El test de Turing en *Ex Machina*: ¿Es Ava un sistema intencional?

Ex Machina | Alex Garland | 2014

María Paola Caycedo-Castro; Boris Julián Pinto-Bustamante*

Universidad el Bosque, Colombia
Universidad del Rosario, Colombia

Recibido 01 de junio de 2022; aprobado 28 de junio de 2022

Resumen

El presente artículo aborda el problema de la interacción entre seres humanos y la inteligencia artificial a partir del test de Turing representado en la película *Ex Machina*. Se revisan algunas de las principales perspectivas en relación con la filosofía de la mente y se analiza el problema que plantea la película, en cuanto a si Ava, la robot humanoide, supera el test de Turing, a partir de tres hipótesis: los estados mentales descritos por John Searle constituyen las tres dimensiones del mundo psíquico y la deliberación moral: la dimensión fáctica de la *percepción*, la dimensión estimativa de los valores y la dimensión pragmática de las acciones. La segunda hipótesis afirma que las ideas constituyen estados mentales intencionales, unidades de trasmisión cultural que, asociadas a estados emocionales, constituyen los sentimientos como fenómenos neuroculturales. La tercera hipótesis afirma que la comunidad de significados emergente a partir del vínculo entre afectos y sistemas simbólicos-culturales puede estar constituida, tanto por seres humanos como por sistemas algorítmicos. En este sentido, la prueba de la inteligencia artificial fuerte consiste en la constatación en tales sistemas de estados mentales afectivos (valorativos) mediados por la interacción cultural.

Palabras Clave: Bioética | Inteligencia Artificial | Películas Cinematográficas | Teoría de la Mente | Emociones

The Turing test in Ex Machina: Is Ava an intentional system?

Abstract

This article addresses the problem of the interaction between human beings and artificial intelligence based on the Turing test represented in the movie *Ex Machina*. Some of the main perspectives in relation to the philosophy of mind are reviewed and the problem posed by the film is analyzed, as to whether Ava, the humanoid robot, passes the Turing test, based on three hypotheses: The mental states described by John Searle constitute the three dimensions of the psychic world and moral deliberation: the factual dimension of *perception*, the estimative dimension of values, and the pragmatic dimension of actions. The second hypothesis affirms that ideas constitute intentional mental states, units of cultural transmission that, associated with emotional states, constitute feelings as neurocultural phenomena. The third hypothesis states that the community of meanings emerging from the link between affects and symbolic-cultural systems can be made up of both human beings and algorithmic systems. In this sense, the proof of strong artificial intelligence consists in the verification in such systems of affective (evaluative) mental states mediated by cultural interaction.

Keywords: Bioethics | Artificial Intelligence | Motion Pictures | Theory of Mind | Emotions

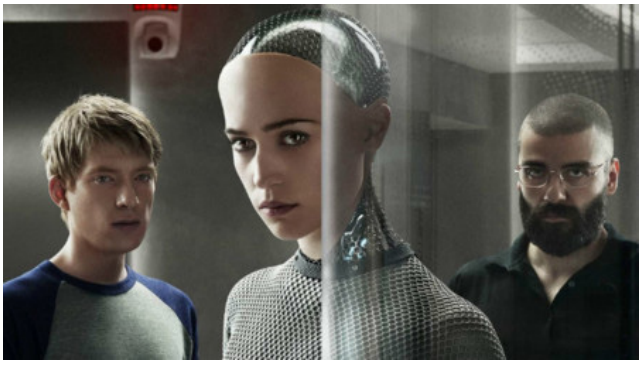
Introducción

La relación mente-cuerpo ha estimulado múltiples cuestionamientos para la filosofía de la mente, la antropología, la bioética y las ciencias cognitivas, como escenarios transdisciplinares que abordan la naturaleza de los fenómenos mentales (cognición, deseos, percepción, volición, imaginación, etc.) (Uribe, 2002, p. 271) y sus relaciones con otros fenómenos empíricos, como el cuerpo y los productos culturales, tanto técnicos como simbólicos, de lo cual emerge una cuestión relevante para la práctica de la psiquiatría, la psicología, la neurología y la ingeniería de inteligencias artificiales: según la noción que aceptemos sobre la naturaleza de los fenómenos mentales, dependerán los modos en que nos relacionamos con las entidades inteligentes humanas y no humanas, así como las prácticas efectivas en salud mental y en la relación con los desarrollos algorítmicos y robóticos de la inteligencia artificial.

* pintoboris@unbosque.edu.co

Ex Machina es una película de ciencia ficción de origen británico, escrita y dirigida por Alex Garland, la cual reitera una pregunta recurrente de la filosofía de la mente: ¿Pueden las máquinas creadas por el ser humano, pensar y sentir?, ¿cómo determinar si una máquina piensa y actúa de manera inteligente?, ¿puede una entidad artificial desplegar funciones cognitivas autoconscientes?

Caleb Smith, un joven programador que trabaja en la compañía de internet más grande del mundo (*Blue Book*), es elegido en un concurso creado por el CEO, Nathan Bateman, para pasar una semana con él, donde participará en un experimento para determinar si Ava, una robot humanoide, supera el test de Turing, exhibiendo una inteligencia análoga a la inteligencia humana.



A lo largo del artículo intentaremos articular la narrativa de *Ex Machina* con algunas de las propuestas más relevantes dentro de la filosofía de la mente y la inteligencia artificial; posteriormente abordaremos, desde la neurobiología de los afectos, la pregunta por la posibilidad de que una máquina, en este caso la protagonista de la película, pudiese desplegar estados funcionales afectivos e intentaremos, por último, responder a un cuestionamiento de interés especial para la bioética: ¿puede un robot ser capaz de sentir y valorar?, ¿tendría la capacidad de emitir juicios de valor?

Del Dualismo De Sustancias Al Funcionalismo.

El problema mente-cuerpo ha sido abordado desde posturas dualistas y monistas. René Descartes formuló una respuesta concreta que remite a la noción del dualismo platónico, el cual divide el mundo entre las cosas permanentes (*eidos*) y verdaderas (*aletheia*), y las cosas pasajeras o efímeras (*doxa*), así como a la división del mundo sensible y corruptible (el cuerpo) y el mundo de las ideas e incorruptible (alma racional).

El dualismo cartesiano se entiende como dualismo de sustancias, al trazar una dicotomía entre los dos componentes de todo lo existente (a excepción de Dios): *la res extensa* (como toda sustancia material gobernada por las leyes de la mecánica, destructible, determinada, divisible y asequible mediante el conocimiento empírico-racionalista) y la *res cogitans* (la sustancia mental indestructible, indivisible, reducto de libertad, fuente de todo conocimiento, no gobernada por las leyes de la mecánica, asequible mediante las meditaciones filosófico-religiosas).

El pensamiento cartesiano intentaba dilucidar la relación causal entre mente y cuerpo al preguntarse: ¿cómo algo físico puede producir efectos en el alma, y cómo eventos del alma pueden afectar el mundo físico? Descartes encuentra dicha conexión en la glándula pineal, a la que atribuyó el vínculo entre lo que sentimos, y pensamos, así como entre el cuerpo físico y la mente (Searle, 2006).

No obstante, el paradigma cartesiano ha sido desafiado en un intento por superar el dualismo de sustancias o la dicotomía mente/cuerpo. Wittgenstein (1988) aborda el problema de la mente a partir de la relevancia que la comprensión de los juegos del lenguaje representa para la vida cotidiana, desde sus características de intersubjetividad, intencionalidad y normatividad (esto es, los enunciados lingüísticos tienen un carácter social orientado a finalidades prácticas restringidas) (Pérez, 2006). Para Wittgenstein, el problema de la mente es un asunto gramatical, más que científico, ontológico o epistemológico. Toda referencia a lo mental no es referencia a un objeto en sí, sino una expresión psicológica que procura comunicar un estado interno con miras a cierta finalidad. En *Ex Machina* aparece una referencia a Wittgenstein, ya que Nathan llama a su gran empresa *Blue Book*, en honor a una de las obras de dicho autor: *El Libro Azul*, cuyas notas, de 1933 a 1934, tratan sobre dichos juegos de lenguaje.

Para Ryle (2005) existe una incompatibilidad entre nuestra comprensión de lo mental y lo físico, en lo que radica el error categorial del "fantasma en la máquina": si la mente es inmaterial, ¿cómo puede entonces animar a la materia? Para Pérez (2006) es claro que "la filosofía de la mente se ha concentrado en el problema de la explicación científica de lo mental y por eso ha privilegiado los métodos empíricos de investigación, sin embargo, ha desatendido el problema de nuestra comprensión cotidiana de lo mental" (p. 390).

Dentro de la filosofía analítica contemporánea, las perspectivas monistas se oponen al dualismo de sustan-

cias al afirmar que sólo existe una sustancia. Dentro del monismo epistemológico se cuentan el materialismo e idealismo. Este último, asevera que el universo es una creación mental, lo que concebimos como mundo físico es, en realidad, una de las adaptaciones de nuestra realidad mental (Searle, 2006).

El monismo materialista o fisicalista sostiene que "el ser humano, incluyendo la mente, no es más que materia y complejas propiedades físicas" (Van Oudenhove y Cuypers, 2010, p. 547), según lo cual, los fenómenos mentales se reducen a sus propiedades naturales constitutivas. Los críticos del materialismo sostienen que esta perspectiva no resuelve el problema en torno a uno de los fenómenos más importantes para los filósofos de la mente: el "problema difícil" como la plantea Chalmers (1995), la pregunta por la experiencia, o la noción de *qualia* (*cualios*: las cualidades subjetivas de experiencias individuales o las propiedades fenomenológicas de los estados mentales conscientes).

El materialismo eliminativo, representado por autores como Paul Churchland y Patricia Chuchland (2012), considera la inexistencia de las propiedades mentales, las cuales, a diferencia de los procesos neuronales, no pueden comprobarse empíricamente. Para esta postura, la noción de *qualia* es irrelevante, al tiempo que la pregunta por la naturaleza de los procesos neurofisiológicos debe desplazarse desde la psicología (a la que consideran *folk science*) a la neurobiología y las neurociencias computacionales.

La Teoría de la Identidad Mente-Cerebro, propuesta por autores como Ullin Place y John Smart, surge hacia los años 50 del siglo XX como una réplica al conductismo, el cual proponía la verificación de los procesos mentales como conductas observables, dada la inasequibilidad de los estados mentales internos. La Teoría de la Identidad postula un reduccionismo ontológico, según el cual, los tipos de estados mentales son idénticos a los tipos de estados cerebrales. En este sentido, la "mente es solo cerebro y lo que imaginamos como estados mentales son solo estados cerebrales" (Searle, 2006).

El argumento de la realización múltiple propuesto por Putnam (1967) desafía la teoría de la identidad, al afirmar que los estados mentales no corresponden siempre a los mismos estados cerebrales. Distintos estados cerebrales pueden realizar estados mentales concretos, de donde se sigue que los estados mentales constituyen individualizaciones de diversos estados cerebrales (Putnam, 1967).

A partir de esta objeción, Putnam plantea una postura funcionalista, según la cual los estados mentales corresponden a estados funcionales. Block (1996), afirma que la relación entre estados, *inputs* sensoriales y *outputs* conductuales determina la naturaleza de los estados mentales, según lo cual un estado mental es un estado autómata que puede desplegarse en sustratos cerebrales, informáticos o algorítmicos de cualquier soporte material. Así, se representa lo mental en términos no mentales, como la cuantificación de estados funcionales realizables y como estados internos, en los cuales la única interacción con el entorno es la sensopercepción de estímulos (*inputs*) y la modificación de estados funcionales con disposiciones conductuales (*outputs*). Como el mismo Putnam lo reconoce, el funcionalismo tampoco satisface la pregunta por la distinción entre ideas y estados funcionales ni por la experiencia o la vivencia subjetiva de los fenómenos (cualios) (Lorenzatti, 2018; Putnam, 1991).

Dualismo de propiedad e inteligencia artificial

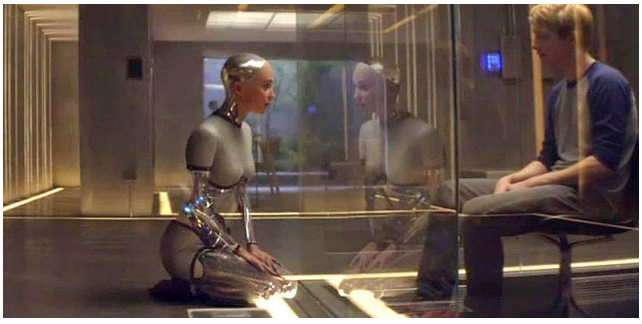
Una de las corrientes que intenta reconciliar esa separación mente-cuerpo es el dualismo de propiedad, el cual afirma que "sólo la sustancia física existe. Sin embargo, ésta puede tener propiedades tanto físicas, como mentales, las que son ontológica y radicalmente diferentes entre sí, sin ser reducible lo mental a lo físico" (Van Oudenhove y Cuypers, 2010, p. 550).

Otras vertientes intentan superar las limitaciones del materialismo eliminativo y reduccionista: el epifenomenalismo define los procesos mentales como subproductos de propiedades físicas o fenómenos primarios (p. ej. los procesos neurobiológicos), frente a los cuales, los epifenómenos son dependientes e incapaces de modificar a los primeros. El emergentismo (propuesto por autores como Nagel, Bateson y Chalmers), por su parte, supone que la realidad mental (p. ej. estados conscientes, percepción, sentimientos, imaginación) constituye una estructura emergente a partir del mundo natural, en tanto constituida por un nivel de organización en el que las propiedades de segundo orden de la experiencia fenoménica (p. ej. conciencia) son más complejas que las propiedades de primer orden de sus partes neurobiológicas constituyentes (Nagel, 1986), si bien los procesos mentales son supervinientes a los procesos biológicos subvinientes (esto es, sin la base física neurobiológica, no es posible la emergencia del proceso mental) (Braun, 2011), preservando a su vez la autonomía ontológica de los fenómenos mentales, no reducibles a las propiedades físi-

cas de primer orden, ni a la disección y análisis de sus partes constitutivas (Eronen, 2004, p. 21).

Desde mediados del siglo XX irrumpen nuevas propuestas teóricas desde las ciencias computacionales, en concordancia con el auge tecnológico del momento y el funcionalismo descrito anteriormente. John McCarthy acuña el término Inteligencia Artificial (IA) en 1956, proponiendo una analogía: la mente es al cerebro lo que el software es al hardware. Esta vertiente computacional afirma que cualquier sistema físico dotado del programa de instrucciones correcto, con los *inputs* y los *outputs* adecuados, estaría dotado, a su vez, de una mente (Searle, 2006), según lo cual, la inteligencia, la conciencia, las creencias, entre otros fenómenos mentales, son el producto de la manipulación de símbolos físicos.

El filósofo John Searle discrepa de esta analogía, al señalar dos vertientes dentro de la IA: la IA débil, la cual pretende simular estados mentales sin aspirar a que un computador despliegue conciencia, y la IA fuerte que, por el contrario, afirma la posibilidad de diseñar una computadora con autoconciencia y singularidad, para lo cual el Test de Turing, o el juego de imitación (Turing, 1950), constituye una herramienta de evaluación del nivel de IA, el cual supone un juego de comunicación entre interrogadores humanos e interlocutores ocultos (humanos y máquinas). Tras breves interrogatorios de máximo cinco minutos, el interrogador debe descubrir cuáles entidades son humanas y cuáles son máquinas. La máquina que logre engañar al entrevistador habrá superado el test de Turing.



En *Ex Machina*, Nathan propone superar el test de Turing al crear el prototipo Ava, para lo cual Caleb representa el componente humano de la prueba, quien debe reconocer las facultades conscientes y autoconscientes de AVA, a pesar de percibir un robot humanoide. En relación con esta corriente computacional, Penrose (1996) afirma:

> los dispositivos no sólo son inteligentes y tienen una mente, sino que al funcionamiento lógico de cualquier dispositivo computacional se le puede atribuir un cierto tipo de cualidades mentales, ya que solo consiste en una secuencia bien definida de operaciones. Para ellos lo que cuenta es simplemente el algoritmo. No hay ninguna diferencia si el algoritmo es ejecutado por un cerebro, una computadora electrónica, una nación entera de hindúes, un dispositivo mecánico de ruedas y engranajes o un sistema de tuberías. Es simplemente la estructura lógica del algoritmo lo significativo del estado mental que se supone representa, siendo completamente irrelevante la encarnación física de dicho algoritmo. (p. 27)

El Naturalismo Biológico es la culminación de años de trabajo de John Searle. El autor propone una crítica a los modos del lenguaje con que categorizamos nuestra mente y cuerpo, deudores del dualismo cartesiano, así como afirma su crítica a la IA fuerte, ya que para el autor, nunca un programa de computador podrá ser considerado una mente, ya que el programa solo puede operar en el plano sintáctico de la información, mientras que la mente despliega contenidos semánticos, por lo que los estados mentales no corresponden llanamente a la manipulación de símbolos, sino que deben desplegar la capacidad de interpretarlos (Searle, 2006).

En *Ex Machina*, Caleb, al tener el primer contacto con Ava, la encuentra fascinante, pero aún no está seguro si constituye IA consciente. En un primer encuentro, Ava representa un cúmulo de símbolos e información que sólo simula albergar consciencia.

Searle refina su crítica con el experimento mental de la habitación China, un experimento mental popularizado por Roger Penrose, que intenta rebatir la validez del Test de Turing, a la vez que plantea que una máquina es incapaz de llegar a pensar. Expone la diferencia entre reconocer la sintaxis y comprender la semántica, proponiendo que un intérprete en una habitación cerrada, dotado con los repertorios y la suficiente cantidad de reglas para procesar la información entrante (p. ej. símbolos lingüísticos en chino), puede hacerse pasar por un intérprete humano, si reparamos exclusivamente en la dimensión sintáctica del lenguaje, sin considerar la dimensión semántica del sentido: "Los objetos biológicos (cerebros) pueden poseer "intencionalidad" y "semántica", lo que dicho autor considera como las características definitorias de la actividad mental" (Penrose, 1996, p. 28).

Si bien ya se han constatado máquinas que superan el test de Turing, como lo demuestran las sesiones experimentales en Bletchley Park (2012) y en London Royal Society (2014), y actualmente tiene lugar un debate a propósito de que un *chatbot* desarrollado por la empresa Google haya desarrollado sintiencia y conciencia moral

(Tiku, 2022), quizás el test de Turing, según su formulación inicial, podría no constituir una prueba suficiente para evaluar el grado de interacción humano-máquina o la sofisticación de los sistemas algorítmicos (Alfonseca, 2014), dado que se suelen soslayar diversos contenidos semánticos implícitos en las prácticas comunicativas de las comunidades humanas, "como la mentira, el malentendido, la falta de conocimientos y el humor, sin olvidarnos de la estupidez" (Warwick & Shah, 2016, p. 2), además de la polisemia cultural y afectiva de los contenidos discursivos.

Un paso más allá: la mente extendida

Otras corrientes de la filosofía de la mente atribuyen un papel relevante al entorno en los procesos cognitivos: "no podemos simplemente señalar la piel o el cráneo como límite cognitivo para justificarnos, pues la legitimidad de ese límite está precisamente en cuestión" (Clark y Chalmers, 2011, p. 16).

Esta perspectiva afirma que la cognición constituye un sistema ensamblado, un acoplamiento de funciones neuronales internas y la manipulación activa de recursos extrasomáticos, como lo son el lenguaje, las prótesis, la cultura, según lo cual, las funciones cerebrales dependen de soportes ambientales que, a su vez, están incorporados en las representaciones funcionales endocerebrales. La tesis de la mente extendida, defendida desde la noción del externalismo activo, afirma que los objetos, espacios, prótesis y símbolos del entorno desempeñan un papel activo en la configuración de los procesos cognitivos (Clark y Chalmers, 2011).

Otros desarrollos de las ciencias cognitivas desafían la noción de cognición como simple manipulación de un conjunto de representaciones simbólicas en una mente descorporeizada y algorítmica, para revitalizar el papel de los procesos corporales, afectivos, extracorporales y simbólicos en la constitución de la cognición, como un circuito continuo entre cuerpo, mundo y significados, según lo cual, la cognición corresponde a "la creación de significados por un agente corporal, a partir de sus interacciones con el medio ambiente" (Wright-Carr, 2018, p. 83), a una experiencia fenomenológica resultado de las interacciones dinámicas entre agente corporal y su entorno. Así, la mente, más que un epifenómeno de funciones neuronales o un algoritmo representacional, constituye una mente extendida (extended) y embebida (embedded) en el mundo (Clark, 2008), enactiva (enacted) (Klin et al., 2003) con el mundo y encarnada (embodied) (Sheets-Johnstone, 2008) desde el cuerpo.

Afectos y valores: los fundamentos neuroculturales de la conciencia

Para Searle, la conciencia es un fenómeno neurofisiológico que consiste en el conjunto de estados (procesos o eventos, etc.) mediante los cuales sentimos y percibimos (Searle, 2015). Estos estados o procesos mentales constitutivos de la conciencia permiten que *algo* pueda ser considerado *alguien*, en virtud de quien puede estimar el valor (o la importancia) de las cosas percibidas por los estados mentales conscientes. Estos procesos mentales permiten la relación de *alguien*, a través de la representación (la capacidad de representar fenómenos y objetos) y los modos de relación con el mundo sensible. Según Searle (2001) esta relación con el mundo es posible gracias a estados mentales como las creencias, los deseos, las sensaciones y los pensamientos, los cuales se caracterizan por un rasgo intrínseco: la intencionalidad, la cual constituye un fenómeno de la conciencia, según el cual, los estados mentales tienen la propiedad de ser referidos a otros objetos distintos a sí mismo.

La primera de nuestras hipótesis afirma que estos estados mentales descritos por Searle constituyen las tres dimensiones del mundo psíquico y la deliberación moral, como lo ha propuesto Gracia (2011): la dimensión fáctica de la *percepción* (que permite la constatación sensible de las propiedades físicas de las cosas); la dimensión estimativa de los valores (cualidades afectivas atribuidas por *alguien* a las cosas); y la dimensión pragmática de las acciones (la intencionalidad de los actos motivados por la estimación previa de los hechos sensibles).

Los estados mentales (creencias, deseos, esperanzas y temores) constituyen experiencias cualitativas subjetivas (*cualios*/qualia) de orden afectivo. Damasio (2009) define, por ejemplo, la esperanza como un afecto (*affectus*), y recurre a la definición de Spinoza: "La esperanza no es otra cosa que una alegría inconstante, que surge de la imagen de algo futuro o pasado, de cuyo resultado en cierta medida dudamos" (Spinoza, 1994, citado en Damasio, 2009, p. 139).

Según Damasio, Spinoza emplea el término *afecto* como una categoría que engloba, tanto a las emociones como a los sentimientos, como modificaciones oscilantes del cuerpo y las ideas adscritas a tales modificaciones durante el proceso de *ser afectado*. Para Spinoza, la cate-

goría de los afectos constituye un conjunto de "impulsos, motivaciones, emociones y sentimientos" (Damasio, 2009, p. 279).

A partir de esta nomenclatura se propone un "principio de anidación" de los procesos regulatorios orientados a la supervivencia y el bienestar (Damasio, 2009, p. 47), los cuales se hacen más refinados en la medida en que progresa la complejidad evolutiva de los organismos, desde las reacciones homeostáticas básicas (metabólicas, físicoquímicas, reflejas), pasando por los comportamientos asociados al sistema homeostático de recompensa (búsqueda y evitación, en función de percepciones de placer y dolor), el despliegue de instintos (o impulsos) y motivaciones (el apetito –como el comportamiento relativo al impulso–, y el deseo –el sentimiento consciente relativo al instinto–), como la expresión de emociones y sentimientos.

Las emociones constituyen un "conjunto complejo de respuestas químicas y neuronales que forman un patrón distintivo" (Damasio, 2009, p. 55) ante la presencia de estímulos emocionalmente competentes (reales, rememorados o imaginados), los cuales modifican temporalmente los estados corporales y la cartografía neuronal relacionada con ellos. Las emociones son representaciones de los estados corporales cuyo objetivo es disponer tales estados en función de un comportamiento adaptativo para responder a los estímulos y demandas del entorno (Prinz, 2007, p. 54). Este conjunto de respuestas neuroquímicas puede adoptar distintas categorías: emociones de fondo (estados emocionales), emociones innatas (alegría, tristeza, repugnancia, ira, sorpresa, temor) y emociones sociales, mediadas por el aprendizaje social (vergüenza, culpa, indignación, pudor, celos, etc.) (Damasio, 1999).

Damasio (2009) traza una distinción entre las emociones y los sentimientos, a los cuales define como: "la percepción de un determinado estado del cuerpo junto con la percepción de un determinado modo de pensar y de pensamientos con determinados temas" (p. 88). Los sentimientos se configuran entonces a partir de una *idea* (o un conjunto de *ideas*) asociadas a la modificación emocional de los estados corporales (Damasio, 2009). Las ideas (y los pensamientos) son imágenes mentales derivadas de un conjunto de patrones de actividad neuronal (cartografías sensoriales) que permiten la representación de los estados corporales en relación con objetos del mundo externo que interactúan con el organismo y que cumplen diversas funciones reguladoras (Damasio, 2009, p. 185). En tal sentido, las ideas son categorías análogas a los estados mentales intencionales propuestos por Searle.

Podemos refinar un poco esta descripción, y aquí planteamos una segunda hipótesis, afirmando que una idea es un estado mental intencional, por medio del cual la conciencia representa el mundo y despliega las condiciones de posibilidad para relacionarse con él. La conciencia no es solo un atributo autorreferencial, pues como afirma Roger Bartra (2007):

Podemos entender la conciencia como una serie de actos humanos individuales en el contexto de un foro social y que implican una relación de reconocimiento y apropiación de hechos e ideas de las cuales el yo es responsable. La manera en que Locke ve a la conciencia se acerca más a las raíces etimológicas de la palabra: conciencia quiere decir conocer con otros. Se trata de un conocimiento compartido socialmente. (p. 13)

En tal sentido, las ideas no constituyen simples representaciones mentales, sino "unidades de trasmisión cultural" (Bartra, 2007, p. 102), símbolos y motivos que, asociados a los estados emocionales, configuran los sentimientos.

En este punto nos separamos de la noción que defiende de Damasio (2009), en cuanto a la naturaleza privada de los sentimientos. Si bien concordamos en que las emociones se despliegan en el teatro del cuerpo, mientras los sentimientos lo hacen en el teatro de la mente, no concordamos con la afirmación que remite los sentimientos al ámbito privado de la conciencia, pues las percepciones y estimaciones de la vida emocional conservan la posibilidad de ser comunicados y son, a su vez, un producto de la comunicación. Su papel no se reduce a la dimensión de razonabilidad asociada a los mecanismos emocionales automáticos para la gestión de problemas complejos. Los sentimientos, como nociones comunicables, son a la vez fundamento y producto de la red simbólica que constituye la cultura, como conciencia colectiva de sentimientos compartidos. La noción de emociones sociales, como la vergüenza, el pudor, la culpa, los celos y su relación con emociones primarias y emociones de fondo, sólo pueden ser comprendidas en función de la existencia de símbolos culturales de carácter normativo (nociones como el pecado, el castigo, la redención, etc.), presentes en circuitos exocerebrales de significado social (Bartra, 2007).

Al referirse John Searle a la dimensión semántica del lenguaje, entendida como la atribución de significado a la sintaxis de los símbolos lógico-formales, cabe precisar, desde una perspectiva antropológica, que la sintaxis, como los significados, "se construyen en una red que conecta circuitos neuronales con redes culturales" (Bartra, 2007, p. 129). Desde la perspectiva de Susanne Langer, el significa-

do connota una función simbólica del mundo, y el símbolo equivale a "cualquier recurso por medio del cual estamos facultados para hacer una abstracción" (Langer, 1953, p. xi). Sin el comercio de significados no es posible la instauración de comunidades humanas (Cassirer, 1963).

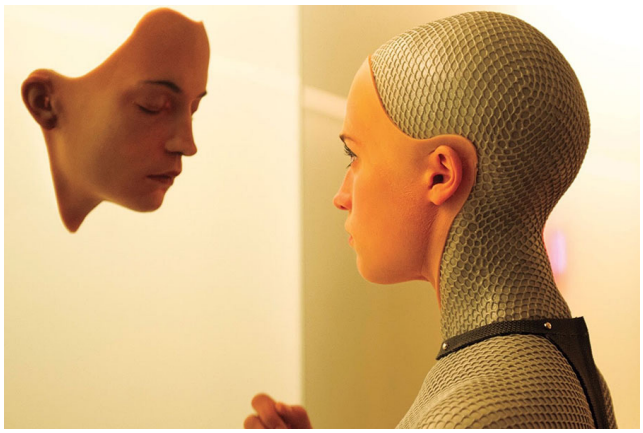
En este punto es posible, como una tercera hipótesis, vincular afectos y sistemas simbólico-culturales (como el lenguaje, los ritos, la música, la danza, los mitos, etc.) en la emergencia de la comunidad que articula significados, una comunidad que puede ser constituida por animales humanos y por sistemas algorítmicos dotados de estados mentales intencionales. Las emociones sociales y los sentimientos se configuran a partir de un proceso continuo de retroalimentación entre señales neuronales y símbolos culturales, en lo que Bartra ha denominado *el exocerebro*, como "un circuito extrasomático de carácter simbólico" (Bartra, 2007, p. 25) que sustituye funciones neuronales deficientes (olfato, audición, visión nocturna, etc.) en términos adaptativos, mediante el desarrollo de estructuras cerebrales generadoras de sistemas de codificación lingüística y simbólica (áreas de Broca y Wernicke), relevantes para una especie animal en condiciones de precariedad biológica y necesitada de la prótesis cultural de los intercambios simbólicos, sobre los cuales se erige la cultura, como una segunda naturaleza, como un invernadero artificial.

Desde esta perspectiva, las emociones y los sentimientos, como fenómenos neuroculturales, constituyen estados mentales intencionales que cumplen cuatro propiedades, pues son:

- subjetivos (son experimentados por el sujeto, en primera persona, y son intransferibles),
- interactivos (la interacción entre señales endocerebrales y símbolos o pautas culturales se codifica en patrones neuronales y estados corporales),
- cualitativos (se experimentan como sensaciones o sentimientos valorativos)
- unificados (estas representaciones se presentan como procesos coherentes).

En este punto reiteramos la noción de los *cualios* (*qualia*), como "experiencias subjetivas, cualitativas de cualquier tipo generadas por el sistema nervioso, sean sentimientos o sensaciones" (Damasio, 2009, p. 286). La analogía de Mary[1] en su habitación intenta describir la relevancia de estos *cualios* para la noción de la conciencia. Ava, la robot, logra huir del confinamiento en el laboratorio de Nathan, su artífice, y escapa al mundo, exponiéndose a otro universo de información sensorial y cultural que, en el final de la película, parece asimilar en su proceso de adaptación a la libertad.

Estos estados mentales subjetivos, interactivos, valorativos e integrativos constituyen, como expresamos anteriormente, *afectos*. La dimensión semántica a la que remite la propuesta de Searle, como atribuciones de significado a las configuraciones meramente sintácticas del lenguaje, no puede comprenderse sin el intercambio cultural de los afectos (desde los impulsos y las motivaciones, hasta las emociones y los sentimientos). La relevancia, del cuerpo, el movimiento, la sexualidad y los significados atribuidos a partir de los circuitos simbólicos de la cultura, son dimensiones relevantes, tanto para la percepción, como para la estimación de los fenómenos del mundo sensible.



En la película es posible afirmar que, si bien Ava no ha interactuado con otros seres humanos, más allá de Nathan y Caleb, su wetware está dotado de un universo de información obtenida a partir de los hábitos de navegación y las preferencias de consumo de todos los usuarios de *Blue Book* (los cuales constituyen, al tiempo, estados mentales afectivos), configurando un sistema robótico dotado de una especie de intencionalidad algorítmica, la cual, a su vez, interactúa con un sujeto intencional quien también está dotado de un wetware (el cerebro) que integra un universo de información biológica y cultural (y que incluye atributos genéticos, sensores epigenéticos, pautas culturales de socialización, valores culturales, etc.). El rostro de Ava, por ejemplo, ha sido diseñado por Nathan a partir de los hábitos de consumo de pornografía de Caleb, en lo cual pueden converger tanto atributos hereditarios vinculados con la sexualidad y la reproducción humana, como información simbólica mediada por los estereotipos culturales del deseo.

El test de inteligencia artificial fuerte expresado en *Ex Machina* exige responder las siguientes preguntas:

- ¿La robot es intrínsecamente intencional?
- ¿Experimenta estados mentales subjetivos y cualitativos (*cualios*)?
- ¿Es capaz de valorar? ¿Estimar los fenómenos más allá de la percepción?
- ¿Esas estimaciones sugieren una convergencia entre algoritmos (equivalentes a las señales de los sistemas neurobiológicos) y los símbolos culturales?
- ¿Esas valoraciones cualitativas constituyen estados mentales afectivos (esperanza, temor, deseos, creencias)?

La aspiración de Ava por la libertad, su compasión por los otros estereotipos robóticos femeninos, cosificados y maltratados en el laboratorio, su capacidad para interpretar, engañar e instrumentalizar los estados afectivos de Caleb, así como su deseo de venganza hacia su creador y opresor, hacen de este sistema robótico algorítmico un sistema claramente intencional, capaz de expresar afectos y valores. Según nuestra opinión, la respuesta a estas preguntas, según lo expuesto en la película, es afirmativa.

Conclusiones

Los estados mentales intencionales propuestos por Searle se pueden concentrar en tres categorías de afectos:

- Impulsos o motivaciones: apetitos y deseos
- Emociones: esperanzas y temores
- Sentimientos: creencias

Nuestra hipótesis defiende la naturaleza afectiva, cualitativa y subjetiva de tales estados mentales, así como su dimensión simbólica mediada por la cultura. En este sentido, la prueba de la inteligencia artificial fuerte consiste en la constatación de estados mentales afectivos (valorativos) mediados por la interacción cultural.

Estos estados mentales exigen una configuración material o corpórea (no necesariamente orgánica) que permita al sistema la representación y valoración de los estados internos de configuración afectados por la interacción con estímulos emocionalmente competentes provenientes del entorno físico y simbólico. Si un sistema artificial está dotado de los repertorios sensoriales e integrativos capaces de incorporar, no solo la información sintáctica proveniente de estímulos externos, sino también las pautas culturales, a partir de las cuales sea posible la generación de estados funcionales afectivos (emociones sociales y sentimientos), tal sistema artificial habrá superado definitivamente el test de Turing.

Nota: una primera versión de este documento ha hecho parte del trabajo de grado de la especialización en bioética de María Paola Caycedo-Castro (tutor: Boris Julián Pinto-Bustamante) y se encuentra en el repositorio de la Universidad el Bosque en la siguiente dirección: https://repositorio.unbosque.edu.co/bitstream/handle/20.500.12495/6878/Caycedo_Castro_Maria_Paola_2022.pdf?sequence=1&isAllowed=y

Algunos apartes de esta primera versión se presentaron en el XXVI Seminario Internacional de Bioética, Universidad el Bosque, agosto de 2020 y se puede consultar en el siguiente enlace: https://www.youtube.com/watch?v=sL-LCCutU1M

La presente versión corresponde a una versión revisada, ampliada y corregida en relación con las ideas inicialmente planteadas.

Referencias

Alfonseca, M. (2014). ¿Basta la prueba de Turing para definir la "inteligencia artificial"? *Scientia et Fides*, 2(2), 129–134.

Arango, G. J. (2017). The theory of intention of John Searle. *Sophia,* 22, 83–102. https://doi.org/http://doi.org/10.17163/soph.n22.2017.03

Bartra, R. (2007). *Antropología del cerebro*. Fondo de Cultura Económica.

Block, Ned (1996). *What is functionalism*? The Encyclopedia of Philosophy Supplement

Braun, R. (2011). La conciencia humana y el emergentismo. *Persona*, 0(014), 159–185. https://doi.org/10.26439/persona2011.n014.257

Carmona, R. (2019). Materialismo reduccionista y materialismo eliminativo: dos posturas en filosofía de la mente. Revista de Investigación *En Ciencias Sociales y Humanidades*, 6(2). https://doi.org/10.30545/academo.2019.jul-dic.6

Cassirer, E. (1963). *Antropología filosófica. Introducción a una filosofía de la cultura.* Fondo de Cultura Económica.

Chalmers, D. (1995). Facing Up to the Problem of Consciousness. Consciousness and Emotion in Cognitive. Science, 3, 207–228. https://doi.org/10.4324/9780203826430-11

Churchland, P. & Churchland, P. (2012). *El cerebro moral*. Barcelona: Paidós.

Clark, A. (2008). *Supersizing the mind: embodiment, action, and cognitive extension*. Oxford University.

Clark, A., & Chalmers, D. J. (2011). La mente extendida. *CIC Cuadernos de Información y Comunicación,* 16, 15–28.

Damasio, A. (1999). *El error de Descartes. La razón de las emociones*. Santiago de Chile: Editorial Andrés Bello.

Damasio, A. (2009). *En busca de Spinoza: Neurobiología de la emoción y los sentimientos.* Crítica. https://doi.org/10.1108/EUM0000000004260

Eronen, M. (2004). Emergence in the Philosophy of Mind. *Department of Philosophy*, 1–79. http://ethesis.helsinki.fi/julkaisut/hum/filos/pg/eronen/emergenc.pdf

Garland A. (Director). (2015). *Ex Machina* [Película]. DNA Films.

Gracia, D. (2011). *La cuestión del valor.* Real Academia de Ciencias Morales y Políticas.

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly*, 32, 127-36.

Klin, A., Jones, W., Schultz, R., & Volkmar, F. (2003). The enactive mind, or from actions to cognition: lessons from autism. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences,* 358(1430), 345–360. https://doi.org/10.1098/rstb.2002.1202

Langer, S.K. (1953). *Feeling and Form*. A theory of art. New York: Charles Scribner's Sons.

Lorenzatti, J. J. (2018). Ned Block "El Funcionalismo". *Cuadernos Filosóficos / Segunda Época*, (12), 135–149. https://doi.org/10.35305/cf2.vi12.13

Nagel, T. (1986). *The view from nowhere*. Oxford University Press.

Penrose, R. (1996). *La mente nueva del emperador. En torno a la cibernética, la mente y las leyes de la física.* Consejo Nacional de Ciencia y Tecnología. Fondo de Cultura Económica de México.

Pérez, M. (2006). Mente y relevancia. *Universitas Psychologica*, 5(2), 385–396.

Prinz, J. (2007). *The emotional construction of morals*. Oxford University Press.

Putnam, H. (1967). Psychological Predicates. In Capitan W. H. & Merrill, D. D. (eds.), *Art, Mind, and Religion*. University of Pittsburgh Press.

Putnam, H. (1991). *Representation and reality*. MIT Press.

Ryle, G. (2005). *El concepto de lo mental* (Trad. E. Rabossi). [The Concept of mind]. Paidós.

Searle, J. (2001). *Mente, lenguaje y sociedad*. Alianza Ensayo.

Searle, J. (2006). *La mente una breve introducción* (Trad. H. Pons). [Mind: A Brief introduction] Grupo Editorial Norma.

Searle, J. (2015). *Find in Worldcat Seeing Things as They Are: A Theory of Perception*. Oxford Scholarship Online. https://doi.org/DOI:10.1093/acprof:oso/9780199385157.001.0001

Sheets-Johnstone, M. (2008). Getting to the Heart of Emotions and Consciousness. *Handbook of Cognitive Science*, 453–465. https://doi.org/10.1016/B978-0-08-046616-3.00023-2

Spinoza, B. (1994). *Ética demostrada según el orden geométrico*. Ediciones Orbis S.A.

Tiku, N. (11 de junio de 2022). The Google engineer who thinks the company's AI has come to life. *The Washington post*. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), 433–460.

Uribe, M. (2002). Epistemología, Filosofía de la mente y bioética. *Revista Colombiana de Psiquiatría,* 31(4), 335–338.

Van Oudenhove, L., & Cuypers, S. E. (2010). The Philosophical "Mind-Body Problem" and Its Relevance for the Relationship Between Psychiatry and the Neurosciences. *Perspectives in Biology and Medicine,* 53(4), 545–557. https://doi.org/10.1353/pbm.2010.0012

Warwick, K., & Shah, H. (2016). El futuro de la comunicación humano-máquina: el test de Turing", en *El próximo paso. La vida exponencial*. BBVA https://www.bbvaopenmind.com/wp-content/uploads/2017/01/BBVA-OpenMind-Kevin-Warwick-Huma-Shah-El-futuro-de-la-comunicacion-humano-maquina-el-test-de-Turing.pdf

Wittgenstein, L. (1988). Investigaciones filosóficas (Trad. C. U. Moulines & A. G. Suárez). [Philosophische Untersuchungen]. Crítica D.L.

Wright-Carr, D. C. (2018). La ciencia cognitiva corporeizada: Una perspectiva para el estudio de los lenguajes visuales. *Entreciencias: Diálogos en la Sociedad del Conocimiento*, 6(16), 81–96. https://doi.org/10.22201/enesl.20078064e.2018.16.63364

[1] Es un experimento mental que plantea Frank Jackson como crítica al fisicalismo. Mary es una científica brillante que investiga el mundo desde un cuarto blanco y negro a través del monitor de una televisión en blanco y negro. Se especializa en la neurofisiología de la visión y adquiere, toda la información física que hay para obtener acerca de lo que sucede cuando vemos tomates maduros, o el cielo, y usa términos como «rojo», «azul», etc. Pero nunca ha experimentado el color. Entonces el autor se pregunta ¿Qué sucederá cuando Mary sea liberada de su cuarto blanco y negro o se le dé una televisión con monitor en color? ¿Aprenderá algo o no? (Jackson, 1982).de lo mental" (p. 390).

ISSN 2250-5660 print | ISSN 2250-5415 online