto no hay respuesta neurobiológica totalmente satisfactoria (por más que algunos neurofármacos y neurocirugías puedan lograr algo de esto), aunque ninguna ley física impide que dicho escenario pudiera presentarse, al menos de forma parcial. Sentir una emoción y no otra, dirigida hacia un objeto y no otro, en un mayor o menor grado es, nomológicamente hablando, susceptible de ser manipulado, aunque la complejidad de los fenómenos cerebrales que acontecen lo hace difícilmente regulable. Además, en la medida en que lo emocional es muchas veces inseparable de lo cognitivo, la supuesta manipulación debería intervenir igualmente otras áreas

neuroanatómicas y circuitos neuronales responsables de dichos procesos cerebrales superiores. Por último, no se debe olvidar la importancia de determinadas claves externas que sirven como disparadores de las emociones, además de las experiencias pasadas, de las expectativas futuras o del contexto socio-cultural, que también formarían parte de la ecuación emocional.

Si bien estas últimas cuestiones filosóficas son de indudable interés, no se insistirá más en ellas por no ser ámbito de la filosofía de la IA de la que trata este ensayo sino, más bien, del debate transhumanista (neuro)tecnológico y de la neuroética aplicada y fundamental[18].

2.4. IA QUE SIENTE EMOCIONES

He dejado para el final el punto de mayor complejidad; también el más controvertido desde casi cualquier punto de vista tecno-científico y filosófico que se considere. Me refiero a la posibilidad de que, algún día, una IA fuerte pueda sentir emociones de forma genuina, al modo de como lo hacen los seres biológicos.

Para un mejor tratamiento de este último apartado dividiré el argumentario en cuatro secciones que se construyen desde las siguientes cuatro preguntas de partida: (2.4.1.) ¿es conceptualmente posible que una IA tenga estados mentales y emocionales? (2.4.2.) ¿Podría saber ella, también nosotros, si los está experimentando? (2.4.3.) ¿Sería de utilidad que una IA sienta emociones? (2.4.4.) ¿Cuál sería el impacto social de una emocionalidad robótica emergente?

2.4.1. *Emergentismo emocional en la IA*

El debate en torno a la posibilidad de la emergencia de emociones en un ser artificial nos retrotrae a la clásica discusión sobre la naturaleza de los procesos mentales y al problema mente-cerebro. Recordemos que ya en el primer apartado concluimos con Broncano (2001: 1) que las emociones eran "territorios intermedios en la mente", es decir, fenómenos con ciertas características propias de los estados mentales, por lo que el debate de lo mental podrá extrapolarse, con matizaciones, al territorio emocional. Dicho debate propio de la filosofía de la mente, de la gnoseología y de la filosofía del lenguaje gira en torno a cuestiones

18 Para una lectura de algunos de estos temas se recomiendan las obras *Transhumanismo. La búsqueda tecnológica del mejoramiento humano* (Diéguez, 2017) y *Neuroética. Retos para el siglo XXI* (Levy, 2014), o acudir a la revisión sobre neurotecnología y neuromejora de Biscaia (2021).

óntico-epistémicas al respecto de la existencia, naturaleza y conocimiento de la mente y de los estados mentales. En líneas generales, a fin de poder seguir mi argumentación, resumo sucintamente las principales posturas al respecto, siguiendo lo recogido por los tratados de Hierro-Pescador (2005) y Martínez-Freire (2002).

A groso modo, hay posturas que niegan de forma radical la existencia de lo mental: bajo este ideario encontramos el materialismo o monismo ontológico de algunos fisicalistas, especialmente aquellos consideramos como eliminacionistas por reconocer que únicamente existe lo material (el cerebro)[19]; igualmente, se situaría aquí cierto tipo de conductismo radical. Además, hay posturas que niegan la sustanciación de lo mental, aunque le reconocen una cierta existencia lingüística. Por su parte, hay posiciones que aceptan la sustanciación ontológica o funcional de lo mental: independientemente de lo físico (con o sin leyes que unen al conjunto mente-cerebro)[20]; aunque, también, vinculado nomológicamente con lo físico (una suerte de epifenomenalismo de lo mental desde lo cerebral). Finalmente, el funcionalismo considera que los estados mentales son estados funcionales cuyo órgano es el cerebro (aunque para algunos funcionalistas el sustrato material podría ser cualquier otro, distinto del tejido nervioso).

Algunos de los fisicalistas más destacados, cada cual con sus propias matizaciones al respecto del tema que nos ocupa, son: Carnap (1931, 1932), el cual reduce el lenguaje psicológico al físico, pues la realidad material es la única existente y lo mental es, en todo caso, algo disposicional; o Rorty (1982), quien niega lo mental, admitiendo que la mente no es siquiera un estado natural; o Feigl (1958), cuyo monismo ontológico descansa, hay que admitirlo, en un cierto dualismo epistemológico sobre el modo de conocer lo psicológico-cerebral, que culmina en su "Teoría del doble acceso"; o Smart (1959), quien identifica estados mentales con estados físicos por la dificultad de engranar leyes psicofísicas que unan los procesos de ambos estados; o Lewis (1994), para quien los estados mentales son idénticos a los estados neuronales. Por su parte, el conductismo ontológico (con John Watson como principal representante) únicamente reconoce a la conducta (derivada de la interacción estímulo-respuesta) como digna de estudio, puesto que niega la existencia de estados mentales[21].

19 En el fisicalismo reduccionista lo mental se reduce a lo cerebral, produciéndose una identificación sustancial (que no descriptivo-lingüística) mente-cerebro.

20 Por ejemplo, el dualismo cartesiano o el dualismo interaccionista del fisiólogo John Eccles y del defensor de la "Teoría de los tres mundos", el filósofo Karl Popper. Para una revisión de sus ideas consultar la obra *El yo y su cerebro* (Barcelona: Labor Sa, 1993).

21 El conductismo metodológico (con Burrhus Skinner como mayor exponente), a diferencia del radical, no niega de forma tajante los estados mentales, pero reconoce que son inobservables y que por tanto no se pueden estudiar de forma directa.

El conductismo lógico de Ryle (1967) considera que el dualismo cartesiano es un error lógico y categorial. Además, defiende que los procesos mentales no son actuales sino disposicionales para la conducta. En una línea similar, Ludwig Wittgenstein reconoce que el problema de la mente es un problema lingüístico[22]. Por su parte, el monismo anómalo o fisicismo de ejemplares de Davidson (1980) plantea que habría dos tipos de descripciones –física y mental– para la misma sustancia, aunque no hay leyes correlacionales entre lo cerebral y lo mental.

Entre los más reconocimos funcionalistas se encontraría Putnam (1960), con su "Teoría del autómata probabilístico" y su analogía con los estados de máquina de Turing, donde hace una comparación cerebro-*hardware* y mente-*software*. El filósofo estadounidense sostiene que los estados mentales no son estados del cerebro ni disposiciones para la conducta, sino estados funcionales que operan según una cadena causal; o Fodor (1968, 1981), quien propone una "Teoría representacional de la mental", donde habría un lenguaje del pensamiento formado por representaciones mentales que se relacionan funcional y computacionalmente de forma causal. Este autor afirma que no deben identificarse estados mentales (lo funcional) y cerebrales (lo estructural), pues operan a diferentes niveles (se refiere a ello como macropropiedades y micropropiedades, respectivamente). Por su parte, el emergentismo de Searle (1987) sostiene que los procesos mentales no son independientes de los físicos, aunque no se reducen a los procesos cerebrales. Serían procesos o propiedades que emergen desde lo cerebral, manteniendo entre ellos una relación de implicación lógica. Lo mental (sea algo meramente físico o no) causaría la conducta, y las macropropiedades psicológicas se realizarían gracias a las micropropiedades del soporte material.

Tras este sucinto repaso de las principales posturas (a riesgo de que, por la brevedad expositiva, se puedan hacer de forma legítima cualesquiera matizaciones y ampliaciones), se puede concluir que las tesis funcionalistas son las que mejor pueden soportar la emergencia de propiedades mentales (y, de ahí, también de estados emocionales) en una IA. Dicho funcionalismo propone que lo importante para la aparición de estados mentales son las estados y relaciones funcionales, no el soporte material que sustenta lo mental (así, una máquina tiene estados lógicos (*software*) y estructurales (*hardware*), como un humano tiene mente y cerebro, respectivamente). En este sentido, una máquina podría perfectamente ser capaz de poseer estados mentales, toda vez que, al menos (y no es poca cosa), imite los estados funcionales de un cerebro. Desde luego, es necesario un soporte material para ello, pero éste, a diferencia de lo que sostienen algunas tesis fisicalistas radicales, no tiene porqué ser necesariamente un cerebro, sino cualquier otro

22 Para un adecuado acercamiento a dicha postura consultar el capítulo 4 del tratado de filosofía de la mente de Hierro-Pescador (2005).

material. Por tanto, según el funcionalismo, si una IA tuviera una configuración causal-funcional semejante a la de un cerebro, nada impediría que pensase (y sintiese). Aunque, desde luego, ninguna ley física impide que otros posibles estados funcionales distintos a los cerebrales, que ni siquiera conocemos, puedan generar emergencia o superveniencia mental[23].

2.4.2. *Test de Voight-Kampff*

El filósofo Thomas Nagel (1974) publicó un artículo titulado *¿Cómo es ser un murciélago?* que, para el caso, puede ser de utilidad. En él, se plantea la dificultad epistemológica de conocer la vivencia subjetiva de los estados mentales del otro. En el núcleo de su debate se encuentra la clásica discusión al respecto de los *qualia* (el modo como algo se experimenta) y, también, del "problema de las otras mentes" (cómo acceder –y comprender– a la mente de los otros). Chalmers (1995) ha querido referirse a esta dificultad explicativa sobre "cómo es ser uno" como el "problema difícil de la conciencia".

Parece claro que para que una máquina experimente emociones subjetivamente debe poseer algún tipo de conciencia. Al menos, como mínimo, equivalente a la de un animal. Block (1994a, 1994b y 1995) diferencia varios tipos de conciencia: fenoménica (sería la experiencia subjetiva), de acceso (en la que hay representación de un contenido representacional para el control racional de la acción o el habla), monitora (responsable de la introspección y de los pensamientos de alto nivel sobre un estado mental) y conciencia del yo. Para que una IA experimentase emociones básicas se exigiría, de entre todos estos tipos (como mínimo), que tuviera conciencia de tipo fenoménica[24], pues es la que mejor correspondencia establece con la categorización subjetivo-afectiva de las emociones descrita en el apartado 1. Lo que parece evidente es que, de ser posible la emergencia o superveniencia emocional (de forma intencional, mediante diseño y programación o, como ha especulado la ciencia-ficción, de forma "espontanea", al modo de otros tantos cambios morfo-funcionales surgidos en el mundo vivo durante el transcurso de la evolución), posiblemente surgiría de forma paulatina: en primer lugar podrían manifestarse emociones básicas o primarias (como el miedo o la alegría),

23 La superveniencia es un concepto ampliamente utilizado por Davidson (1973), entre otros autores, en el marco de su "Teoría psicofísica de la identidad", según el cual no es posible que haya dos eventos iguales en los aspectos físicos pero que difieran en algún aspecto mental.

24 Se conocen casos de conciencia fenoménica sin otros tipos de conciencia, como la de acceso. Lo vemos, por ejemplo, en la prosopagnosia (tipo de agnosia visual en la que el sujeto es incapaz de reconocer rostros familiares; a veces, incluso, es incapaz de reconocerse así mismo).

que son las más sencillas (en el caso biológico, surgieron pronto en la filogenia y tienen una extensión universal entre los vertebrados). Y quizá, más adelante, podrían aparecer emociones secundarias o "autoconscientes" (como la vergüenza o el orgullo), aunque éstas exigirían, como su nombre indica, que se organizaran alrededor de la unidad de conciencia que confiere un "yo".

El título de este apartado, "Test de Voight-Kampff", hace mención a una prueba ficticia que aparece en la película de ciencia-ficción *Blade Runner* (1982), ideada para la detección de replicantes (una suerte de seres artificiales). Dicho test, al modo del famoso test de Turing, era utilizado para saber si una máquina poseía emociones, al basarse en la detección de respuestas fisiológicas y conductuales asociadas. El problema aquí es que las experiencias subjetivas son fenoménicamente intransferibles, y sólo podríamos inferir estados emocionales genuinos en una IA mediante conductas o respuestas correlacionales (serían como "criterios"[25] de la presencia de dichos estados), analogías e inferencias. No se quiere decir con esto que sea nomológicamente imposible que una super-IA experimente emociones, sino que desde un punto de vista gnoseológico no tendríamos forma de estar seguros. Aunque esto no debe ser mayor impedimento epistemológico que el que, de facto, se produce con respecto al conocimiento de los estados emocionales de otros seres vivos no humanos o, incluso, de nuestros propios congéneres, pues, al fin y al cabo, el núcleo del "problema de las otras mentes"[26] bien se puede aplicar a la cuestión que ahora nos ocupa: en su vertiente puramente epistemológica, el acceso a nuestros estados mentales (emocionales) es radicalmente diferente del acceso a los estados de los demás; en su vertiente conceptual, la dificultad estriba en cómo es posible formar un concepto de los estados mentales de los otros. En definitiva, podemos concluir que no sería posible tener un conocimiento directo de los estados mentales de cualesquiera otras mentes, sean animales o robóticas, aunque esto no impide tener un cierto grado de conocimiento (indirecto) sobre, al menos, la existencia y modo de operar de tales estados.

A pesar de las dificultades (para muchos) aparentemente insalvables en relación al desarrollo de capacidades cognitivas y emocionales similares a las humanas, algunos científicos y pensadores creen que en el futuro será posible alcanzar lo que se conoce, en palabras de Vernor Vinge (1993), como la

25 Sin duda, el filósofo Ludwig Wittgenstein es quien ha hecho las mayores aportaciones con respecto al concepto de "criterio". No obstante, para muchos pensadores, el emplear los criterios como prueba de la existencia de estados mentales es totalmente deficiente, dado el abismo explicativo entre la conducta y los estados internos que la sustentan. Sin conexión conceptual y en ausencia de inferencias inductivas, parece difícil sostener con pleno derecho esta conexión (McDowell, 1982).

26 La *Stanford Encyclopedia of Philosophy* (2019) ofrece una magnífica exposición en relación al concepto de las "otras mentes" (https://plato.stanford.edu/entries/other-minds/) (consultado el 17/7/21).

"singularidad tecnológica". Este nuevo estadio supondría, en su versión débil, la aparición de una IA general equivalente a la humana (Tegmark [2018] y otros expertos lo consideran altamente posible para mediados del presente siglo) y, en su versión fuerte, significaría el advenimiento de una super-IA con capacidades cognitivas superiores a las nuestras (Kurzweil, 2012). Para justificar la llegada de dicha singularidad, el autor se basa en la conocida como "Ley de Moore" (1965), que él adapta en forma de su "Ley de rendimientos acelerados": ambas, en todo caso, se basan en el crecimiento exponencial de la tecnología cibernética que se observó en los primeros años de la era informática. No obstante, para superar el estancamiento que parece haberse producido en los últimos tiempos (Waldrop, 2016), el científico sostiene que se tendrá que dar un cambio de paradigma tecnológico una vez alcanzada la actual fase de meseta (Kurzweil, 2012), lo cual podría llegar cuando el primer sistema super-inteligente fuese capaz de perfeccionarse a sí mismo y/o de fabricar otras IA iguales. En esta línea, la utilización de nuevos lenguajes lógicos (lógica trivalente, intuicionista y de la vaguedad) que fuesen capaces de programar el "sentido común", el empleo de sistemas de IA conexionista (que operen de forma paralela y distribuida) o la utilización de nuevos algoritmos de aprendizaje podría acelerar el progreso[27]. Y, sobre todo, la llegada de la computación cuántica podría ser ese esperado hito que esperan los singularistas más voluntariosos: aunque todavía incipiente, este sistema trabaja con "qubit" (del término inglés *quantum bit*), lo que contempla la superposición de estados cuánticos (1, 0, o los dos a la vez). Además, a este elemento de complejidad se le suma el entrelazamiento cuántico, lo que multiplica exponencialmente (millones de veces) la capacidad de computación frente a cualquier ordenador actual (Diéguez, 2017). En definitiva, aunque muchos científicos ortodoxos prefieren no hablar de singularidad tecnológica, pues lo consideran improbable dadas las actuales dificultades técnicas (Stone y Hirsh, 2006)[28], el experto nacional en transhumanismo, Antonio Diéguez (2017: 77), considera que "no hay razones irrefutables para pensar que la creación de una super-inteligencia artificial no es ni será jamás posible". Por su parte, el experto en IA Max Tegmark (2018) indica que no hay leyes físicas que impidan el surgimiento de una super-IA, y plantea un escenario al que deno-

27 Un interesante proyecto de conciencia artificial es LIDA (*Learning Intelligent Distribution Agent*). Es, a su vez, en palabras de Boden (2017: 123): (1) "un modelo conceptual (una teoría computacional expresada verbalmente) de conciencia (funcional)" y (2) "una implementación parcial y simplificada de ese modelo". LIDA funcionaría de manera distribuida, con varios subsistemas trabajando en paralelo y de forma jerarquizada (lo que se conoce como *Global Work-space Theory*), es decir, imita el funcionamiento del cerebro según la "Teoría neuropsicológica" desarrollada por Bernard Baars (1988).

28 Así lo señala un informe de la Universidad de Stanford, titulado *Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence* (2016). https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai_100_report_0831fnl.pdf (consultado el 1/2/2021).

mina "3.0", con máquinas cuyo *hardware* sería auto-replicativo y cuyo *software* produciría mejoras recursivas.

2.4.3. *Mente caliente VS mente fría*

La película de ciencia-ficción *Morgan* (2016) plantea la pertinencia de crear una IA general con auténtica capacidad emocional. Para algunos autores, desarrollar una "mente caliente", es decir, programar un algoritmo emocional, podría ser útil y necesario para desarrollar una conducta ética en la IA (Latorre, 2019). En este sentido, tal y como se comentó más arriba, ya existe un precedente tecnológico (el programa en inteligencia artificial MINDER) capaz de regular sus respuestas en el cuidado de bebés en base a la simulación (funcional) de estados de ansiedad. La idea subyacente es que la programación de emociones también podría ser ventajosa en la toma de decisiones (pues, como se discutió ampliamente en el primer capítulo, existe un continuo cognitivo-emocional). La película *Morgan* (2016), no obstante, nos señala los riesgos asociados a esta posibilidad: la aparición de emociones y sentimientos negativos, como el miedo o la hostilidad, que podrían volverse contra sus creadores.

Siguiendo la "Teoría funcional del sistema emotivo" de Oakley y Jonhson-Laird (1987), las emociones ayudan a evaluar aspectos relevantes del mundo (interno o externo) de forma rápida y global, para que la respuesta que se requiere no esté limitada por la lentitud de procesos cognitivos que exigen la participación de la conciencia y la memoria. Las emociones activan un plan de acción rápido y eficaz orientado a la deliberación, aunque, también, un plan menos discriminador que el activado por otros módulos cognitivos. Así pues, en un primer momento podría parecer útil, tal y como se acaba de indicar, que un sistema de IA (débil) lo poseyera. Sin embargo, si se considera que, por defecto, la cognición de una super-IA se presume hiper-rápida y compleja, quizá este argumento no sea del todo convincente en el escenario hipotético de dicha singularidad tecnológica.

Por su parte, una "mente fría" (únicamente con capacidad analítica y juicio racional) tampoco está exenta de riesgos, tal y como, por ejemplo, plantean obras cinematográficas como *Terminator* (1984) o *Yo, robot* (2004), donde la ausencia de "color emocional" en las máquinas puede provocar que la elección más óptima según la lógica de un algoritmo no sea la más conveniente según las normas éticas, los convencionalismos sociales o el sentido común[29]. En concordancia con

29 Aunque esto puede ser discutido: el libro *La sabiduría de los psicópatas* de Kevin Dutton (2020) afirma que algunas personas que tienen trastorno antisocial de la personalidad y psicopatía

esta idea, los científicos españoles López de Mántaras y Meseguer (2017: 149) opinan al respecto del sentido común que "es requisito fundamental para conseguir IA similar a la humana en cuanto a generalidad y profundidad".

2.4.4. *Mundo Silico Sapiens*

Esta última sección será meramente apuntada, con la optimista idea de desarrollarla de forma más detenida en sendas reflexiones. Pues la introducción de IAs genuinamente afectivas (robóticas o no) en nuestro mundo tendrá un considerable impacto en diferentes planos de la praxis humana, como por ejemplo la convivencia, la cultura, el derecho, el arte, la economía o la política. Plantear un escenario en el que convivan diferentes seres inteligentes, sensitivos y conscientes nos arroja hacia un abismo reflexivo, similar al que tendríamos con el conocimiento de otros seres inteligentes extraterrestres[30]. Así pues, dejo abiertas una serie de preguntas generales con el objetivo de invitar a la especulación y el propósito de profundizar en su análisis en posteriores investigaciones: ¿podríamos considerar "vivo" a un ser artificial dotado de emociones e inteligencia; qué características ayudarían a dicha adscripción? ¿Tendría una IA general capacidad de libre albedrío y responsabilidad moral; cuál sería su posicionamiento ético en el mundo? ¿Qué estatus jurídico-legal tendría un ser vivo artificial, sintiente y dotado de inteligencia? ¿Cuáles serían las ocupaciones y tareas de dicha inteligencia robótica; cómo se integraría en el sustrato económico y productivo de los estados? ¿Cómo serían las relaciones jurídico-administrativas entre máquinas inteligentes-sintientes y humanos? ¿Qué tipo de relaciones afectivas y convivenciales se establecerían entre humanos y máquinas emocionales e inteligentes? ¿Qué peligros o amenazas deberíamos considerar, y qué medidas sería conveniente adoptar frente a una super-IA?

(caracterizados por empobrecimiento emocional y falta de empatía; diríamos, pues, que son "mentes frías"), tienen una alta capacidad racional dirigida a metas y, por tanto, no es infrecuente que alcancen el éxito profesional.

30 Hay que señalar que este escenario ya se produjo, salvando las distancias, en nuestro pasado evolutivo, con una situación de convivencia (no está claro si pacífica) entre dos especies inteligentes, como el *Homo sapiens* que somos y el extinto *Homo neanderthalensis* (si bien, en cierto modo, somos también Neandertales, en la medida en que compartimos acervo genético con ellos). La extensa y magnífica obra divulgadora del paleontólogo Juan Luis Arsuaga ha profundizado en muchos de estos aspectos.

CONCLUSIONES

El correcto abordaje de la interacción IA-humano exige aproximarse desde la perspectiva de tres grandes áreas de conocimiento dentro del campo de las ciencias y las humanidades: (1) la psicología y la neurociencia; (2) la robótica y las ciencias de la computación; y (3) la filosofía teórica (sobre todo la filosofía de la mente) y práctica (especialmente la ética aplicada). Sólo así podrá alcanzarse un conocimiento profundo e integrado del complejo fenómeno de la emocionalidad artificial.

Como se ha podido comprobar, no son pocos los retos a los que se enfrentarán los futuros desarrollos de la inteligencia artificial en el ámbito afectivo. Su interactividad humana tendrá un papel muy destacado en el reconocimiento y simulación emocional, especialmente en los robots de compañía con funciones socio-sanitarias o en cualesquiera sectores en los que el empleo de programas de detección de expresión emocional (fisiológica o conductual) suponga algún tipo de beneficio.

Aunque por el momento sean meras propuestas vanguardistas, los interfaces cerebro-ordenador regulados por IA también tendrán, presumiblemente, un destacado papel en la manipulación terapéutica de los afectos, más en concreto, en el tratamiento de trastornos neuropsiquiátricos que cursan con alteraciones emocionales como la depresión, la ansiedad o la esquizofrenia.

Pese a la problemática tecnológica y la controversia conceptual apuntada, el advenimiento de una IA general o, incluso, de una super-IA con conciencia y vivencia emocional genuina podría, a juicio del transhumanismo más optimista, ser una realidad en un futuro a medio o largo plazo. Desde un punto de vista tecnológico, son varias las alternativas propuestas para lograr algunos de estos hitos, como el desarrollo de nuevos lenguajes lógicos y algoritmos de aprendizaje, de sistemas integrados y arquitecturas cognitivas o de la prometedora computación cuántica. Desentrañar los misterios de la estructura y funcionamiento cerebral sin duda ayudará en esta progresión tecnológica.

Pero no sólo hay retos científico-tecnológicos, pues desde la filosofía de la mente y de las ciencias cognitivas el transhumanismo más transgresor seguirá debatiendo acaloradamente sobre el núcleo ontológico y gnoseológico de la IA, al respecto de la posibilidad de estados mentales artificiales, de la creación de una inteligencia general con "conciencia artificial" o de la emergencia de genuinos *qualia* emocionales. Posiblemente el funcionalismo sea la corriente filosófica que mejor pueda sustentar este conjunto de controvertidas propuestas.

A todas las dificultades comentadas se une, sin duda, cierta urgencia por establecer una vigilancia que limite los riesgos bioéticos y sociales inherentes a una

tecnología tan sobrecogedora, capaz de modificar nuestro mundo de un modo que superará con creces los cambios tecno-sociales de la Revolución Industrial. Se hace imprescindible, pues, un desarrollo responsable y sostenible de dichas tecnologías, con sistemas de control garantistas, habida cuenta del impacto bio-psico-social que la llegada de la emocionalidad artificial presumiblemente tendrá en las próximas décadas. Afortunadamente, contamos ya en nuestro entorno inmediato con iniciativas reguladoras, como las Directrices éticas de la Comisión Europea aprobadas en junio de 2018, que proponen una serie de normas básicas que pivotan en torno al respeto de la autonomía humana, la prevención del daño, la equidad y la explicabilidad[31].

REFERENCIAS BIBLIOGRÁFICAS

APARICIO-GARCÍA, R.; JUÁREZ-GRACIA, G.; ÁLVAREZ-CEDILLO, J.; SANDOVAL-GUTIÉRREZ, J y TOVAR-CORONA, B., "Evaluation of the Design of a Brain-Computer Interface for Emotion Detection". *DYNA*, 93(5), 2019, 468.

BAARS, MJ., *A Cognitive Theory of Conscience*. Cambridge: Cambridge University Press, 1988.

BISCAIA, JM., "Neuromejora: de la vanguardia científica y tecnológica a las dificultades y límites planteados por la filosofía de la mente y la bioética". *Revista Iberoamericana de Bioética*, 16, 2021, 1-17. doi: 10.14422/rib.i16.y2021.003.

BLAKEMORE, C. y GREENFIELD, S. (ed.), *Mindwaves*. Oxford: Blackwell, 1987.

BLOCK, N., "Consciousness". En: GUTTENPLAN, S. (Ed.), *A Companion to the Philosophy of Mind*. Oxford: Blackwell, 1994a.

—, "Qualia", En: GUTTENPLAN, S. (ed.), *A Companion to the Philosophy of Mind*. Oxford: Blackwell, 1994b.

—, "On a Confusion about a Function of Consciousness". *Behavioural and Bran Sciences*, 18, 1995.

BLOCK, N. y FODOR, JA., "What Psychological States are not". *Philosophical Review*, 81, 1972 159-181.

BODEN, MA., *Inteligencia Artificial.* Madrid: Turnes Publicaciones, 2017.

BRAIDOT, N., *Neuromarketing, neuroeconomía y negocios*. Madrid: Editorial puerto norte-sur, 2005.

BRONCANO, F., "Las emociones: territorios intermedios en la mente". *Contrastes. Revista internacional de filosofía*, suplemento VI, 2001.

31 Normativa europea: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (consultado el 19/7/21).

CANO-VINDEL, A., "Orientaciones en el estudio de la emoción". En FERNÁNDEZ-ABASCAL, EG. (ed.), *Manual de Motivación y Emoción*. Madrid: Centro de Estudios Ramón Areces, 1995, 337-383.

CARNAP, R., "Die Physikalische Sprache als Universalsprache der Wissenschaft". *Erkenntnis*, 1931-1932.

CASACUBERTA, D. y VALLVERDÚ, J., "Emociones sintéticas". *Páginas de Filosofía*, XI, 13, 2010, 116-144.

CASADO, C. Y COLOMO, R., "Un breve recorrido por la concepción de las emociones en la filosofía occidental". *A parte Rei. Revista de Filosofía*, 47, 2006, 1-10.

CHAKRISWARAN, P.; VINCENT, D.; SRINIVASAN, K.; SHARMA, V.; CHANG, C-Y. y GUITIÉRREZ, D., "Emotion AI-Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues". *Appl. Sci*, 9(24), 2019, 5462. doi: org/10.3390/app9245462.

CHALMERS, D., "Facing up to the Problem of Consciousness". *Journal of Consciousness Studies*, 2, 1995, 200-219.

CORTINA, A., *Neuroética y neuropolítica. Sugerencias para la educación moral*. Madrid: Tecnos, 2011.

DAMASIO, AR., *Descartes'error. Emotion, reason, and the human brain*. Nueva York: Grosset/Putman Book, 1994.

DAVIDSON, D., "Material Mind". *Studies in Logic and the Foundations of Mathematics*, 74, 1973.

—, Ensayos sobre acciones y sucesos. Barcelona: Crítica, 1980.

DIÉGUEZ, A., *Transhumanismo. La búsqueda tecnológica del mejoramiento humano*. Barcelona: Herder Editorial, 2017.

DIXON, ML.; THIRUCHSELVAM, R.; TODD, R.; CHRISTOFF, K., "Emotion and the prefrontal cortex: An integrative review". *Psychology Bulletin*, 143(10), 2017, 1033-1081. doi: 10.1037/bul0000096.

DUTTON, K., *La sabiduría de los psicópatas*. Madrid: Ariel, 2020.

EKMAN, P., "An argument for basic emotions". *Cognition and Emotions*, 6(3-4), 1992, 169-200.

ELSTER, J., *Alchemies of the Mind. Rationality and the Emotions*. Cambridge: Cambridge University Press, 1999.

ESPERIDIAO-ANTONIO, V.; MAJESKI-COLOMBO, M.; TOLEDO-MONTEVERDE, D.; MORAES-MARTINS, G.; FERNANDES, JJ.; BAUCHIGLIONI DE ASIS, M., MONTENEGRO, S., SIQUEIRA-BATISTA, R. "Neurobiology of emotions: an update". *International Review Psychiatry*, 29(3), 2017, 293-307.

EVERS, K., *Neuroética. Cuando la materia se despierta*. Madrid: Katz Editores, 2011.

FARAH, MJ., "Neuroethics: The Practical and the Philosophical". *Trends in Cognitive Sciences*, 9, 2005, 34-40. doi: 10.1016/j.tics.2004.12.001.

FEIGL, H.; MAXWELL, G. y SCRIVEN, M. (Ed.), *Concept, Theories and Mind-Body Problem.* Minneapolis: University of Minnesota Press, 1958.

FERNÁNDEZ-ABASCAL, EG.; GARCÍA, B.; JIMÉNEZ, MP.; MARTÍN, MD. y DOMÍNGUEZ FJ., *Psicología de la Emoción*. Madrid: Editorial Universitaria Ramón Areces, 2010.

FODOR, J., *Psychological Explanation*. Nueva York: Random House, 1968.

—, *Representations.* Brighton: The Harvester Press, 1981.

GAZZANIGA, MS., *The Ethical Brain*. Nueva York: Dana Press, 2005.

GIBBARD, A., *Wise Choices, Apt Feelings. A Theory of Normative Judgement*. Oxford: Clarendon, 1990.

GOLEMAN, D., *Inteligencia emocional*. Madrid: Kairos, 1995.

GONZÁLEZ, M. y GARCÍA, J., "Arquitecturas emocionales en inteligencia artificial: una propuesta unificadora". *Teoría de la Educación. Educación y Cultura en la Sociedad de la Información*, 7(2), 2006, 156-168.

GRIFFITHSS, P., *What Emotions Really Are*. Chicago: The University of Chicago Press, 1977.

GUTTENPLAN, S., *A Companion to the Philosophy of Mind*. Oxford: Blackwell, 1994.

HANSON, R., (2017). "Cuando los robots gobiernen la Tierra: el legado humano". En *El próximo paso: la vida exponencial*. OpenMind BBVA, 2017. doi: 10.1080/0952813X.2015.1055826.

HIERRO-PESCADOR, J., *Filosofía de la mente y de la Ciencia cognitiva*. Madrid: Akal, 2005.

HOOK, S. (Ed.), *Dimension of Mind*. Nueva York: Collier Books, 1960.

IACOBONI, M., "Imitation, empathy and mirror neurons". *Annu Rev Psychol*, 60, 2009, 653-670. doi: 10.1146/annurev.psych.60.110707.163604.

IENCA, M. y ANDORNO, R., "Towards new human rights in the age of neuroscience and neurotechnology". *Life Sciences, Society and Policy*, 13, 2017, 5. doi: 10.1186/s40504-017-0050-1.

JANAK, PH. Y TYE, KM., "From circuits to behaviour in the amygdala". *Nature*, 517(7534), 2015, 284-292. doi: 10.1038/nature14188.

KLEINGINNA, PR. y KLEINGINNA, AM., "A Categorized List of Emotion Definitions, with Suggestions for a Consensual Definition". *Motivation and Emotion*, 5, 1981, 345-379.

KURZWEL, R., *La singularidad está cerca. Cuando los humanos trascendamos la biología.* Berlín: Lola Books, 2012.

LATORRE, JI., *Ética para máquinas*. Barcelona: Ariel, 2019.

LEDOUX, JE. "Emotion circuits in the brain". *Annual Review of Neuroscience*, 23, 2000, 155-184.

LEVY, N., *Neuroética. Retos para el siglo XXI*. Barcelona: Avarigani Editores, 2014.

LEWIS, D., "Reduction of Mind". En: GUTTENPLAN, S. (Ed.), *A Companion to the Philosophy of Mind*. Oxford: Blackwell, 1994.

LEWIS, M., "Self-conscious emotions: Embarrassment, pride, shame, and guilt". En: LEWIS, M. Y HAVILAND-JONES, JM. (Ed.), *Handbook of Emotions* (623-636). New York: The Guilford Press, 2000.

LÓPEZ DE MÁNTARAS, R. y MESEGUER, P., *Inteligencia Artificial*. Madrid: Consejo Superior de Investigaciones Científicas, 2017.

MARTÍNEZ-FREIRE, P., *La nueva filosofía de la mente*. Barcelona: GEDISA, 2002.

MCCULLOCH, W. y PITTS, W., "A Logical Calculus of the Ideas Immanent in Nervous Activity". *The Bulletin of Mathematical Biophysics*, 5, 1943, 115-133.

MCDOWELL, J., "Criteria, Defeasibility, and Knowledge". *Proceedings of the British Academy*, 68, 1982, 455-79.

MOORE, GE., "Cramming more components onto integrated circuits". *Electronics,* 38(8), 1965.

NAGEL, T., "What is it Like to be a Bat?". *Philosophical Review*, 83, 1974, 435-456.

NEWELL, A. y SIMON HA., "Computer Science as Empirical Enquiry: Symbols and Search". *Communications of the Association for Computing Machinery*, 19, 1976.

OAKLEY, K. y JOHNSON-LAIRD, PN., "Toward a Cognitive Theory of Emotion". *Cognition and Emotion,* 1, 1987, 29-50.

PICARD, RW., *Affective Computing.* New York: MIT Press, 1997.

PINEDO IA. y YÁÑEZ CJ., "Las emociones: una breve historia en su marco filosófico y cultural en la época antigua". *Cuadernos de Filosofía Latinoamericana*, 39(119), 2018, 13-45. doi: 10.15332/25005375.504.

PUTNAM, H., "Minds and Machines". En HOOK, S. (Ed.). *Dimensions of Minds*. New York: New York University Press, 1960, 138-164.

RAPAPORT, WJ., "Syntactic Semantics: Foundations of Computational Natural Language Understanding". En FETZER, JH. (Ed.), *Aspects of Artificial Intelligence*. Dordrecht: Kluwer Academic Publishers, 1988, 81-131.

REEVÉ, J., *Motivación y Emoción*. Madrid: MCGraw-Hill, 1994.

REISENZEIN, R., "Cognition and emotion: a plea for theory". *Cognitive Emotion*, 33(1), 2019, 109-118. doi: 10.1080/02699931.2019.1568968.

RORTY, R., "Contemporary Philosophy of Mind". *Synthese*, 52, 1982, 323-348.

RUSSELL, S. y NORVIG, P., *Artificial Intelligence: A Modern Approach*. New Jersey: Prentice Hall, 2009.

RYLE, GILBERT., *El concepto de lo mental*. Buenos Aires: Paidós, 1967.

SCHULLER, D. y SCHULLER BW., "The Age of Artificial Emotional Intelligence". Computer, 51(9), 2018, 38-46. doi: 10.1109/MC.2018.3620963.

SEARLE, JR., "Minds, Brains and Programs". *Behavioral and Brain Sciences*, 3(3), 1980, 417-457. doi: doi.org/10.1017/S0140525X00005756.

—, "Minds and Brains without Programs". En BLAKEMORE, C., (Ed.), *Mindwaves*, Blackwell, 1987.

SENTENTIA, W., "Neuroethical Considerations: Cognitive Liberty and Converging Technologies for Improving Human Cognition. *Annals of the New York Academy Sciences,* 1013, 2004, 221-228. doi: 10.1196/annals.1305.014.

SMART, J., "Sensations and Brain Processes". *The Philosophical Review*, 68 (2)2, 1959, 141-156.

SPINOLA, J. y QUEIROZ, J., "Artificial Emotions: Are We Ready for Them?". *Advances in Artificial Life*, 4648, 2007.

STONE, M., y HIRSH, H., "Artificial Intelligence: The Next Twenty-Five Years". *IA Magazine*, 26(4), 2006, 85-97. doi: doi.org/10.1609/aimag.v26i4.1852.

TANTAM, D., "The flavour of emotions". *Psychology Psychotherapy*, 76(1), 2003, 23-45. doi: 10.1348/14760830260569229.

TEGMARK, M., *Vida 3.0*. Barcelona: Editorial Taurus, 2018.

TIRAPU-USTÁRRUZ, J.; PÉREZ-SAYES, G.; EREKATXO-BILBAO, M. y PELEGRÍN-VALERO, C., "¿Qué es la teoría de la mente?" *Revista de Neurología*, 44(8), 2007, 479-489. doi: doi.org/10.33588/rn.4408.2006295.

TORRES, E.; TORRES, A.; HERNÁNDEZ-ÁLVAREZ, A. y YOO, S., "EEG-Based BCI Emotion Recognition: a Survey". *Sensors*, 20(18): 5083, 2020. doi: 10.3390/s20185083.

VINGE, V., "The Coming Technological Singularity: How to Survive in the Post-Human Era". Ponencia presentada en *VISION-21 NASA Lewis Research Center and the Ohio Aerospace Institute*, 1993.

VOGT, BA., "Cingulate cortex in the three limbic subsystems". *Handbook of Clinical Neurology*, 166, 2019, 39-51. doi: 10.1016/B978-0-444-64196-0.00003-0.

WALDROP, MM., "The Chips are down for Moore's Law". *Nature,* 530(7589), 2016, 144-147. doi: 10.1038/530144a.

WENGER, MA.; JONES, FN. y JONES, MH., "Emotional Behavior". En CANDLAND, DK., (ed.). *Emotion: Bodily Change*. Princeton: van Nostrand, 1962.