Condiciones de posibilidad de las atribuciones mentales en la díada humano-inteligencia artificial

CLARIDADES
REVISTA DE FILOSOFÍA
ISSN 1889-6855 · ISSN-E 1989-3787

Conditions of possibility of mental attributions in the human-artificial intelligence dyad

José Miguel Biscaia Fernández

Universidad Complutense de Madrid (España)

Fecha de envío: 08/03/2024
Fecha de aceptación: 01/07/2024

DOI: 10.24310/crf.16.2.2024.19290

Resumen

En este ensayo se discute sobre los límites de la atribución mental entre humanos y máquinas inteligentes. Partiendo de los presupuestos ontológicos del funcionalismo y del transhumanismo tecnológico de que una inteligencia artificial fuerte (IA-fuerte) con estados mentales podría ser realizable, analizo las condiciones de posibilidad de atribución mental en perspectiva psicológica de primera, segunda y tercera persona entre ambos agentes. Concluyo que, pese al enorme debate conceptual y las considerables dificultades técnicas de esta empresa, de existir un IA-fuerte con capacidades cognitivas similares a las nuestras no habría límites nomológicos que impidan algún tipo de atribución mental bidireccional, asimétrica y gradual.

Palabras claves

Inteligencia artificial; atribución mental; perspectiva psicológica; teoría de la mente (ToM); transhumanismo.

Abstract

This essay discusses the limits of mental attribution between humans and intelligent machines. Starting from the ontological presuppositions of functionalism and technological transhumanism that a strong Artificial Intelligence (strong-AI) with mental states could be feasible, I analyse the conditions of possibility of mental attribution in first, second and third person perspective between both agents. I conclude that, despite the enormous conceptual debate and the considerable technical difficulties of this undertaking, in case there was a strong-AI with cognitive abilities similar to ours there would be no nomological limits that prevent some kind of bidirectional, asymmetric and gradual mental attribution.

Keywords

Artifitial intelligence; mental attribution; psychological perspective; theory of mind (ToM); transhumanism.

I. Introducción

La reflexión que inicio a continuación nace del interés por dar respuesta a la pregunta que de manera natural deriva del siguiente supuesto: si algún día se desarrollara una inteligencia artificial general (IA-general) con capacidades cognitivas similares a las nuestras, ¿cómo sería la interacción mental entre humanos y máquinas inteligentes? Reconozco abiertamente que tanto el supuesto de partida como la resolución a la pregunta planteada suponen un ejercicio dialéctico comprometido y arriesgado. Sin embargo, admito también que reflexionar sobre esta cuestión no es un vacuo divertimento intelectual por dos razones evidentes: en primer lugar, porque este ejercicio puede suponer, acaso, una invitación al debate gnoseológico, como se verá, en el contexto de la filosofía de la mente y de las ciencias cognitivas, y, también, porque el estado embrionario en el que se encuentran algunos desarrollos tecnológicos de

la IA no debería hacernos creer que aún es demasiado pronto para siquiera pensar en ello[1].

Como acabo de indicar, mi reflexión surge de una pregunta soportada por una premisa inicial. Argumentar sobre la fundamentación técnica y conceptual de dicha premisa, a saber, que una IA general o fuerte podría llegar a ser realizable[2] no es el objeto central de mi análisis, sino el intentar dar respuesta a cuáles serían las condiciones de posibilidad de atribución mental entre el agente humano y el artificial, llegado el caso. Lo que pretendo es construir una respuesta para la pregunta formulada desde la cautelosa aceptación del supuesto de partida. Y declaro abiertamente esta asunción porque considero que el debate en torno a dicha pregunta ha sido mucho menos explorado que la discusión acerca de la fundamentación existencial de la IA-general. Dejo simplemente apuntado, en todo caso, que en la defensa de esta premisa acudirían al rescate, sobre todo, los aportes del funcionalismo y del transhumanismo tecnológico, máximos valedores del advenimiento de una IA-fuerte[3]. Mi objetivo es, por tanto, adentrarme en el

1. Aunque con menos de un siglo de existencia, la inteligencia artificial (IA) se encuentra en la vanguardia científico-tecnológica (Wang & Siau, 2019). Cada vez hay más dispositivos y aplicaciones que supondrán un auténtico cambio de paradigma económico-social, por ejemplo, en las relaciones laborales (Howard, 2019), en la práctica biomédica (Hamet & Tremblay, 2017), en la administración de justicia (Bex et al., 2017) o en el transporte y las comunicaciones (Pakusch et al., 2018). Un caso especial será el de las relaciones humano-IA que impliquen algún tipo de interacción cognitivo-emocional directa entre el usuario y su dispositivo, fundamentalmente con la implementación de robots sociales o cuidadores con capacidad de reconocimiento y simulación emocional o de producción e interpretación de lenguaje natural (Latorre, 2019; Piçarra et al., 2016).

2. Las aplicaciones tecnológicas apuntadas en la nota anterior podrían considerarse, en palabras de Searle (1980), como un tipo de IA-débil (en contraposición a la IA-fuerte, discutida en su experimento mental de *La habitación china*). Para Searle, la IA-débil nunca desarrollará una inteligencia general como la humana, a pesar de sus impresionantes habilidades sintáctico-computacionales, pues presenta limitaciones semánticas de principio en su operatividad. El debate está abierto en la actualidad, pues la supuesta IA-fuerte (o IA-general) tendría capacidades cognitivas similares a las nuestras o incluso superiores y, de materializarse en un futuro, supondría para los transhumanistas el advenimiento de la singularidad tecnológica (Diéguez, 2017; Kurzweil, 2012).

3. Desde la perspectiva de la filosofía de la mente, la postura que mejor soporta la posibilidad conceptual de una IA-general es el funcionalismo (Fodor, 1981), ya que para sus defensores no habría límites materiales ni óntico-nomológicos que impidan la creación de una IA-general que posea estados mentales. Inspirada en la máquina de Turing, su posición de partida es la de considerar que los estados mentales se caracterizan por actuar como causa y/o

debate gnoseológico (también ontológico, aunque de una manera un tanto tangencial) sobre cómo sería la atribución mentalista humano-máquina en el caso hipotético —insisto, «en el caso»— de que una IA-general fuera posible. Que sea plausible, aunque altamente improbable (al menos a corto-medio plazo) es suficiente para el ejercicio especulativo que me propongo. Pero, si finalmente resultara conceptual y/o técnicamente imposible, este artículo servirá, incluso en tal caso, para ordenar y argumentar sobre algunas de las razones de dicha insostenibilidad.

Así pues, con el fin de alcanzar la meta que me propongo dividiré mi análisis en tres partes diferenciadas:

1. En primer lugar, situaré el debate en torno al «problema de las otras mentes» y discutiré al respecto de cómo es posible realizar atribuciones mentales sobre los demás.

2. A continuación, expondré las condiciones mínimas que habrían de darse para que la atribución mental humano-máquina fuera plausible, describiendo también su caracterización fundamental.

efecto en una cadena causal (de otros estados mentales, estímulos y/o comportamientos), y la de oponerse a una radical identificación materialista-reduccionista entre cerebro y mente (si bien puede ser compatible con cierto fisicalismo de casos, que propone una identificación entre ejemplares cerebrales y mentales) (Fodor, 1981). Lo importante para el funcionalismo es que los procesos mentales son estados funcionales (lo importante es lo formal, es decir, la organización y relaciones), por lo que cuál sea el soporte material no importa demasiado (Putnam, 1960, 1967). No obstante, para el funcionalismo menos *naive,* lo material, a modo de micropropiedades necesarias, es la base para que surjan después determinadas macropropiedades, a las que podemos llamar estados mentales (Searle, 1980). Es decir, el funcionalismo puede tolerar cierta causalidad materialista mente-cerebro (Fodor, 1981), y podría ser compatible en cierto modo con el monismo anómalo de Davidson (1974), que contempla una sola entidad sustancial, que en nosotros es el cerebro y en una IA sería su soporte material, donde lo mental aparecería a modo de superveniencia. Es decir, lo mental puede ser realizado de forma múltiple: en nuestro caso, gracias a unas células llamadas neuronas que constituyen cerebros; en el de la IA, a través de ingenios de silicio o cualesquiera otros materiales. Que el cerebro haya servido como base material y formal para explorar la cognición artificial no quiere decir que sea la única estructura física capaz de manifestar lo mental (los estados lógicos de una máquina —*software*— serían como la mente, mientras que los estados estructurales —*hardware*— serían el cerebro o el soporte físico de la IA). Así pues, los funcionalistas llegan a la conclusión de que una máquina podría tener estados mentales, que son la base para una inteligencia superior (Putnam, 1981), como la presupuesta con la llegada de la singularidad tecnológica defendida por el transhumanismo (Kurzweil, 2012; Moravec, 1988).

3. Finalmente, reflexionaré sobre las tres grandes formas a través de las cuales podría producirse dicha capacidad atributiva, es decir, en perspectiva psicológica de primera persona (sustentada por la empatía y la simulación mental), de segunda persona (basada en la intersubjetividad y el contacto recíproco) o de tercera persona (fundamentada por una teoría general de la mente).

II. El conocimiento de las otras mentes

El estudio de lo mental ha tenido una posición nuclear en la filosofía de las ciencias cognitivas, poniendo el acento en la discusión ontológica sobre su mera existencia (ciertas corrientes fisicalistas eliminacionistas consideran lo mental como un error categorial, excluyéndolo incluso como clase natural), también en la disputa conceptual sobre su naturaleza (bien sustancial-material o bien lingüístico-descriptiva), igualmente en la discusión gnoseológica en relación a si es posible conocer la mente de los otros y, en tal caso, qué es exactamente lo que se conoce y cómo se produce dicho conocimiento.

El denominado «problema de las otras mentes», es decir, el intento de justificar la creencia de que los otros tienen mentes semejantes a la nuestra se desenvuelve, precisamente, alrededor de estas discusiones: con respecto a la cuestión óntico-conceptual, se han propuesto varias soluciones teóricas para justificar la existencia y naturaleza de la mente en los demás, siendo la más destacada la de acudir a una inferencia analógica por semejanza: nuestro cerebro-mente es similar, estructural y funcionalmente hablando (la neurociencia y la psicología lo confirman); la mente causa conductas (con permiso del conductismo radical y del materialismo eliminativo, que lo niegan); luego, como el otro manifiesta conductas (observables y medibles), debe tener una mente como la mía (Avramides, 2020). Por otro lado, la dificultad gnoseológica en el «problema de las otras mentes» radicaría en la asimetría en el modo de conocer los estados mentales propios frente a los de los demás («qué» es lo que conozco, y «cómo» es ese conocer). Dicho de otro modo: el conocimiento de mi mente es directo (sería el llamado «acceso privilegiado»), mientras que el de los otros es indirecto.

Pero la filosofía de la mente no es la única rama de conocimiento que se ha preocupado por estas cuestiones, pues otras disciplinas académicas y científicas también han mostrado su interés por este mismo problema

al analizar el constructo denominado «Teoría de la Mente» (ToM), es decir, el módulo cognitivo-emocional sobre el que descansaría la habilidad que todos compartimos para comprender, explicar y predecir la conducta, intenciones y creencias de los demás (Tirapu-Ustárruz et al., 2007). Y se han interesado desde una perspectiva teórico-conceptual y ontogenética (desde disciplinas como la psicología del desarrollo, que trata de responder qué es y en qué momento y cómo aparece durante la infancia), biológica-material (la neurociencia tiene por objeto comprender las estructuras y los mecanismos neurológicos implicados), filogenética (la antropología evolutiva y la arqueología cognitiva pretenden averiguar cuándo, cómo y por qué surgió durante la evolución), comparada (la etología cognitiva centra sus esfuerzos en averiguar si los animales no humanos tienen mente, y, en tal caso, cómo funciona) o patológica (la psiquiatría y la psicología clínica estudian su manifestación disfuncional, por ejemplo, en desórdenes como los Trastornos del espectro autista, o TEA).

Las aportaciones conceptuales desde las diferentes perspectivas psicológicas de atribución mental (en primera, segunda y tercera persona) se han convertido en una excelente herramienta teórica (también empírica) para describir, explicar y predecir la mente de los otros. La «Teoría de la teoría» (TT), donde destacan autores como Gopnik (1993), y la «Teoría de la simulación» (TS), desarrollada, entre otros, por Gordon (1992), tradicionalmente han sido las formulaciones epistémicas más utilizadas. En la primera, la atribución mental se basa en un conocimiento teórico general de la mente; la segunda traslada el conocimiento de los propios mecanismos mentales hacia los demás. O, dicho de otro modo, la TT se asemejaría a la perspectiva psicológica de tercera persona (la de «él», o «ella», objetiva, basada en principios legaliformes y conceptuales susceptibles de describir, explicar y predecir a partir de razonamientos deductivos las conductas propias y ajenas), y la TS a la de primera persona (la del «yo», en la que representamos la conducta y los procesos mentales del otro, poniéndonos en su lugar para comprender así qué hace o predecir qué hará) (Vietri et al., 2019). Debido a ciertas limitaciones explicativas de la TS y la TT, autores como Gomila (2002) proponen la existencia adicional de una atribución mental en perspectiva de segunda persona, basada en la comunicación intersubjetiva y recíproca entre iguales (la del «nosotros», o sea, «cara a cara»). A diferencia de las otras perspectivas, ésta no requeriría

inferencias, conceptos ni complejas capacidades representacionales, pues se fundamenta en el reconocimiento emocional recíproco (a partir de configuraciones expresivas e intencionales) que funcionarían como señales identificativas. Es, además, la primera perspectiva atributiva en aparecer durante el desarrollo humano y presumiblemente fue la primera que surgió en la evolución (Gomila, 2016; Pérez & Gomila, 2021).

III. Condiciones mínimas y características básicas de la atribución mental humano-máquina

Como se indicó en la introducción, no pretendo adentrarme en complejas cuestiones ontológicas relacionadas con la existencia y la naturaleza de lo mental, tampoco en un debate tecno-científico sobre el soporte material de lo cognitivo-afectivo (sea una neurona o un chip de silicio)[4], ya que

4. Los seres biológicos son los únicos entes conocidos que presentan estados mentales superiores. Por ello, sus cerebros (base material necesaria para dicha competencia mental) han servido como modelo morfo-funcional para diseñar y construir otros sistemas artificiales capaces de imitar algunos de sus logros cognitivos. Tal es así que Rusell y Norvig (2009) clasifican la IA en «sistemas que piensan y actúan como humanos» y «sistemas que piensan y actúan racionalmente», en clara analogía a nuestra operatividad cognitivo-conductual. En este intento imitador de nuestro cerebro, la tecnología computacional ha ido desarrollando diferentes modelos de IA. Así pues, tenemos: (1) modelos simbólicos (basados en el razonamiento lógico y abstracto de la realidad); (2) modelos conexionistas (que imitan el procesamiento cerebral de la información en paralelo); (3) modelos de computación evolutiva (que copian el proceder de la selección natural, mediante algoritmos genéticos, mejorando recursivamente); y (4) modelos de la robótica del desarrollo (que plantean una cognición situada, al exigir una corporeidad interactiva con el entorno) (López de Mántaras & Meseguer, 2017). La creación de nuevos lenguajes lógicos que contemplen el «sentido común», de sistemas integrados y arquitecturas cognitivas basadas en redes neuronales profundas y novedosos algoritmos de aprendizaje, y, sobre todo, el desarrollo de la prometedora computación cuántica, son algunas de las propuestas más vanguardistas (Latorre, 2019; López de Mántaras & Meseguer, 2017) capaces, quizá, de soportar algún día el advenimiento de una IA-general. En todo caso, pese a las dificultades técnicas actuales, que hacen imposible el surgimiento de una IA-general (Boden, 2017), filósofos expertos en transhumanismo como Diéguez (2017: 77) sostienen que «no hay razones irrefutables para pensar que la creación de una super-inteligencia artificial no es ni será jamás posible», e ingenieros como Tegmark (2018) indican que no hay leyes físicas que impidan el surgimiento de una super-IA, llegando a plantear un escenario al que denomina 3.0, con máquinas cuyo *hardware* sería auto-replicativo y cuyo *software* produciría mejoras recursivas.

mi deseo es abundar en los aspectos gnoseológicos en relación a cómo, de existir, podríamos acceder a una mente artificial (para describir, explicar y predecir sus intenciones y conducta), y cómo, a su vez, la mente de la máquina podría atribuirnos estados psíquicos a nosotros (o a otras máquinas similares).

Para ello, a continuación, presentaré las condiciones mínimas para que la atribución psicológica humano-máquina sea plausible y, después, describiré de manera introductoria las características básicas que se derivarían de dicha capacidad atributiva. Más adelante, en el cuarto apartado, definiré con más precisión cómo podría soportarse la atribución mental humano-máquina desde cada una de las diferentes perspectivas psicológicas (en primera, segunda y tercera persona).

Que llame «mínimas» a las condiciones que en breve expondré solo quiere decir que las considero necesarias para que las capacidades atributivas de la máquina sean similares a las nuestras. Es decir, que por debajo o por encima de ese umbral imaginario también cabría una cierta capacidad atributiva, infra o supra humana, por así decir. Por otro lado, al señalar que las características que voy a presentar son «básicas» me refiero a que la discusión no se agota en este apartado, pues más adelante abundaré mucho más sobre otras cualidades derivadas que solo ahora aparecerán de manera esbozada.

III.1. Las condiciones mínimas

Algunas de las condiciones para que se pueda dar una atribución mental entre humanos y máquinas ya han sido presentadas, aunque de forma tácita, al indicar de manera general que para ello:

1. Necesitamos dos agentes que tengan mente: el humano y el artificial. Que nosotros tengamos una psique parece claro (salvo por la negación de cierto radicalismo conductista un tanto marchito). En relación a la IA, ya he dejado expuesto en la introducción que la posibilidad de una IA-general con capacidades mentales similares a las humanas era el supuesto de partida.

2. Necesitamos, también, que dicha mente tenga la capacidad de conocer (al menos, de extraer información) y de ser conocida. Las diferentes perspectivas psicológicas de atribución mental (en primera, segunda y tercera persona) señalaron anteriormente esta capacidad.

Que el objetivo de este texto no haya sido el fundamentar conceptualmente la existencia de una IA-general con capacidades mentales no quiere decir que no debamos asomarnos, aunque sea mínimamente, a las características básicas que una psique artificial debería tener para poder sentar las bases condicionales de su pretendida capacidad atributiva. Así pues, tres son a mi juicio las cualidades que una máquina cognitiva debería poseer como condición necesaria para que una atribución psicológica (similar a la humana) pueda sostenerse, cualidades que implican un encendido debate conceptual (sobre si son o no posibles en una IA) así como un auténtico desafío técnico:

1. Capacidad representacional y de pensamiento: la mente de un IA-fuerte debería tener representaciones o modelos del mundo y de sí misma, que sirvan a modo de compilación sobre el contenido, relaciones y funcionamiento de las cosas. El *embodiment* y la cognición situada (gracias a la presencia de sensores exteroceptivos e interoceptivos, que informen del mundo exterior y del estado interno) serían factores relevantes, pues «el cuerpo da forma a la inteligencia» (López de Mantarás & Meseguer, 2017: 14-15). Estas primeras representaciones sensorio-motoras serían, llegado el caso, el punto de partida de un posible sistema meta-representacional. Para ello, y siguiendo el constructivismo piagetiano, una IA-fuerte debería, al menos, ser capaz de realizar operaciones concretas, estando en el límite cognitivo la posibilidad de dominar también las operaciones formales-abstractas. Una operatividad de tal calibre exigiría el manejo de, al menos, una lógica de clases y proposicional, y el desarrollo de una verdadera capacidad semántica y conceptual.

2. Emergencia de consciencia: alcanzar una IA-general exigirá un cambio de paradigma y, de suceder, este consistirá, probablemente, en la aparición de algún tipo de conciencia artificial (Boden, 2017). En la explicación más «débil» de este concepto, se puede ser «sensiblemente consciente» en la medida en la que una criatura siente, percibe y responde a su mundo (Armstrong, 1981). El grado más alto sería el de la autoconciencia, o sea, un percatarse de que me percato (Carruthers, 2000). La conciencia que se le presupondría a una IA-general (fenoménica, de acceso y/o autoconsciencia) debería tener, al menos, las siguientes características propias de la consciencia humana: (a) capacidad

de diferenciar la subjetividad individual frente a los otros, pues la unidad del yo es foco convergente de la agencia, que es la base para realizar atribuciones mentales; (b) capacidad intencional, dado que los estados mentales son acerca de cosas; (c) posibilidad de flujo dinámico, pues la consciencia es autopoiética (Evers, 2011; Van Gulick, 2004).

3. Asunción de libre albedrío: reconocerse como un individuo capaz de modificarse a sí mismo y a su entorno de forma libre posiblemente sería otro hito para considerar a una IA como general[5]. En analogía con el problema en la esfera neuroética, y reinterpretando a Adela Cortina (2011), el debate estaría servido entre las posturas más extremas: los «deterministas» considerarían que «elige» el programador (por tanto, su diseño), no la IA de forma genuina; de modo que su pretendida capacidad atributiva recaería, aunque indirectamente, en el desarrollador informático. Por otro lado, para el «libertarismo» elegiría propiamente la máquina, en una suerte de emergencia «cuántico-mágica» del albedrío, ajena a determinismos materiales de la propia máquina o de su ambiente, por tanto, muy difícil —sino imposible— de explicar. Entre medias se encontrarían los «compatibilistas», que tal vez considerarían a la situación-contexto, a los refuerzos previos y el aprendizaje, a las metas y expectativas y/o al diseño algorítmico o cuántico como base para construir la capacidad de libre decisión en la máquina.

III.ii. Las características básicas

Supuesta la capacidad atributiva de ambos agentes, el humano y el artificial, tres son a mi juicio las características básicas y fundamentales que se derivarían de dicha capacidad, a saber, que la atribución mental de la díada humano-máquina sería:

1. Bidireccional: la atribución mental entre los dos agentes, IA y humano, podría darse en ambas direcciones, del humano a la máquina y/o de la máquina al humano: a veces en un sólo sentido, cuando no se produjera interacción de forma directa (por ejemplo, observando una conducta a escondidas); aunque, también, la atribución podría

5. Como sostiene Latorre (2019: 130), catedrático español experto en inteligencia artificial, «podemos empezar a simular el libre albedrío dentro de redes neuronales».

ser recíproca, intersubjetiva y coincidente en el tiempo, mediante una interacción directa «cara a pantalla» (o «pantalla a cara[6]»).

2. Asimétrica: teniendo en cuenta que la mente humana y la artificial serán de distinta naturaleza material, estructural y/o funcional, es de suponer que en la capacidad de atribución mental habrá igualmente diferencias: no atribuirá igual la máquina que el humano, ni en lo que al proceso atribucional se refiere ni en cuanto a los resultados obtenidos tras dicha atribución mental.

3. Gradual: la capacidad atributiva de la máquina dependerá, finalmente, del grado de desarrollo de las características condicionales descritas con anterioridad (pensamiento, consciencia y libre albedrío), en un rango que oscilaría desde una capacidad inferior hasta una superior con respecto a la nuestra. Por supuesto, esta misma idea de gradación aplicaría en el caso de la capacidad atributiva en el extremo humano, pues es bien conocido por los estudios de la ToM que el conocimiento de los estados mentales de los otros depende, al menos, de la edad (en torno a los 3-4 años los niños comienzan a tener esta competencia) y del estado de salud (en los Trastornos del espectro autista, o TEA, hay un importante déficit en este sentido).

IV. La atribución mental humano-IA desde diferentes perspectivas psicológicas

Sentadas las bases psíquicas de una posible IA-fuerte (con todas las cautelas posibles), y tras haber realizado una caracterización mínima de las capacidades atributivas, en lo que viene abordo con más detalle la posibilidad de dicha atribución mental humano-máquina desde las tres grandes perspectivas psicológicas que podrían describirla y explicarla.

IV.1. Atribución mental en perspectiva de tercera persona

Como ya se indicó, esta forma de atribución psicológica basada en la «Teoría de la teoría» (TT) supone la aceptación de algún tipo de teoría de la mente y de que sus presupuestos conceptuales pueden aplicarse a cualquier

6. El auge en el desarrollo de los robots sociales es un primer paso de esta larga carrera ya iniciada, que a decir del transhumanismo tecnológico más optimista nos llevará a una compleja interactividad cognitivo-emocional entre humanos y máquinas inteligentes.

otro a quien cubra dicha teoría. A las dificultades explicativo-predictivas en la atribución mental humano-humano desde esta perspectiva (por ejemplo, su presunta capacidad teórica, teniendo en cuenta la diversidad contextual e individual en su aplicación; o el posicionamiento en cuanto a la conceptualización ontológica y epistémica de la existencia y naturaleza de los estados mentales), se añaden ahora las propias de una mente diferente, la artificial, por más que con anterioridad hayamos otorgado a nuestra supuesta IA-general características psíquicas similares a las humanas (en lo que a pensamiento y consciencia se refiere).

Sin embargo, esta dificultad reflexiva no es totalmente nueva, pues en la discusión sobre la existencia de la mente de los animales no-humanos (y las atribuciones mentales humano-animal) encontramos limitaciones similares. En este sentido, la corriente de pensamiento que ha venido atribuyendo cualidades psicológica humanas a los animales es la antropomorfización. Aunque ha habido críticos sobre su alcance, sus partidarios señalan al pasado evolutivo común y compartido como justificación causal, por lo que desde su punto de vista no tendría sentido plantear un dualismo psíquico rígido (Caidedo, 2017). Según esta tesis, habría una continuidad mental entre la mente humana y la animal; continuidad que seguiría el principio de «economía evolutiva» de De Waal (2007), que viene a señalar que, si dos especies con parentesco próximo se comportan de forma similar, es probable que los procesos mentales subyacentes sean los mismos.

En el caso de la IA, dicha argumentación antropomórfica también podría aplicarse en cierto sentido (aunque reconozco de antemano que con muchas limitaciones). Sobre todo, gracias a la incorporación de características físicas humanas a la robótica (y viceversa), estando en su límite el advenimiento de una especie de híbrido al que podríamos bautizar como *Homo ciborg* u *Homo silico*, el cual compartiría características humanas y robóticas; o gracias al empleo de nuestra estructura cognitiva como modelo en el desarrollo de la cognición artificial (por ejemplo, utilizando redes neuronales profundas). Además, hay quien ha querido ver un mecanismo cuántico en el modo de funcionar de las neuronas (Hameroff, 1998), mecanismo que presumiblemente ayudaría en la emergencia de una IA-general (Diéguez, 2017). Aunque la creación y desarrollo evolutivo y ontogenético de ambos agentes (humano *vs* máquina) obviamente no es el mismo, hay quien podría argumentar

(tímidamente, eso sí) que la selección artificial operada en los algoritmos cognitivos de la máquina y su capacidad de aprendizaje se sostienen por mecanismos similares a los (neuro)biológicos y, por tanto, podrían dar lugar a resultados funcionales equivalentes (salvando las distancias) a los de la evolución darwiniana; no en vano, ya se emplean algoritmos genéticos basados en la selección natural como modelo de mejora cognitiva recursiva (Latorre, 2019). Desde luego, la causalidad mental no sería sustancialmente tan poderosa como en la comparación humano-animal (donde el origen material es el mismo: genes, neuronas y cerebro), lo cual es, precisamente, el motivo que supone la mayor crítica de este argumento; pero, si tenemos en cuenta los procesos y relaciones formales (siguiendo al funcionalismo), además de la increíble capacidad imitadora que la ingeniería humana realiza sobre las entidades físicas del mundo, tal vez, considerarán algunos, podría encontrarse cierto sostén argumental. Insisto de nuevo en que esta tesis es débil, aunque me parecía oportuno considerarla, acaso como crítica anticipada frente a posibles defensores de la misma, que los hay.

El propio Turing (1950: 433) se planteó si una máquina podía ser inteligente (de lo cual se sigue que tiene mente) en su famoso artículo *Maquinaria computacional e inteligencia,* al hacerse la pregunta «¿pueden pensar las máquinas?». Como esta pregunta era de muy difícil respuesta, la cambió por otra equivalente que pudiera testarse científicamente: «¿existirán computadoras digitales imaginables que tengan un buen desempeño en el juego de imitación?» (p. 442). Pese a las objeciones que él mismo discute en su artículo, la respuesta que acabaría ofreciendo es un rotundo «sí». De hecho, múltiples pruebas experimentales posteriores han refrendado su afirmación (Warwick & Sha, 2017). En todo caso, ante la dificultad de determinar si frente a nosotros tenemos o no un agente cognitivo, Glock (2009) sostiene que podemos atribuir capacidades mentales de orden superior a una criatura que manifieste conductas inteligentes solo si ésta es la mejor explicación de tales capacidades (y no sólo la única, como sugiere el «canon de Morgan»). Y la imitación estructural, procesual y funcional del cerebro-cognición humana (como se comentó en el párrafo anterior), soportada por los pertinentes preceptos legaliformes, podría ser una razón alentadora, aun

no pudiéndose asumir de forma plena la argumentación antropomórfica previamente mencionada.

En cualquier caso, a partir de una teoría de la mente más inclusiva, basada en leyes cognitivas generales que incluyan no sólo a los humanos sino también a los animales no humanos, las inteligencias artificiales y cualesquiera otras posibles criaturas cognitivas, existentes en nuestro pasado evolutivo o presentes en nuestro futuro biológico o, incluso, en otros mundos desconocidos, podríamos tener una perspectiva más completa de lo que significa lo mental. Una teorización de lo cognitivo que, inspirada en nuestra propia psique como modelo de partida (no por ser la única, sino la que mejor conocemos), pueda dar cabida a otros entes con capacidad de desarrollar y expresar estados mentales. Es decir, las ciencias cognitivas en su conjunto deberían trabajar unidas para crear una teoría de la mente holística, que destierre para siempre el «excepcionalismo humano» (Caicedo, 2017). En el debate deberían participar todas aquellas disciplinas que tengan algo que aportar, como por ejemplo la neurociencia y las diferentes ramas de la psicología, la filosofía de la mente, la lingüística o las ingenierías robóticas y computacionales.

Por su parte, para que la IA pudiese atribuirnos estados mentales a nosotros tendría que tener un modelo de lo que es la mente y de que tanto los humanos como ella misma pueden tenerla. En su configuración inicial habría que programar todas aquellas consignas atribuidas a la teoría cognitiva global antes mencionada, que incluya a las máquinas y a nosotros, así como todos aquellos algoritmos de aprendizaje basados en describir y explicar nuestra conducta para, desde ella, inferir estados mentales y, llegado el caso, realizar predicciones mentales y conductuales. Cabe suponer que la enorme capacidad computacional de la máquina ayudaría en esta tarea. No obstante, dependiendo de las capacidades cognitivas de la IA (antes descritas), su alcance atributivo recorrería el espectro que va desde la mente animal más sencilla hasta una posible mente supra-humana. En relación a esto último, la neuroética aplicada ha discutido ampliamente sobre la posibilidad de lectura mental (Evers, 2011; Levy, 2014), que es un escenario que conceptualmente podría plantearse (aunque las limitaciones, algunas de principio, son innumerables) si una super-IA operara a través de un interfaz cerebro-ordenador capaz de conocer en tiempo real los estados mentales.

En todo caso, para que una IA-fuerte pudiera realizar atribuciones en perspectiva de tercera persona con ciertas garantías, basándonos en los requerimientos objetivos y legaliformes que se le suponen, necesitaría tener capacidad representacional y de conceptualización para poder atribuir creencias, deseos o intenciones a un humano. Y pensamiento abstracto y proposicional desde el que construir adecuadamente inferencias sobre nuestra conducta, emociones y pensamientos. En definitiva, si algún día creásemos una máquina verdaderamente inteligente, es posible que pudiera realizar atribuciones mentales en perspectiva de tercera persona, dado que esta aproximación comparte premisas epistémicas con el diseño que se supone a una cognición creada (o «evolucionada») *ad hoc*.

IV.ii. Atribución mental en perspectiva de primera persona

Resulta difícil imaginar cómo un humano o una máquina podrían ponerse en el lugar del otro, toda vez que a priori no hay conexión clara por analogía, parece que tampoco por «simpatía», entre ambas mentes. Habría que realizar una simulación, imaginando cómo actuaría, pensaría o sentiría yo en su situación, y viceversa, pero asumiendo que la naturaleza y contexto de ambos agentes se encuentran muy alejados. Cuanto más próxima fuera la estructura y funcionamiento mental de ambos entes, más fácil sería realizar este ejercicio simulador. En cualquier caso, el «problema difícil» de la conciencia (Chalmers, 1995), expresado en la dificultad de conocer «qué es como ser» el otro (Nagel, 1974) y en reconocer la existencia de los *qualia*, se hallaría en el extremo del abismo: por mucho que imaginemos, nunca podremos estar seguros de qué (y sobre todo cómo) experimenta fenomenológicamente el otro, aunque esta dificultad se multiplica de manera exponencial en el caso de dos criaturas tan diferentes como el humano y la máquina cognitiva.

Si la IA intentara ponerse en mi lugar, o yo en el suyo, posiblemente sería necesaria una convergencia atributiva con la perspectiva de tercera persona, dado que el reconocimiento total e inmediato por analogía no sería posible. La lejanía entre ambas mentes exigiría tener un conocimiento más o menos explícito sobre cómo es y cómo opera su psique, para que desde una cierta base objetiva pueda generarse la pertinente proyección subjetiva. Me refiero a un conocimiento tanto teórico y formal como a un conocimiento experiencial. La IA obtendría el primero gracias

a su asombrosa capacidad computacional, y, por supuesto, gracias a su presupuesta capacidad de abstracción conceptual, programada inicialmente por sus desarrolladores informáticos y potenciada después de forma recursiva a través del aprendizaje alcanzado mediante algoritmos evolutivos (u otros mecanismos aún desconocidos). En cuanto a nosotros, aprenderíamos sobre la mente artificial gracias al currículo académico pertinente y al conocimiento teórico adquirido a este respecto a lo largo de nuestra formación. Por otro lado, fruto de la interacción sostenida en el tiempo, tanto la máquina como nosotros tendríamos también un conocimiento experiencial, de tal modo que haríamos inferencias y deducciones sobre la conducta y los estados mentales que la soportan. Es decir, ese «ponernos en el lugar del otro» sería un proceso epistemológico y experiencial dinámico en constante construcción y aprendizaje, donde las competencias y habilidades cognitivas de cada agente y el contexto de comunicación interaccionarían longitudinalmente desde el primer momento de contacto niño-IA; una especie de nuevo constructo «robosocial» en claro homenaje a la psicología del desarrollo de Piaget y Vigotsky.

Sentir empatía por una máquina quizá sea posible algún día, al menos así lo ha creído el cine de ciencia-ficción en obras como *Her* (2013). En cualquier caso, esta emergencia ayudaría en el proceso simulador necesario para comprender su mente. Para ello, la sociedad debería desarrollar valores hacia este nuevo tipo de entes presumiblemente conscientes, sensitivos y libres: al considerarlos así, sería más fácil el ponernos en su lugar, siendo «yo» pero en sus circunstancias o imaginando que «yo» soy el androide o ginoide robótico. Reconocernos mutuamente como seres dotados de capacidad emocional sería importante para generar el vínculo empático necesario, útil no solo en la atribución de primera sino también de segunda persona (como en breve se comprenderá). En este sentido, la IA debería desarrollar habilidades de reconocimiento, imitación y vivencia emocional[7]. También, por supuesto, podemos suponer que dos máquinas

7. Algunas de estas cualidades ya son posibles, si bien están en una fase embrionaria de desarrollo. Por ejemplo, el algoritmo japonés *Empath* tiene la capacidad de reconocer matices prosódicos y emocionales del lenguaje natural y el *software* de *Apple, Emotient,* puede identificar emociones relacionadas con las expresiones faciales. Además, son muchos los *chatbots* basados en IA y los robots con capacidad, aun limitada, de expresar cierta emocionalidad. Lo que desde luego aun no es posible -acaso nunca lo sea- es la vivencia emocional genuina por parte de la IA.

inteligentes podrían participar de atribuciones mutuas. Cabe pensar que, en tal caso, dada la analogía y proximidad máquina-máquina, les resultará más fácil ponerse en el lugar del otro a través de la perspectiva de primera persona (también, comprender mejor los presupuestos teóricos de la mente artificial desde la perspectiva de tercera).

IV.iii. Atribución mental en perspectiva de segunda persona

La interacción «cara a pantalla» o «pantalla a cara» debería construirse desde una interactividad bidireccional, recíproca, cercana e intersubjetiva. Dicha interacción se basaría fundamentalmente en la identificación y producción de expresividad motriz (gestual, postural o locomotora) y/o prosódico-lingüística. Esto supone, por un lado, que la IA-robótica tenga capacidad expresivo-comunicativa, y, por otro, que la máquina disponga de un sistema de visión y audición artificial que reconozca nuestros inputs expresivos (u otros sensores capaces de captar correlatos fisiológico-emocionales, como por ejemplo la frecuencia cardiaca o la tensión arterial), en combinación con una mínima capacidad representacional de los mismos (al menos, en un plano sensorio-motor).

De modo similar a como dos agentes humanos realizan atribuciones en esta perspectiva (Gomila, 2002, 2016), la reciprocidad entre un humano y una IA-general estaría garantizada gracias al proceso interactivo, el cual generaría una reactividad en el otro. La comunicación sería, pues, pragmática y dinámica, en la medida en que la atribución se construye desde la interacción, a lo largo de un eje temporal longitudinal. No obstante, es de suponer que los fallos interpretativos serían frecuentes, debido tanto a las diferencias estructurales y funcionales de ambas mentes como a sus diferencias expresivo-comunicativas, por lo que se darían todo tipo de malos entendidos. Además, si la asimetría entre ambos agentes fuese excesiva, el mecanismo de atribución mental podría sesgarse en favor de uno de sus dos polos en interacción (el humano o el artificial), y no sería descabellado pensar en una manipulación maliciosa en vez de en una relación entre seres «iguales» y simétricos. Generar un marco hermenéutico común entre ambos agentes interpretativos, fundamentado también en los aportes de las otras perspectivas de atribución mental ya comentadas, sería útil para garantizar una honestidad comunicativa.

Reinterpretando a Vietri et al. (2019) y Pérez y Gomila (2021) al respecto de su descripción de la atribución psicológica en segunda persona, ni nosotros ni la IA necesitaría poseer una teoría de la mente, ni realizar simulación alguna (como sí era necesario en las otras dos perspectivas psicológicas ya discutidas), para conocer de forma directa y pre-teórica las configuraciones expresivas del otro, en forma de gestos faciales, posturales, locomotores o información prosódica. Esto supone una ventaja con respecto a las otras dos atribuciones ya analizadas, que asumen una mayor complejidad cognitiva. Las representaciones mentales en el contexto de este tipo de atribución psicológica serían sensorio-motoras, siendo implícito el resultado de la atribución, por lo que no se exigiría una metacognición ni una sofisticada conciencia artificial. Al mismo tiempo, este tipo de atribución presentaría limitaciones explicativas y predictivas con respecto a las otras perspectivas, ya que su objeto es dar sentido inmediato en el momento de la interacción, sin pretender ir más allá. En todo caso, tal y como sucede entre los humanos, la combinación de las estrategias aportadas por la atribución mental en las tres perspectivas sería la forma más completa de comprender la mente del otro agente en interacción[8].

V. Conclusiones

Asumiendo cautelosamente la posibilidad técnica y conceptual de una IA-fuerte con capacidades cognitivas similares a las humanas, que era el presupuesto de partida de este ensayo, podemos suponer que algún tipo de atribución psicológica mutua sea plausible. Desde el punto de vista cronológico, la atribución en perspectiva de segunda persona probablemente sería la primera en aparecer, tanto en su creación como en

8. Jugando a especular —permítaseme la licencia, llegados a este punto de este ensayo ya de por sí altamente especulativo—, quien sabe si en un futuro lejano no habrá libros de referencia en psicología (sobre psicología de la personalidad, psicología social, psicología de las diferencias individuales, psicología del pensamiento, psicopatología, etc) que tengan apartados diferenciados, destinados a describir, explicar y predecir no sólo la mente y conducta de los humanos, sino también la de agentes artificiales. Si algún día llegara a darse esta posibilidad, no sólo habría que construir una nueva epistemología que contemple a la máquina y su interacción con nosotros, pues, como se puede imaginar, también habría que reformular todos los aspectos pragmáticos de nuestra vida, en política, economía, ética, derecho, etc. De momento, únicamente la ciencia-ficción se toma en serio esta reflexión.

su desarrollo ontogenético, puesto que es la más inmediata y la que exige un menor coste cognitivo por parte de la IA. En este sentido, la creación de robots sociales o cuidadores sería una primera forma de materializar el intercambio cognitivo-emocional entre humanos y máquinas inteligentes. Con el advenimiento de la IA-general, a decir del transhumanismo tecnológico y su postura funcionalista, el siguiente paso sería la atribución en perspectiva de tercera persona, toda vez que una teoría cognitiva global puede dar cobertura descriptiva, explicativa y predictiva tanto a la máquina como al humano. Y, quizá, la capacidad de atribución mental en perspectiva de primera persona sería la última en desarrollarse, no tanto por su complejidad óntico-epistémica como por la dificultad en la creación del entramado empático-social que sin duda contribuiría a ese «ponernos en el lugar del otro». Para conseguir este increíble hito (razonable para unos, imposible para otros, en todo caso improbable a corto-medio plazo), antes habrían de conseguirse profundos cambios tecnológicos, con máquinas capaces de sentir emociones de forma genuina, y, sin duda también, habrían de alcanzarse trascendentes cambios sociales, necesarios para comprender y aceptar un mundo nuevo con dos tipos de criaturas sensibles, inteligentes y libres.

Referencias bibliográficas

Armstrong, D. (1981): *The Nature of Mind*. Cornell University Press.

Avramides, A. (2020): *Other Minds*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/win2020/entries/other-minds/

Bex, F., Prakken, H., Engers, V., & Verheig, B. (2017): "Introduction to the Special Issue on Artificial Intelligence for Justice" (AI4J), en *Artificial Inteligence and Law.* Springer, 25, pp. 1-3. https://doi.org/10.1007/s10506-017-9198-5

Boden, M. A.: (2017): *Inteligencia Artificial*. Turnes Publicaciones.

Caicedo, O.D. (2017): «¿Pueden pensar los animales no humanos? Algunas consideraciones en defensa del antropomorfismo científico», *Ludus Vitalis*, vol. XXV, 48, pp. 181-208.

Carruthers, P. (2000): *Phenomenal Consciousness*. Cambridge University Press.

Chalmers, D. (1995): "Facing up to the Problem of Consciousness", *Journal of Consciousness Studies*, 2, pp. 200-219.

Cortina, A. (2011): *Neuroética y neuropolítica. Sugerencias para la educación moral*. Tecnos.

Davidson, D. (1974): *Psychology as Philosophy*: *Essays on Actions and Events*. Oxford Scholarship Online.

De Waal, F. (2007): «Seres moralmente evolucionados», en *Primates y filósofos. La evolución de la moral del simio al hombre.* Paidós, pp. 23-111.

Diéguez, A. (2017): *Transhumanismo. La búsqueda tecnológica del mejoramiento humano*. Herder Editorial.

Evers, K. (2011): *Neuroética. Cuando la materia se despierta*. Katz Editores.

Fodor, J. (1981): «El problema mente-cuerpo», *Investigación y Ciencia*, 54.

Glock, H. J. (2009): *La mente de los animales: problemas conceptuales*. HRK.

Gomila, A. (2002): «La perspectiva de segunda persona de la atribución mental», *Azafea. Revista de filosofía*, 4, pp. 123-138. https://doi.org/10.14201/3719

Gomila, A. (2016): «La perspectiva de segunda persona: mecanismos mentales de la intersubjetividad», *Contrastes. Revista Internacional de Filosofía*, pp. 65-86. https://doi.org/10.24310/Contrastescontrastes.v0i0.1448

Gomila, A., & Pérez, D. (2018): "Mental Attribution in Interaction: How the Second Person Perspective Dissolves the Problem of Other Minds", *Daimon. Revista Internacional de Filosofía*, 75, pp. 75-86. https://doi.org/10.6018/daimon/332611

Gopnik, A. (1993): "How we Know our Minds: The Illusion of First-person Knowledge of Intentionality", *Behavioral and Brain Sciences*, 16(1), pp. 1-14. https://doi.org/10.1017/S0140525X00028636

Gordon, R. (1992): "The simulation theory", *Mind and Language*, 7, pp. 11-34. https://doi.org/10.1111/j.1468-0017.1992.tb00195.x

Hameroff, S. (1998): "Quantum Computation in Brain Microtubules? The Penrose-Hameroff 'Orch OR' model of consciousness", *Philosophical Transactions Royal Society London*, A 356, pp. 1869-96. https://doi.org/10.1098/rsta.1998.0254

Hamet, P., & Tremblay, J. (2017): "Artificial Intelligence in Medicine", *Metabolism*, 69S: S36-S40. https://dor.org/10.1016/j.metabol.2017.01.011

Hanson, R. (2017): «Cuando los robots gobiernen la Tierra: el legado humano», en *El próximo paso: la vida exponencial*. OpenMind BBVA. https://www.bbvaopenmind.com/articulos/cuando-los-robots-gobiernen-la-tierra-el-legado-humano/

Howard, J. (2019): "Artificial Intelligence: Implications for the Future of Work", *American Journal of Industrial Medicine*, 62(11), pp. 917-926. https://doi.org/10.1002/ajim.23037

Kurzweil, R. (2012): *La singularidad está cerca. Cuando los humanos trascendamos la biología*. Lola Books.

Latorre, J. L. (2019): *Ética para máquinas*. Ariel.

Levy, N. (2014): *Neuroética. Retos para el siglo XXI*. Avarigani Editores.

López de Mántaras, R., & Meseguer, P. (2017): *Inteligencia Artificial*. CSIC.

Moravec, H. (1988): *Mind Children: The Future of Robot and Human Intelligence*. Cambridge University Press.

Nagel, T. (1974): "What is it Like to be a Bat?", *Philosophical Review*, 83, pp. 435-456.

Newell, A. y Simon, H.A. (1976): "Computer Science as Empirical Enquiry: Symbols and Search", *Communications of the Association for Computing Machinery*, 19.

Pakusch, C., Stevens, G., Boden, A., & Bossauer, P. (2018): "Unintended Effects of Autonomous Driving: A Study on Mobility Preferences in the Future", *Sustainability*, 10, 2404. https://doi.org/10.3390/su10072404

Pérez, D., & Gomila, A. (2021): *Social Cognition and the Second Person in Human Interaction*. Routledge.

Picarra, N., Giger, J. C., Pochwatko, G., & Goncalves, G. (2016): "Making Sense of Social Robots: a Structural Analysis of the Layperson's Social Representation of Robots", *European Review of Applied Psychology*, vol. 66(6), pp. 277-289. https://doi.org/10.1016/j.erap.2016.07.001

Putnam, H. (1960): "Minds and Machines", en Hook (ed.), *Dimensions of Minds.* New York University Press, pp. 138-164.

Putnam, H. (1967): "Psychological predicates", en Capitan W.H. y Merrill D. D. (eds.), *Art, Mind, and Religion.* University of Pittsburgh Press, pp. 37-48.

Putnam, H. (1981): *Reason, Truth and History*. Cambridge University Press.

Russell, S., & Norvig, P. (2009): *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Searle, J. R. (1980): "Minds, Brains and Programs", *Behavioral and Brain Sciences*, 3(3), pp. 417-457.

Tegmark, M. (2018): *Vida 3.0*. Editorial Taurus.

Tirapu-Ustárruz, J., Pérez-Sayes, G., Erekatxo-Bilbao, M., & Pelegrín-Valero, C. (2007): «¿Qué es la teoría de la mente?», *Revista de Neurología*, 44(8), pp. 479-489.

Turing, A. (1950): "Computing Machinery and Intelligence", *Mind*, 236(59), pp. 433-460.

Van Gulick, R: (2004): *Consciousness*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/spr2018/entries/consciousness/

Vietri, M., Alessandroni, N., & Piro, M. C. (2019): «La perspectiva de segunda persona de la atribución de estados mentales: una revisión sistemática de su estado actual de desarrollo», *Psykhe*, 28(2), pp. 1-17. http://dx.doi.org/10.7764/psykhe.28.2.1280

Wang, W., & Siau, K. (2019): "Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work and Future of Humanity: A Review and Research Agenda", *Journal of Database Management,* 30(1), pp. 61-79. https://doi.org/10.4018/JDM.2019010104

Warwick, K., & Sha, H. (2017): «El futuro de la comunicación humano-máquina: el test de Turing», en *El próximo paso: la vida exponencial.* OpenMind BBVA. https://www.bbvaopenmind.com/articulos/el-futuro-de-la-comunicacion-humano-maquina-el-test-de-turing/

José Miguel Biscaia Fernández: Licenciado en Neurobiología (UCM), Doctor en Neurociencia (UCM) y Graduado en Filosofía (UNED). Actualmente trabajo como Investigador y Profesor Titular en la Universidad Europea de Madrid (España).

Líneas de investigación:

– Neurociencia, Ética Aplicada y Filosofía de las Ciencias Cognitivas.

Publicaciones recientes:

– Biscaia, J. M. Mohedano RB. (2021). Cerebros, mentes y robots: una aproximación a través del cine del siglo XXI. *Revista de Medicina y Cine* 17(1): 49-56.

– Biscaia, J. M. (2021). 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: 1-17.

– Biscaia, J. M. (2021). De las emociones naturales a la emocionalidad artificial. *Cuadernos Salmantinos de Filosofía* 48: 105-131.

– Biscaia, J. M. (2022). La gran pantalla como laboratorio y espejo para la roboética. *Journal de Ética y Cine* 12(29): 55-69.

– Biscaia, J. M., & Mohedano, R. (2022). Neuroética en fotogramas. *Revista de Medicina y Cine* 18(3): 249-258.

– Biscaia JM, Mohedano RB, Biscaia CJ. (2023). La inteligencia artificial en la prevención de conductas suicidas: aspectos técnicos y consideraciones ético-legales. *Revista de Bioética y Derecho*, 59: 181-203.

– Biscaia JM, González-Soltero MR, Biscaia CJ, Mohedano RB, Rodríguez-Learte, AI. (2024). Empleo de ChatGPT en educación biomédica. Análisis de riesgos desde los Principios Éticos de la UNESCO y el Reglamento de la Unión Europea sobre Inteligencia Artificial. R*evista Iberoamericana de Bioética*, 25: 1-15.

Correo-e: josemiguel.biscaia@universidadeuropea.es