MIGUEL HOYUELOS

ES DOCTOR EN FÍSICA, SE DESEMPEÑA COMO DOCENTE EN LA UNIVERSIDAD NACIONAL DE MAR DEL PLATA Y ES INVESTIGADOR DEL CONICET. TAMBIÉN ES ESCRITOR. SICCUS, SU PRIMERA NOVELA (LETRA SUDACA , 2014), OBTUVO UNA MENCIÓN ESPECIAL DEL PREMIO EUROPEO DE CIENCIA-FICCIÓN.

CONCIENCIA ARTIFICIAL EN LA CIENCIA Y EN LA FICCIÓN

PALABRAS CLAVES: INTELIGENCIA ARTIFICIAL, CINE DE CIENCIA FICCIÓN, LÍMITES ENTRE FICCIÓN Y REALIDAD

RESUMEN. Este artículo expone algunas ideas vinculadas con el fenómeno de la inteligencia artificial, empezando con un recorrido por las películas de ciencia ficción que desarrollan el tema desde diferentes perspectivas. Se plantea luego el problema de la conciencia como elemento esencial de una inteligencia artificial, partiendo de la dificultad con que se ha encontrado la ciencia para definir qué es una conciencia y con ello poder reproducirla artificialmente. Mediante un repaso de los paradigmas más importantes, se analiza la plausibilidad de la existencia futura de una conciencia artificial, y se examina la idea del cerebro crítico, haciendo hincapié en el conocimiento del funcionamiento del cerebro como el enfoque más adecuado para desentrañar el problema de la conciencia.

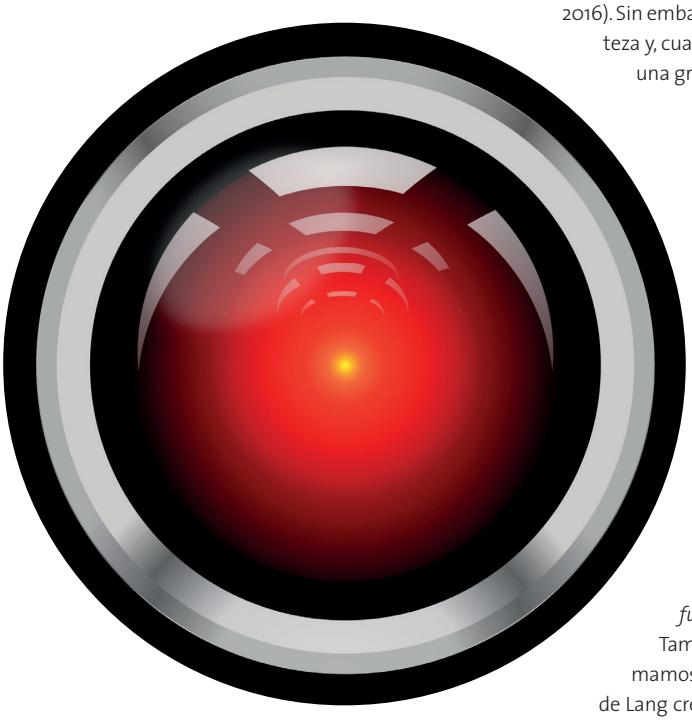
KEYWORDS: ARTIFICIAL INTELLIGENCE, SCI-FI CINEMA, BOUNDARY BETWEEN FICTION AND REALITY

ABSTRACT. This article presents some ideas regarding the phenomenon of artificial intelligence, beginning with an overview on a number of science fiction films that deal with the topic from different perspectives. The problem of consciousness as an essential element of artificial intelligence is then raised, starting with the difficulty science has encountered in defining what a consciousness is, and thus being able to reproduce it artificially. Through a review of the most important paradigms, the plausibility of the future existence of an artificial consciousness is analyzed, and the idea of the critical brain is examined, emphasizing that knowledge of the functioning of the brain might be the most appropriate approach to unravel the problem of consciousness.

La inteligencia artificial ha sido un tópico muy visitado por la literatura y el cine de ciencia ficción. Un par de décadas atrás, máquinas dotadas con algún rasgo de inteligencia o conciencia humana estaban claramente ubicadas en el terreno de la ficción. Hoy, la posibilidad de mantener una conversación, al menos de unas pocas frases, con un teléfono celular ha transformado a esas máquinas en un concepto que no parece puramente ficcional. Es un campo interdisciplinario, que involucra a matemáticos, psicólogos, neurólogos, ingenieros electrónicos, físicos, biólogos, informáticos, etc., que durante las últimas décadas atrae el interés de un número creciente de investigadores y que avanza de forma persistente. A medida que el conocimiento avanza, los límites entre ficción y realidad se desdibujan.

En la primera parte de este artículo voy a mencionar un puñado de películas de ciencia ficción, haciendo énfasis en las que la inteligencia artificial posee, en alguna medida, rasgos humanos. En la segunda parte me voy a referir, también en forma resumida, al problema de la conciencia artificial, que parece estar lejos todavía de ser resuelto. El problema reside, en parte, en la dificultad que existe incluso en definir con claridad qué es la conciencia y cómo se produce, a pesar de que la experimentamos permanentemente.

Una inteligencia que no solo es capaz de jugar al ajedrez o de programar el despertador si se lo pedimos, sino que tiene la versatilidad y la capacidad de adaptación de una inteligencia humana, e iguala o supera su potencial, recibe el nombre de "inteligencia general artificial". Aunque aún es un problema abierto, muchos investigadores suponen que la conciencia debería ser un ingrediente fundamental de este tipo de inteligencia. Una inteligencia artificial que pueda pensar, tener una mente y ser autoconsciente es una "inteligencia artificial fuerte", y por el momento se trata solo de una hipótesis. Aquí voy a usar los términos "inteligencia artificial fuerte" y "máquina consciente" como sinónimos. Los expertos estiman que la inteligencia artificial fuerte podría alcanzarse entre 2040 y 2050 (Müller y Bostrom, 2016). Sin embargo, resulta difícil realizar pronósticos tecnológicos con alguna certeza y, cuando un pronóstico se ubica 20 o 30 años en el futuro, es indicio de una gran incertidumbre.



(Las referencias que se mencionan al final de este artículo son una muy pequeña fracción de la extensa bibliografía sobre el tema; los lectores más entusiastas hallarán más fuentes de información en las citas dentro de esas referencias.)

CINE DE INTELIGENCIA ARTIFICIAL

El robot de aspecto humano, o androide, debutó en el cine en 1927, en el clásico de Fritz Lang *Metrópolis*. La palabra robot, introducida por el escritor checo Karel Čapek en 1920, aún no se había difundido. Por entonces lo llamaban *Maschinenmensh*: máquina humana. En la película de Lang, reemplaza a una mujer llamada María. El primer androide del cine fue, entonces, un ginoide, un robot de aspecto femenino. La palabra "androide" fue introducida en una novela de ciencia ficción, *La Eva futura* del francés Auguste Villiers de l'Isle-Adam, publicada en 1886. También se trató de lo que hoy, para mayor corrección etimológica, llamamos ginoide. Como sucede con frecuencia en ciencia ficción, la historia de Lang crea conflicto y tensión recurriendo al temor y la incertidumbre que produce lo desconocido, lo que queda fuera de control. La falsa María es, finalmente, destruida, para alivio de la audiencia.

El mismo recurso vuelve a aparecer en otras grandes películas de ciencia ficción. En la de Stanley Kubrick de 1968, 2001 *Odisea del espacio*, la inteligencia artificial no toma aspecto humano. Es la computadora HAL 9000, que lo ve todo a través de un ojo-cámara de color rojo. HAL considera "mecanismos fallidos", y prescindibles, a los humanos escépticos o dubitativos. En *Colossus: el proyecto Forbin* (Joseph Sargent, 1970), los Estados Unidos dejan la defensa en manos de una supercomputadora; luego de contactar con su par

soviético, las máquinas deciden unirse; forman una inteligencia consciente que, con la intención de asegurar la paz mundial, envía el siguiente mensaje a la humanidad: "obedézcanme y vivan, o desobedezcan y mueran". *THX* 1138 es una película de 1971 poco conocida que, con el tiempo, se ha transformado en un film de culto; es el primer largometraje de George Lucas; presenta un futuro distópico en el que el uso de drogas para suprimir la emoción y evitar las relaciones sexuales es obligatorio; el control está a cargo de policías androides. En *Blade Runner* (Ridley Scott, 1982) los papeles se invierten, androides rebeldes casi indistinguibles de los seres humanos son perseguidos por un grupo de policías humanos especializados en ese trabajo. Al final surge el interrogante acerca de la verdadera naturaleza del personaje principal, una vuelta de tuerca que no está presente en el libro original de Philip Dick y que agrega interés a la película. En *Terminator* (James Cameron, 1984), una inteligencia artificial llamada Skynet lidera un ejército de máquinas para, nada menos, destruir a la humanidad, aunque para eso sea necesario retroceder en el tiempo. En *Matrix* (Lana y Andy Wachowski, 1999), la población está inmersa, sin saberlo, en un enorme mundo virtual, controlado por programas de inteligencia artificial que adoptan aspecto humano; el objetivo es ocultar el estado de esclavitud de las personas, que son conservadas solo para producir energía. La excusa de generar energía es un detalle secundario, aunque desatinado desde el punto de vista de las leyes de la física, que pone un poco a prueba la capacidad del espectador de inhibir la incredulidad. Pero cualquier defecto es de sobra compensado por las virtudes de la película, tanto por la trama y la acción como por los planteos profundos acerca de la frágil línea que separa lo real de lo virtual. En *Yo, robot* (Alex Proyas, 2004), el protagonista humano debe enfrentar una rebelión de androides; la trama apenas tiene alguna relación con las historias de Asimov publicadas con el mismo nombre.

En todos los casos, la inteligencia artificial, de aspecto humanoide o no, es una amenaza, sale de nuestro control y trae nuestra perdición. Aunque la inteligencia artificial como peligro o amenaza sigue y seguirá siendo un recurso valioso para la ciencia ficción, también se han explorado otros posibles conflictos inherentes a ella, con más frecuencia en los últimos años.

Ghost in the Shell (Mamoru Oshii, 1995) explora las consecuencias de una asociación cada vez más estrecha entre humanos y máquinas. ¿Cómo modificará esa fusión a nuestra identidad, o a nuestra existencia misma? Transcurre en un futuro distópico de alta tecnología: el futuro ciberpunk, donde una red de computadoras inteligentes es omnipresente. Planteos relacionados aparecen en *La máquina* (Caradog W. James, 2013); científicos del Reino Unido realizan implantes cibernéticos en el cerebro de personas para crear soldados cíborgs.

En *El hombre bicentenario* (Chris Columbus, 1999), película basada en una novela de Isaac Asimov, el proceso es invertido: un robot androide va adquiriendo, poco a poco, rasgos, características —y órganos— humanos, con el objetivo de ser reconocido legalmente como tal.

Identidad sustituta (Jonathan Mostow, 2009) no logró buenas críticas a pesar de basarse en una idea interesante. En este caso la máquina no se fusiona con el humano, sino que lo reemplaza. Los androides sustitutos son versiones mejoradas, más jóvenes y atractivas, de los humanos originales, que viven a través de ellos manejándolos por control remoto, experimentando sus mismas sensaciones, pero con la tranquilidad de saber que el dolor que los sustitutos puedan sufrir por algún daño será filtrado.

Steven Spielberg renovó el cine de ciencia ficción, en especial el de las superproducciones taquilleras, haciendo películas donde, por ejemplo, los extraterrestres ya no venían hasta nosotros para someternos o destruirnos, sino para salvarnos (*Encuentros cercanos del tercer tipo,* 1977) o para pedir ayuda (*E.T.*, 1982). Hizo algo similar con la inteligencia artificial. La máquina ya no es el enemigo en I*A. Inteligencia artificial* (2001), donde se plantea la posibilidad de que surjan sentimientos de paternidad hacia un androide-niño programado para demostrar amor. El tema fue planteado antes en *D.A.R.Y.L.* (Simon Wincer, 1985) y antes aún en la serie de televisión japonesa *Astroboy* (Osamu Tezuka, 1963), donde un científico crea un niño-robot para reemplazar a su hijo muerto. Vuelve a aparecer en *EVA* (Kike Maillo, 2011), una muy buena y poco conocida película española.

El sentimiento de amistad entre hombre y máquina aparece en *Frank y el robot* (Jake Schreier, 2012). La película *Her* (Spike Jonze, 2013) da un paso más allá y presenta de manera convincente un hecho que, a primera vista, parece chocante o difícil de aceptar. Se trata de una historia de amor platónico entre un

hombre y un programa, un sistema operativo, llamado Samantha y representado por la voz de Scarlett Johansson. Un acierto de la película es no dar a Samantha un aspecto físico; la relación se desarrolla a través de la voz. El ambiente futurista se logra a través de la arquitectura, la ropa y la tecnología que rodea a los protagonistas. Sin embargo, la organización de la sociedad y el trabajo y las relaciones entre las personas son como las actuales. Este detalle se ha visto como un defecto. Creo, en cambio, que es otro acierto. Mostrarnos una sociedad similar a la nuestra nos facilita la identificación con los personajes y la inmersión en la historia. Además, es probable que en el futuro las cosas sucedan así. Los que esperen acción en *Her* no se verán satisfechos, es una película romántica.

La vigencia del tema se manifiesta también en la televisión. La serie *Humans* tuvo su primera temporada en 2015 con muy buena recepción de la crítica. Las historias sacan provecho de los conflictos sociales, psicológicos o culturales que surgen con la difusión de robots trabajadores o sirvientes domésticos, de aspecto humano, llamados "synths". Una situación similar puede verse en la serie *Westworld*, que este año comienza su tercera temporada; seres artificiales son anfitriones en un parque de diversiones que simula el salvaje oeste. Estos seres se ven casi indistinguibles de los humanos, tanto que parecen tener conciencia y capacidad para el sufrimiento. La serie enfoca la atención en los conflictos éticos que surgen por la forma en que los humanos tratan a los androides.

Volviendo al cine, las últimas que quiero mencionar son *Transcendence* (Wally Pfister, 2014) y *Ex Machina* (Alex Garland, 2015), que me parecen interesantes por la profundidad de algunos planteos y porque creo que marcan el camino de lo que vendrá o, al menos, de lo que me gustaría que viniera en el cine de ciencia ficción. Aproximadamente la primera mitad de *Transcendence* transmite de manera efectiva la lucha, los conflictos y las emociones de un grupo de científicos que sufren ataques de terroristas antitecnología, y que buscan, y logran, transferir la conciencia de uno de ellos, moribundo, a una máquina. La máquina muestra rasgos de la personalidad de su original, pero se mantiene una brecha que la diferencia de los humanos. En algún momento se comenta que la emoción humana involucra una serie de conflictos contradictorios que la máquina no puede reproducir. Lo que conduce a las siguientes preguntas: ¿qué diferencia a las máquinas de los humanos?, ¿cuál es la esencia de la naturaleza humana que una máquina sería incapaz de reproducir? Hasta ahora la respuesta siempre pareció evidente y ni siquiera resultó necesario plantear la cuestión. Pero la capacidad creciente de las máquinas —real o imaginaria— hace que la diferencia sea cada vez más difícil de establecer. La segunda parte de la película decae a una pelea entre buenos y malos y a un salto de la conciencia artificial hacia un estado supertrascendente que resulta difícil, por no decir imposible, de describir. Un final similar se usó en *Lucy* (Luc Besson, 2014) y también en *Her*.

La historia de *Ex Machina* gira en torno a la prueba de Turing, con una modificación. En 1950 Alan Turing buscó dejar a un lado el problema de si las computadoras pueden pensar, por la dificultad en definir qué es pensar, y propuso una prueba para determinar si las computadoras pueden tener un comportamiento inteligente indistinguible del de un humano usando el lenguaje natural a través de una pared. El que hace la prueba no sabe si detrás de la pared hay una persona o una máquina. En la película se modifica la prueba eliminando, nada menos, la pared. Un robot con aspecto de mujer, una ginoide, está a la vista. Aun así, la máquina logra convencer al encargado de llevar adelante la prueba, y también al espectador, de que es una persona con propósitos, deseos y temores. Lo logra de manera brillante usando diálogos breves y gestos mínimos. En un momento dado pregunta qué será de ella si no supera la prueba. La respuesta queda implícita: será destruida. Y el espectador se ve arrastrado a compartir el desasosiego del personaje que está ante ella.

Como se dijo antes, la capacidad creciente de las máquinas, y también el conocimiento creciente sobre el funcionamiento del cerebro, hacen que cada vez sea más complejo distinguir entre máquina y humano. *Transcendence y Ex Machina* hacen referencia a este problema que, a su vez, nos conduce a otra cuestión filosófica y profunda: ¿qué somos? Es decir, el hecho de que una máquina, puramente material, se torne equivalente a un ser humano tendría consecuencias importantes sobre nuestras ideas acerca de nuestra propia naturaleza. La cuestión está planteada y el cine de ciencia ficción está empezando a sacar provecho de ella. Es un gran campo abierto para la exploración. Hasta hoy ha habido una brecha entre los humanos y las máquinas, representada, por ejemplo, por la incapacidad de las últimas de experimentar

emociones, sentimientos o sentir dolor o placer, por una desmedida búsqueda de acumulación de poder, por dificultades en utilizar lenguaje metafórico, por ser un poco tontas, o por alguna otra característica que las distinguía de nosotros. Desde la perspectiva de un autor de ciencia ficción, es complicado cruzar esa brecha, pues pone al lector ante una situación difícil de aceptar. La razón por la que no es tan frecuente hallar en ciencia ficción la idea de una máquina no solo inteligente, sino también autoconsciente y con la capacidad humana para la emoción y el sufrimiento, de la cual no podamos diferenciarnos fácilmente, es que se trata de un concepto tradicionalmente chocante. Sin embargo, es una idea que empieza a aceptarse poco a poco con el paso del tiempo o, al menos, no suena tan disparatada como hace solo unas décadas atrás. Comparemos, por ejemplo, a C-3PO con la ginoide de *Ex Machina*. Otro ejemplo con dos casos en los que no tenemos una representación física: HAL, de 2001, es claramente no humano, en cambio Samantha, de *Her,* parece humana.

LA CONCIENCIA NATURAL

Dejemos ahora la ciencia ficción y veamos qué nos puede decir la ciencia sobre este tema. La inclusión de la conciencia como parte de una inteligencia artificial fuerte tiene un inconveniente muy básico: ni siquiera existe consenso acerca de una definición precisa de la conciencia y de cómo se produce. Esta situación parece sorprendente a primera vista, porque todos experimentamos la conciencia; sabemos qué es al experimentarla. Pero poner en palabras esa experiencia resulta más complicado; mucho más, reproducirla en una máquina.

Aquí solo me propongo aclarar ideas relacionadas con la noción de conciencia usando, en particular, algunos conceptos del artículo de Zeman (2005), *¿What in the world is consciousness? (¿Qué demonios es la conciencia?)*. Conciencia se asocia tanto a la vigilia, a estar despierto, como al discernimiento. El nivel de conciencia, en el sentido de vigilia, se establece observando, por ejemplo, si el individuo tiene los ojos abiertos o es capaz de mantener una conversación. La conciencia, en el sentido de discernimiento, de estar consciente de algo en particular y conocer su significado, va más allá de la vigilia y es más difícil de establecer, pues involucra una experiencia subjetiva. A este nivel básico ya puede observarse la ambigüedad del concepto de conciencia, porque está asociado a estados que se pueden diferenciar. A un nivel mayor de complejidad, pasamos al concepto de autoconciencia, más restringido pero también multifacético; se refiere principalmente a nuestra capacidad para reconocernos no solo como un cuerpo sino también como una mente, como un sujeto con experiencia, con deseos, creencias e intereses propios, lo que nos lleva a interactuar de una forma personal con los objetos y los individuos que nos rodean. Lo que nos lleva, a su vez, a otro nivel de autoconciencia: el discernimiento del discernimiento de los otros, que adivinamos con limitaciones similares a las propias.

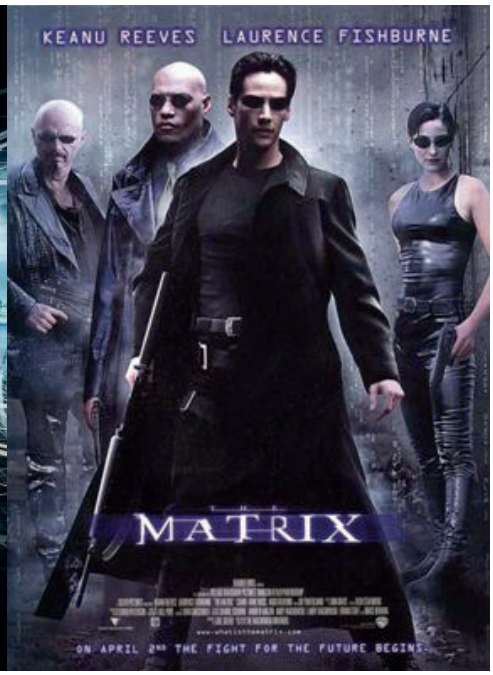
El concepto se hace aún más complejo al observar fenómenos que están en el límite entre lo consciente y lo inconsciente como, por ejemplo, la visión ciega. Los pacientes con ceguera cortical tienen daño en algunas áreas del cerebro asociadas a la visión y perciben vagamente la luz o el movimiento, pero son incapaces de reconocer objetos. Pueden, sin embargo, guiar su mano hacia un objeto, aunque no son conscientes de su presencia, no lo perciben a nivel consciente. Esto significa que nuestro comportamiento (el movimiento de la mano) puede ser guiado por información sensorial de la que no somos conscientes, información que se procesa en lo que se denomina módulos inconscientes de función cognitiva. Hoy en día se supone que el procesamiento consciente ocurre cuando dichos módulos unen fuerzas, se comunican a lo largo del cerebro en un proceso que unifica la información y genera una percepción (Celesia, 2010).

Una concepción frecuente, posiblemente la dominante, acerca de la conciencia es que se trata de un proceso interno, oculto, privado, solo accesible al que lo experimenta e inaccesible para el resto. Esta concepción es el punto de partida desde donde surgen las dificultades en el estudio de la conciencia. Implica que el objeto de estudio está fuera del alcance del escrutinio científico. Es inobservable. Solo podríamos aspirar a identificar la relación entre procesos neuronales y experiencias, lo que se conoce como el correlato neuronal de la conciencia (Tononi y Koch, 2008), insuficiente para alcanzar una explicación (acerca de la posible confusión de este concepto con la dualidad materia-mente, ver Bunge 2010, p. 150).

El problema es difícil. Tanto que ha recibido el nombre, no muy imaginativo, de "el problema difícil de la conciencia" (Chalmers, 1995). Entre la observación de la actividad cerebral y la experiencia interna de la conciencia parece haber una brecha imposible de superar; la literatura sobre el tema se refiere a una instancia individual de experiencia consciente con el nombre de quale (plural: *qualia*). Newton ya se refirió a este problema en 1672 cuando escribió: "no es tan fácil determinar con qué acciones o de qué modo la luz produce el fantasma del color en nuestras mentes". Ese "fantasma del color" es un quale. ¿Cuál es la explicación

que supera la brecha? ¿La existencia del alma? Los científicos son vistos, posiblemente, como personas poco espirituales, pero lo cierto es que el concepto de alma es profundamente insatisfactorio como explicación. Elementos milagrosos o sobrenaturales surgen tradicionalmente para explicar lo inexplicable. La ciencia, en cambio, está comprometida con una explicación natural, no mítica, del mundo. El alma queda afuera y, mientras tanto, uno debería resignarse a admitir que no entiende lo que está pasando. Si nuestra rica vida interior no es pasible de una representación en términos de procesos físicos, entonces no tenemos una explicación; algunos filósofos han intentado la formulación de nuevos conceptos, como la existencia de un constituyente fundamental del universo que formaría la esencia de la conciencia, pero no hacen más que confirmar el carácter elusivo de lo que se pretende definir.

Hay, sin embargo, una alternativa. Dennet (1991) propuso que la conciencia es una ilusión. Creemos que vemos un color, o sentimos un dolor, pero eso no es más que ilusión, es como un truco de magia en el que se usan métodos ordinarios para dar la impresión de un acto sobrenatural; la conciencia hace trucos para hacernos creer que carece de todo sustrato físico. No parece una alternativa atractiva pero, a pesar de ciertas críticas iniciales, se está volviendo popular; no en la sociedad, pero sí en el ambiente científico. Según Dennett, la única teoría posible es una que explique los eventos conscientes en términos de eventos inconscientes (los módulos inconscientes de función cognitiva mencionados antes). En la misma línea, Dehaene (2014) considera inexistente el problema difícil de la conciencia; afirma que "el concepto hipotético de qualia, experiencia mental pura, separada de cualquier función de procesamiento de la información, será visto como una idea peculiar de la era precientífica, muy similar al vitalismo". Una objeción es que resulta difícil aceptar que la conciencia sea una ilusión porque, por lo que respecta a ella, la apariencia es la realidad; el hecho de que esa apariencia exista la hace real (Searle, 1997). Estos son, en forma muy resumida, los elementos de un debate que no termina de resolverse.



LA CONCIENCIA ARTIFICIAL

No existe (todavía) una "inteligencia artificial fuerte", una máquina consciente. Como se mencionó al principio de este artículo, los especialistas suponen, en promedio, de manera quizá optimista y seguramente incierta, que dentro de 20 o 30 años se lograría ese objetivo. Mientras tanto, son varias las propuestas de modelado de la conciencia en una máquina (Aleksander, 2005). Los enfoques posibles se encuentran ubicados entre dos paradigmas extremos. De un lado, se tienen los modelos puramente funcionales que enfocan la atención en el comportamiento de la máquina sin importar cómo está constituida, con la aspiración de que ese comportamiento pueda atribuirse a una conciencia. Por el otro, las propuestas se basan en lo que se conoce de la anatomía y el funcionamiento del cerebro. En todos los casos, se tiene la esperanza de que estos trabajos contribuyan a resolver el problema difícil de la conciencia. Sería largo describir los principales modelos con algún detalle, solo voy a mencionar brevemente una propuesta, más adelante, en la sección siguiente. Pero antes veamos la cuestión de la plausibilidad.

¿Es posible una máquina consciente? Según Bunge (2010, p. 230), no: "las computadoras carecen de espontaneidad o libre albedrío: están a merced de sus usuarios. En particular, carecen de la plasticidad, la libertad y la fuerza de voluntad requeridas para decidir aprender nuevos asuntos". Luego agrega: "A diferencia de las computadoras (...), los humanos tienen la capacidad de innovar, desobedecer y hacer trampa. En particular, pueden elaborar, discutir, criticar y poner en práctica reglas de conducta. (...) Tales normas están motivadas y representadas por emociones sociales, como la empatía, la simpatía, la compasión, la vergüenza, el orgullo, la confianza y la desconfianza, que están más allá del alcance de las máquinas." Buttazzo (2001) resume el mismo argumento de la siguiente manera: "La objeción más común para otorgar a las computadoras impulsadas por circuitos electrónicos el estado de autoconsciencia es la percepción de

que, trabajando en un modo completamente automático, no pueden exhibir creatividad, emociones o libre albedrío. Una computadora, como una lavadora, es un esclavo operado por sus componentes." Ni siquiera valdría la pena intentarlo. Sin embargo, como sucede con la mayor parte de las afirmaciones de este artículo, hay posiciones divergentes. El concepto de libre albedrío resulta problemático, quizá aún más que el de la conciencia. Para un materialista eliminativo —algo así como un materialista extremo— tanto el concepto de conciencia como el de libre albedrío serían ficcionales, conceptos precientíficos al estilo del vitalismo o el flogisto. Según esta posición, nosotros solo aparentamos tener libre albedrío, algo que una máquina también podría hacer. Una analogía puede servir para aclarar esta idea. Una computadora funciona de forma automática, determinista, obedeciendo los pasos de un programa. No puede tener la impredecibilidad de un comportamiento aleatorio. Pero puede simularlo muy bien, incluso superar muchos tests estadísticos, usando un generador de números seudoaleatorios; podría, por lo tanto, tener un comportamiento casi indistinguible de uno impredecible, aunque en el fondo no lo sea. Lo mismo sucedería con nosotros: podemos tener un comportamiento casi indistinguible de uno que involucre conciencia y libre albedrío, pero en el fondo no tendríamos una cosa ni la otra.

Algunos investigadores, especialmente los que se dedican al diseño de modelos conectivistas, o de redes neuronales, suponen que una estructura lo suficientemente compleja como para procesar sensaciones, emociones, estados internos, placer y dolor, sería capaz de experimentar conciencia. La conciencia emergería en forma espontánea. Una hipótesis no probada, quizá demasiado optimista. Los modelos de redes neuronales han resultado muy útiles, sin embargo, para una mejor comprensión de diversos aspectos del funcionamiento del cerebro, como el aprendizaje o la memoria. En el artículo de Wedemann y Plastino (2016), por ejemplo, se describe un modelo, basado en una red neuronal, para describir el comportamiento de la memoria a nivel consciente e inconsciente. Otras alternativas, en las que no voy a extenderme porque creo que poseen un poder explicativo menor, son: la conciencia es un fenómeno sobrenatural más allá del alcance de la ciencia; o es un fenómeno que se desarrolla a un nivel esencialmente cuántico; o involucra no solo el cerebro sino todo el cuerpo humano; o no solo el cuerpo sino todo el universo (panpsiquismo). Creo que, en el futuro, lo más provechoso será insistir en un enfoque de la conciencia como fenómeno biológico confinado en el cerebro, de modo que, para comprenderlo, debemos seguir investigando el interior de nuestros cráneos. El hecho de que aún no se sepa qué es eso que suponemos que debería estar ahí, eso que representaría la conciencia, no significa necesariamente que el problema sea invulnerable al escrutinio científico.

EL CEREBRO CRÍTICO

Volvamos a Bunge y sigamos uno de sus consejos: "En general, para imitar cualquier cosa, comience por aprender sobre el artículo genuino" (Bunge, 2010, p. 237). El artículo genuino, una red de neuronas vivientes, exhibe patrones de actividad que pueden observarse con distintos métodos; uno de los más conocidos es, por ejemplo, la tomografía computada, que detecta las zonas más activas, es decir, las que reciben mayor irrigación sanguínea. Los patrones son, típicamente, oscilaciones, sincronización de distintas regiones u ondas que se propagan. Durante las últimas décadas ha llamado la atención otro tipo de comportamiento que se observa, por ejemplo, en neuronas corticales (Beggs y Plenz, 2003). Se trata de fluctuaciones en la actividad neuronal que tienen la forma de avalanchas. Si se grafica la probabilidad con la que se observa una avalancha en función de su tamaño (o sea, de la cantidad de neuronas involucradas), se obtiene que las avalanchas pequeñas son más frecuentes que las grandes. Lo más interesante es que, en escala log-log, ese decaimiento tiene la forma aproximada de una recta. Se trata de una ley de potencia, donde la pendiente de la recta es el exponente. Las fluctuaciones con ley de potencia son un indicio importante de lo que se conoce como comportamiento crítico, frecuente en muchos sistemas físicos cerca de una transición de fase (por ejemplo, el cambio brusco en la magnetización que se observa al bajar la temperatura de un sistema de espines). Hay sistemas que espontáneamente evolucionan hacia la criticalidad. El ejemplo más conocido es el de la pila de arena que se forma al dejar caer un flujo de granos en un punto. La probabilidad de las avalanchas de arena en función de su tamaño sigue una ley de potencia. No es necesario sintonizar un parámetro para que se obtenga, en este caso, el comportamiento crítico (como sería la temperatura en el caso de la magnetización,). Esta situación se conoce como criticalidad autorganizada (Bak et al., 1987). El parámetro, en el caso de la pila de arena, es la pendiente que se forma a medida que la pila crece. Esa pendiente toma espontáneamente un valor crítico, el valor en el cual la arena empieza a desplazarse y las avalanchas empiezan a caer. Según la hipótesis del cerebro crítico, dentro de nuestras cabezas sucede algo similar (Chialvo, 2010). La actividad cerebral está en el punto crítico, en el límite entre dos fases; en una de ellas la actividad decae y muere, nada sucede en el cerebro; en la otra, la actividad crece y crece hasta saturar la red. La criticalidad optimiza la capacidad de procesamiento de información (Shew et *al.*, 2009). Las fluctuaciones críticas en la actividad cerebral cumplen un rol importante en nuestra capacidad de adaptarnos y de responder a un entorno cambiante. Se ha verificado que la hipótesis del cerebro crítico

predice correctamente los patrones de la actividad que se observan en un cerebro en reposo (Haimovici et al., 2013). Una actividad cerebral que se aparte del comportamiento crítico podría constituir un síntoma de una enfermedad del sistema nervioso (Shew y Plenz, 2013); por lo tanto, el análisis del grado de criticalidad puede dar lugar a nuevas herramientas de diagnóstico.

¿Qué hay acerca de la conciencia? La hipótesis de criticalidad no resuelve las dificultades mencionadas en la sección anterior, pero es un avance en otros aspectos. Se ha propuesto a la criticalidad autorganizada como el fundamento de una arquitectura operacional de la actividad cerebral (Fingelkurts et *al.,* 2013), un marco teórico que podría ser útil, por ejemplo, para determinar grados de conciencia, comparar estado consciente mínimo y estado vegetativo, o para predecir qué pacientes en esos estados podrían recuperar la conciencia (Fingelkurts y Fingelkurts, 2018). Por otro lado, como se mencionó antes, las fluctuaciones críticas del cerebro tienen relación con nuestra capacidad de reaccionar ante un entorno continuamente cambiante, es decir, con nuestro estado de alerta, con nuestra conciencia en el sentido de vigilia.

Uno de los proyectos más interesantes relacionados con la idea del cerebro crítico es el desarrollo de redes con interruptores atómicos de plata. El interruptor atómico tiene propiedades similares a las de las sinapsis neuronales; consiste, sin embargo, en un dispositivo puramente inorgánico. El punto de partida es una red regular de 128 electrodos de platino distribuidos en una superficie de 2 mm de lado; son los puntos desde los cuales se tiene acceso a distintas regiones de la red de interruptores. Luego de un proceso, cuyos detalles pueden encontrarse en Demis et al. (2016), se logra el crecimiento de una intrincada red de nanocables de plata, "un plato de fideos altamente interconectado", según uno de los autores, con un aspecto similar a lo que se observa en fotografías de tejido nervioso cortical. Se forman alrededor de mil millones de contactos por centímetro cuadrado entre nanocables, los llamados interruptores atómicos, que consisten en interfaces plata-sulfuro de plata-plata (Ag-Ag2S-Ag). Al aplicar una tensión se forma un filamento de átomos de plata que crece hasta cerrar el contacto; si la tensión se invierte, se produce el efecto contrario: el filamento se reduce y el interruptor se abre. Al usar el interruptor con más frecuencia, la conexión se activa más fácilmente. Si deja de usarse por un tiempo, se desconecta solo. Su funcionamiento depende, por lo tanto, de su historia. Algo similar sucede con las conexiones sinápticas: se refuerzan con el uso durante, por ejemplo, el proceso de aprendizaje. Usando una cámara infrarroja se ha monitoreado el recorrido de la corriente a través de esta red intrincada. Observaron que la corriente cambiaba continuamente su recorrido a través del dispositivo, una actividad no localizada reminiscente a la del cerebro. Lo más llamativo fue, sin embargo, cuando observaron indicios de comportamiento crítico: áreas pequeñas de actividad eran más frecuentes que áreas grandes. El comportamiento crítico en una red artificial de este tipo fue recientemente reportado en Scharnhorst et al. (2018). Algunos experimentos preliminares sugieren que también tiene la capacidad de resolver tareas computacionales, a pesar de estar muy lejos de ser una computadora tradicional (von Bubnoff, 2017). Y todo a partir de una red que emerge de forma espontánea, autorganizada, sin un diseño previo. Por lo pronto, sus capacidades son más bien modestas, pero su potencial podría ser bastante grande.

CONCLUSIONES

La conciencia es uno de los más grandes misterios, quizá el más grande, con el que la humanidad se ha topado; hoy en día muchos investigadores están empeñados en desentrañarlo. Es el más grande misterio y también el más frecuente; lo experimentamos cada uno de nuestros días, a cada momento, y no sabemos cómo se produce. La cuestión de la conciencia está directamente relacionada con una de las preguntas más antiguas, más simples, más profundas y más difíciles de responder: ¿qué somos? Lamento decepcionar a los lectores que llegaron hasta aquí con la esperanza de una respuesta directa. Lo más cercano a una respuesta al problema de la conciencia que uno puede encontrar en la literatura científica son los indicios de existencia, en nuestros cerebros, de módulos, especie de unidades con capacidad cognitiva, que funcionan a nivel inconsciente y que, cuando varios de ellos se combinan y unen fuerzas, dan lugar a un evento consciente, un proceso que crea en nosotros la ilusión, según Dennett (1991), de que percibimos o pensamos algo. Las extensas discusiones que genera esta propuesta suelen dar como resultado la sensación algo frustrante de que uno no ha avanzado mucho en resolver la cuestión principal: ¿qué somos? Mientras tanto, el conocimiento del funcionamiento del cerebro progresa poco a poco, y a paso firme. Desde hace algunos años es posible, por ejemplo, identificar la actividad de neuronas individuales en un cerebro vivo. Este conocimiento incentiva y alimenta el desarrollo de modelos teóricos, simulaciones computacionales y la construcción de redes artificiales que emulan una red neuronal.

Como dice Mizraji (2017): "Si los cerebros son computadores celulares, cuya función se basa en la electroquímica y en señales moleculares, entonces, una vez comprendida plenamente la mecánica de su función puede no haber obstáculos para la creación de robots conceptualizadores y racionales". Una idea

de la ciencia ficción que se toma cada vez con más seriedad en el ambiente científico, a tal punto que varios especialistas opinan que una inteligencia artificial fuerte, o una máquina consciente, sería posible en las próximas décadas. La literatura y el cine de ciencia ficción también cambian y se adaptan a la evolución de estas ideas. Antes siempre estuvo clara la distinción entre máquina y humano. En la actualidad, esa distinción, aunque aún está presente, se va desdibujando poco a poco, como se manifiesta en películas como *Transcendence* o *Ex Machina*. La ciencia ficción ha sabido sacar partido de miedos asociados al avance imprevisible y descontrolado de la tecnología, nuestra o de otros mundos, para crear historias atrapantes. Con respecto a la inteligencia artificial y las cuestiones acerca de nuestra propia naturaleza, creo que las posibles respuestas que nos depare el futuro serán lo suficientemente inquietantes como para producir buenas historias.

REFERENCIAS BIBLIOGRÁFICAS

- Aleksander, I. (2005). "Machine consciousness". En Progress in Brain Research (ed. S. Laureys), Vol. 150, p. 99.
- Bak, P., Tang, C. y Wiesenfeld, K. (1987). "Self-organized criticality: an explanation of 1/f noise". Physical Review Letters, 59(4), pp. 381–384.
- Beggs, J. M. y Plenz, D. (2003). "Neuronal Avalanches in Neocortical Circuits". Journal of Neuroscience, 23(35), pp. 11167-11177.
- Bunge, M. (2010). Matter and Mind: A Philosophical Inquiry. Dordrecht: Springer.
- Buttazzo, G. (2001). "Artificial consciousness: Utopia or real possibility?". Computer, 34(7), pp. 24-30.
- Celesia G. (2010). "Visual perception and awareness: a modular system". Journal of Psychophysiology, 24(2), pp. 62–67.
- Chalmers, D. (1995). "Facing up to the problem of consciousness". Journal of Consciousness Studies, 2, p. 200–219
- Chialvo, D. R. (2010). "Emergent complex neural dynamics". Nature Physics, 6, pp. 744–750.
- Dehaene, S. (2014). Consciousness and the brain: deciphering how the brain codes our thoughts. Nueva York: Penguin Books.
- Demis, E. C., Aguilera, R., Scharnhorst, K., Aono, M., Stieg, A. Z., y Gimzewski, J. K. (2016). "Nanoarchitectonic atomic switch networks for unconventional computing". Japanese Journal of Applied Physics ,55, p. 1102B2.
- Dennett, D. (1991). Consciousness Explained (ed. A. Lane). The Penguin Press.
- Fingelkurts, A. A. y Fingelkurts, A. A. (2018). "Actual Physical Potentiality for Consciousness". AJOB Neuroscience, 9(1), pp. 24–25.
- Fingelkurts, A. A., Fingelkurts, A. A., y Neves, C. F. H. (2013). "Consciousness as a phenomenon in the operational architectonics of brain organization: Criticality and self-organization considerations". Chaos, Solitons & Fractals, 55, pp. 13–31.
- Haimovici, A., Tagliazucchi, E., Balenzuela, P., y Chialvo, D. R. (2013), "Brain Organization into Resting State Networks Emerges at Criticality on a Model of the Human Connectome", Phys. Rev. Lett., 110, 178101.
- Mizraji, E. (2017). "El cerebro, las palabras y los razonamientos". Revista Núcleos (UNNOBA), 5, p. 45.
- Müller, V. C. y Bostrom, N. (2016). "Future progress in artificial intelligence: A survey of expert opinion". En Fundamental issues of artificial intelligence (pp. 555-572). Cham: Springer.
- Scharnhorst, K. S., Carbajal, J. P., Aguilera, R., Sandouk, E. J., Aono, M., Stieg, A. Z., y Gimzewski, J. K. (2018). "Atomic switch networks as complex adaptive systems". Japanese Journal of Applied Physics, 57, 03ED02.
- Searle, J. R. (1997). The mystery of consciousness. Nueva York: New York Review of Books.
- Shew, W. L., y Plenz, D. (2013). "The functional benefits of criticality in the cortex". Neuroscientist, 19, pp. 88-100.
- Shew, W. L., Yang, H., Petermann, T., Roy, R. y Plenz, D. (2009). "Neuronal avalanches imply maximum dynamic range in cortical networks at criticality". The Journal of Neuroscience, 29, pp. 15595-15600.
- Tononi, G. y Koch, C. (2008). "The neural correlates of consciousness: an update". Annals of the New York Academy of Sciences, 1124, p. 239.
- von Bubnoff, A. (septiembre, 2017). "A Brain Built From Atomic Switches Can Learn". Quanta Magazine. Recuperado de https://www.quantamagazine.org/a-brain-built-from-atomic-switches-can-learn-20170920/.
- Wedemann, R. S. y Plastino, Á. R. (2016). "Física estadística, redes neuronales y Freud". Revista Núcleos (UNNOBA), 3, p. 4.
- Zeman, A. (2005). "What in the world is consciousness?". En Progress in Brain Research (ed. S. Laureys), Vol. 150, p. 1.