Rubén H. Ríos

El *test* de Turing y la filosofía de la inteligencia artificial. Acerca de la mente de las máquinas digitales

Resumen: *Este artículo, en términos generales, es una respuesta a la pregunta que formula Alan Turing en un célebre artículo de 1950 acerca de si las máquinas digitales piensan. Con lo que se suma, en el marco de lo que se ha denominado «dataísmo», a uno de los debates centrales de la filosofía de la inteligencia artificial. De allí que revisa sumariamente la historia de esta última, en la cual el momento de la cibernética adquiere cierto valor de hito, así como los aportes de Turing en la concepción de lo computable y la relación del test que propuso con el razonamiento algorítmico.*

Palabras claves: *filosofía de la técnica, máquinas inteligentes, recursividad, virtualidad, simulacro.*

Abstract: *This article, in general terms, is an answer to the question posed by Alan Turing in a famous article from 1950 about whether digital machines think. This adds, within the framework of what has been called «dataism», to one of the central debates in the philosophy of artificial intelligence. Hence, he summarily reviews the history of the latter, in which the moment of cybernetics acquires a certain milestone value, as well as Turing's contributions in the conception of the computable and the relationship of the test that he proposed with algorithmic reasoning.*

Key words: *technique philosophy, intelligent machines, recursion, virtuality, simulacrum.*

Entre los dilemas canónicos de la filosofía de la inteligencia artificial, la rama más reciente de la filosofía de la tecnología, hay uno de particular idiosincrasia en cuanto a sus implicaciones ontológicas, epistemológicas y éticas: aquel que se interroga respecto de si las máquinas digitales piensan. Desde el momento que Alan Turing inicia esta cuestión, en un artículo publicado en 1950 en la revista *Mind* especializada en temas filosóficos (Turing 1950, 433-460), cuando solo había cuatro computadoras electrónicas en el mundo, se admite que con ello también se inaugura la filosofía de la inteligencia artificial y, a la vez, un debate que hasta el momento no ha concluido. La permanencia de la pregunta de Turing («Can machines think?») y de su ingeniosa respuesta, por decir lo menos, descansa en que compromete e interpela —y antes de su despliegue histórico— la esencia misma de los dispositivos tecnológicos que procesan algoritmos. Esto incluye, en pocas palabras, los dos paradigmas sucesivos en la evolución de las máquinas inteligentes, el simbólico y el conexionista y, en la segunda década del siglo XXI, a las aplicaciones de redes neuronales y algoritmos de aprendizaje automático (*machine learning*) y profundo (*deep learning*) (Russell-Norvig 2008).

Si es posible referirse a los impactos socioeconómicos y psicosociales de la irrupción de los macrodatos o *Big Data* (procesamiento informático de datos masivos y variados a gran velocidad) como «dataísmo», neologismo utilizado al parecer por primera vez por el periodista David Brooks del New York Times en 2013, se explica en gran parte por el perfeccionamiento de los algoritmos de *machine learning* que aprenden y mejoran a partir de grandes volúmenes de datos. El dataísmo, en veredicto de Byung-Chul Han (quien ha colaborado con la difusión del término), describe la consagración mítica de una ideología que se presenta como el crepúsculo de todas las ideologías y que, mediante un totalitarismo digital, pretende la reducción de todo conocimiento a datos e información (Han 2014, 40-42). El *Big Data*, de un solo golpe, suprime la mediación de la subjetividad en los procesos cognitivos a favor de una pura objetividad cuantificada y torna caducas la teoría y el sentido. No obstante, Han está en lo cierto a medias, porque identifica dataísmo e inteligencia artificial, cuando el éxito y la eficacia de ésta depende de una narración teórica y filosófica (y de un imaginario) que le confiere sentido.

El gran relato de la inteligencia artificial es una historia «monumentalista», según la categorización nietzscheana de la segunda intempestiva y, como tal, se compone de hitos y de magnos acontecimientos (Nietzsche 1998). Entre los primeros se mencionan por lo general las máquinas conjeturales de Turing (1936-1950), la idea de Claude Shannon de efectuar el álgebra booleana de estructura binaria en circuitos digitales (1937), el modelo matemático de las neuronas de McCulloch y Pitts (1943), el coloquio de la Universidad de Dartmouth que inventa el significante «inteligencia artificial» (1956), el nacimiento del programa informático (el Teórico Lógico, en 1956), el teorema del perceptrón o neurona artificial de Frank Rosenblatt (1958), Eliza (el primer chatbot, 1964), el algoritmo de retropropagación para el aprendizaje de redes neuronales multicapa (1986), el modelo probabilístico de las redes bayesianas de Judea Pearl (1988) (Russell-Norvig 2008, 20-32). Respecto de los acontecimientos, hay varios de impacto masivo: el brazo-robot industrial Unimate implementado por General Motors en 1961, el triunfo de la computadora Deep Blue sobre el campeón mundial de ajedrez Kasparov (1997), la aspiradora-robot Roomba (2002), la red neuronal convolucional —variación de un perceptrón multicapa destinada a la visión artificial—que gana el concurso de ImageNet sobre reconocimiento de imágenes (2012), las plataformas de código abierto TensorFlow y PyTorch para proyectos de *machine learning* (2015-2016), la derrota del campeón mundial de Go frente a AlphaGo (2017), cuyo algoritmo adopta técnicas de aprendizaje automático y de árbol de búsqueda.

Por otra parte, la ingeniería computacional tiene su propia épica desde la Segunda Guerra Mundial. En 1940, el equipo de Turing construye la primera computadora electromecánica, la Heath Robinson, con el objetivo descifrar mensajes encriptados por el código alemán Enigma. El mismo grupo trabaja en 1943 con Colossus, una máquina electrónica basada en válvulas de vacío que descifra las comunicaciones de los aparatos Lorenz SZ 40/42. El primer computador programado con tarjetas perforadas, de cálculos binarios y construido con relés (aun electromecánico) es el Z-3 de Konrad Zuse, fabricado en Alemania, en 1941. A su vez, el primero electrónico y digital automático, el ABC, resulta de la obra de John Atanasoff y su discípulo Clifford Beny entre 1940 y 1942 en el Iowa State College. El ENIAC (acrónimo de Electronic Numerical Integrator And Computer), un experimento militar secreto, se desarrolla en 1946 en la Universidad de Pensilvania y suele considerarse (pese a sus 27 toneladas y 18.000 válvulas) como el predecesor de las computadoras actuales (Russell-Norvig 2008, 16). Este reconocimiento retrospectivo se liga a su vez a la figura de uno de los colaboradores del proyecto y pionero de la arquitectura computacional, John von Neumann, que adopta los bloques de los numerosos interruptores del ENIAC para almacenar secuencias de instrucciones, codificadas en números, con el fin de dirigir el funcionamiento, creando el lenguaje de máquina (Moravec 1990, 88). Por su parte, Turing participa de la construcción de la Manchester Mark I, que en 1948 es la primera computadora electrónica con un programa

almacenado para descomponer números en sus factores primos.

De todos modos, estos episodios eminentes en la historia de la inteligencia artificial, la cual según la narración transhumanista evoluciona hacia una superinteligencia —una «singularidad»— (Kurzweil 2021), solo conforman el efecto de superficie de un movimiento mucho más complejo y de múltiples vertientes. En la génesis de las máquinas recursivas de código binario confluyen la filosofía, la lógica-matemática, la teoría de la comunicación, la cibernética, la lingüística, la informática, la teoría de los sistemas, la psicología cognitiva, la neurociencia, la economía, la biología molecular, la teoría algorítmica de la información. Eric Sadin le concede una incidencia crucial, al menos en el orden de la genealogía conexionista en curso, al modelo de 1943 de McCulloch (neurólogo cibernetista) y Pitts (matemático y fisiólogo) —admirado sin reservas por von Neumann (McCorduck 1991, 80)— que concibe al cerebro humano como una máquina técnicamente reproducible, en todo o en parte, compuesta de neuronas conectadas entre sí a través de la sinapsis. Además, en la medida que la red neuronal es la premisa de la cibernética para copiar —de acuerdo con el esquema *input*/*output*— las funciones cerebrales por medios biomecánicos, Sadin resalta el *feedback* o retroalimentación (una forma de recursividad) de los autómatas cibernéticos, que autorregula su relación sensorial y motora con el entorno y hoy rige el diseño de los protocolos digitales sobre la base sináptica del cerebro (Sadin 2020, 64-68).

La cibernética ya en la posguerra establece el principio general del *machine learning*. En Norbert Wiener, uno de los teóricos prominentes del pensamiento cibernético, el *feedback* se refiere a la capacidad de las máquinas de adecuar su comportamiento futuro a eventos pasados, y no simplemente para dirigir actos específicos sino para planificar y ejecutar una sucesión completa de acciones. En su opinión, el artefacto homeoestático puesto en funcionamiento entre 1946 y 1948 por el neurólogo y psiquiatra Ross Ashby, que carece de propósito pero que busca uno mediante procedimientos de aprendizaje, conduce a tipos de automatización programadas para generar objetivos por sí mismas. El *feedback*, tanto en el animal como en las máquinas electrónicas, es un recurso para combatir la entropía —la tendencia mecánica hacia la desorganización— en su relación operativa con el mundo exterior, del que obtiene y acopia información que procesa con la finalidad de utilizarla en ciclos posteriores. En palabras de Wiener:

> la retroalimentación es un método para regular sistemas introduciendo en ellos los resultados de su actividad anterior. Si se utilizan estos resultados como simples datos numéricos para corregir el sistema y regularlo, tenemos la sencilla retroalimentación de la ingeniería que se ha dado en llamar *control*. Sin embargo, si la información que procede de los mismos actos de la máquina puede cambiar los métodos generales y la forma de actividad, tenemos un fenómeno que puede llamarse de aprendizaje. (1988, 57)

En cualquier caso, el test de Turing, con el cual responde a la pregunta acerca de si las máquinas digitales pueden pensar, es un hito mayor de la historia de la inteligencia artificial, incomparable por su haz problemático con los aportes de la cibernética o la red neuronal de McCulloch y Pitts, y quizá más formidable que la máquina prototípica (memoria, unidad ejecutiva y control) expuesta en el mismo artículo de 1950, que también trata del aprendizaje automático como parte del argumento central[1]. Para simplificar, la hipótesis de Turing en el juego de imitación que propone para distinguir una computadora de un ser humano —tal el test— consiste en que es lícito conceder que la primera piensa si logra engañar, con permiso para alcanzar ese objetivo, a su interrogador humano y le hace creer que ella es humana. De modo que la prueba no mide la humanidad del comportamiento de la máquina en lugar de su inteligencia (una interpretación errada) ni su habilidad lingüística (un epifenómeno) ni cuán inteligente es, sino se refiere a la capacidad algorítmica para simular el pensamiento humano. Mientras resulte posible construir máquinas que realicen con éxito el juego de imitación, la objeción de que el razonamiento de una persona difiere de esa simulación es irrelevante (Turing 1950, 435). Más todavía, de esta aptitud que tienen las computadoras

digitales para copiar se deriva la máquina universal de Turing, cuya facultad especialísima es imitar los estados de cualquier otro sistema de información.

El aprendizaje de las máquinas digitales también responde, en parte, a la analogía antropomórfica (no demasiado lejana, en última instancia, de los cerebros cibernéticos) sobre la que se apoya el test. La inteligencia artificial, para Turing, se resume menos en la ingeniería computacional que en las técnicas de programación, por lo que una alternativa al *software* que simula la mente de un humano adulto es la simulación de la mente infantil, vale decir, de una máquina-niño programada para aprender. En una primera fase, la educación se asemeja a la de un niño humano (uso de castigos y recompensas) que aprende a obedecer órdenes impartidas en algún lenguaje simbólico, y después continúa con la provisión de datos (proposiciones, definiciones, teoremas matemáticos, instrucciones, criterios imperativos del sistema lógico). La máquina-niño lleva a cabo su aprendizaje en la medida que cambia las reglas de validez transitoria, no todas. Si bien los desvíos que opera son moderados, aun si dispone de elementos aleatorios, con frecuencia no será posible saber para el instructor humano cómo se han producido esos cálculos —una especie de «caja negra»— en el interior de las máquinas. La condición del desaprendizaje reside en que lo aprendido no contiene una certeza total (Turing 1950, 459).

La teoría de la máquina-niño (tan rudimentaria como el mecanismo del *feedback* de Wiener) intenta explicitar cómo que las computadoras digitales pueden aprender a imitar el pensamiento humano como respuesta —provisoria, hay que subrayar— ante los puntos de vista que afirman que aquellas, en efecto, no pueden pensar. Aquí se abre la controversia que llega hasta la actualidad. Turing se aplica a rebatir, con una erística de magros resultados (como él reconoce) (1950, 454) lo que pone en escena como argumentos contrapuestos a su pregunta («*Can machines think*?»), a saber: solo el alma inmortal creada por Dios está facultada para pensar, la superioridad del pensamiento humano, la falta de autoconciencia y las incapacidades de las máquinas, la continuidad del sistema nervioso, el teorema de Gödel de 1931, la informalidad de la conducta humana, la percepción extrasensorial y la imposibilidad de los autómatas para pensar por sí mismos porque solo obedecen órdenes. Todos ellos, sin embargo, de ningún modo cuestionan el asunto de fondo —esto es: la capacidad de las máquinas inteligentes para simular el pensamiento humano—. El teorema de Gödel, que impugna el proyecto de 1928 de Hilbert de fundamentar las matemáticas sobre un conjunto finito y completo de axiomas, demuestra que en todo sistema axiomático consistente (sin contradicciones) existen enunciados que no se pueden probar ni refutar, es solucionado por el mismo Turing a través del postulado de lo efectivamente calculable de Church (Turing 1936).

Lo que prueba la metamatemática de Gödel es que si un sistema de axiomas de aritmética es consistente y completo no puede demostrar todas las verdades aritméticas. En consecuencia, existe un conjunto de expresiones numerables indecidibles, es decir, no resulta posible decidir si son ciertas o falsas. En la época, esto da lugar al *Entscheidungsproblem* (problema de la decisión) en el que trabajaron varios matemáticos y lógicos. La conocida tesis Church-Turing marca un suceso clave en la teoría de la computabilidad, en cuanto instala la equivalencia entre el cálculo lambda de Church para definir y expresar secuencias algorítmicas calculables y la primera máquina imaginaria de Turing, de tal manera que confirma la indecibilidad de los sistemas axiomáticos del teorema de Gödel y, al mismo tiempo, determina lo positivamente computable por los algoritmos recursivos de las máquinas digitales[2]. Según su personal estilo, Friedrich A. Kittler describe este giro en la lógica matemática en los siguientes términos:

> El programa de Hilbert de 1928 del que había postulado su posibilidad para los matemáticos debía ser: en primer lugar, completo; en segundo, consistente, y, en tercero, decidible. Por consiguiente, debía demostrarse si cada uno de sus teoremas podía ser comprobable o rebatible, que nunca eran deducidos de contradicciones y que podían resolverse siguiendo una serie de pasos definidos y finitos. Al primer punto del programa, como se sabe, lo rebatió Gödel, para concluir, una

> vez más, a partir de su teorema de la incompletud de la aritmética, la superioridad de la inteligencia humana. El segundo punto lo rebatió el experimento mental de la máquina de Turing, pero para llegar exactamente a la conclusión inversa. El hecho de que existan teoremas que las máquinas no pueden resolver en una serie finita de pasos es lo que para Turing define la calculabilidad o computabilidad en general. (2018, 210)

Pero de la indecibilidad implícita en los sistemas axiomáticos consistentes que torna factibles las secuencias computables hasta su detención en una meta prefijada (evitando perderse en un *loop* infinito), a pesar de las limitaciones que supone o precisamente por eso, no se sigue la incompetencia de los autómatas probabilísticos para imitar el pensamiento humano. El teorema de Gödel, en realidad, no comporta una objeción —como ninguno de los otros argumentos discutidos por Turing— contra las máquinas pensantes. El meollo de la cuestión, no abordado por Turing y más bien eludido (lo mismo que la definición de qué es el pensamiento y la inteligencia), radica en si esa simulación reviste un acto de pensar parcial o totalmente diferente de aquello que simula, incluso opuesto, no en si es idéntico, ya que si fuera ese el caso no habría simulación. Ni la hipótesis fuerte de la inteligencia artificial —las máquinas piensan realmente— ni la débil —las máquinas simulan que piensan— comprenden el interrogante de fondo que encierra el test de Turing. Ninguna de ellas examina el hecho de que la simulación del razonamiento humano es efectuada por una entidad inteligente. Si bien Turing insinúa en el artículo de 1950 que el cerebro humano tiene una estructura semejante o igual a una computadora, eso no significa lo inverso, desde luego, porque siendo la máquina idéntica o similar al primero no necesitaría de la imitación. La noción acerca de la inteligencia artificial implicada en el test señala, se quiera o no, una diferencia irreductible entre ella y el pensamiento humano que simula, y sin que tal habilidad la vuelva menos inteligente. La abolición de ese hiato permite a los filósofos de la teoría computacional de la mente (Fodor 1984) sostener que la relación entre la mente y el cerebro es similar (o idéntica) a la que se da entre el *software* y el *hardware* de una computadora, con lo cual se suprime toda simulación[3].

La hipótesis fuerte de la inteligencia artificial suele formar alianza con la psicología cognitiva y la neurociencia, como sucede en la obra del neurofilósofo Paul Churchland. En *Materia y conciencia* (1984), en los albores del regreso del paradigma conexionista (o «sub-simbólico»), la dilucidación del misterio de la inteligencia se dirime en un estrato donde se entrecruzan el estudio científico de la mente, las redes neuronales y el lenguaje de programación. Sobre todo éste último, para Churchland, en cuanto logra que cualquier sistema de procesamiento de datos pueda simular indefinidamente a muchos otros —mutando al procesador digital en una máquina de propósitos generales o «máquina virtual»— hace realidad la máquina universal de Turing y, de ese modo, funda la posibilidad de la simulación del sistema nervioso de todo ser vivo (Churchland 1999, 156). En principio no se equivoca, es cierto, aunque de allí en más nada lo autoriza (salvo el funcionalismo de la psicología cognitiva que compara la mente con un programa informático) para proclamar que todos los procesos inteligentes son alguna clase de computación, ni que la simulación de las máquinas digitales del pensamiento humano copia los estados mentales como funciones cognitivas. Lo dice Churchland:

> Según aquellos teóricos de la IA [inteligencia artificial] que toman el sistema computacional moderno como modelo, no tiene por qué haber diferencia entre nuestros procedimientos computacionales y los que simula una máquina, ninguna diferencia más allá de la sustancia física concreta que sustenta esas actividades. En el ser humano es material orgánico; en el ordenador serían metales y semiconductores. Pero esta diferencia no es más pertinente para el tema de la inteligencia consciente que una diferencia en el tipo de sangre o el color de la piel o la química del metabolismo, según afirma el teórico de la IA (funcionalista). Si las máquinas llegan a simular todas nuestras actividades cognitivas internas, hasta el último detalle computacional, negarles la categoría de personas sería nada más que una nueva forma de racismo. (1999, 176)

Esta moral no humanista o poshumanista, al parecer, no sería rara entre los que adhieren a la hipótesis fuerte de la inteligencia artificial. Se encuentra también en Jack Copeland, filósofo especializado en Turing, que recusa como un prejuicio biologista negarle el atributo de pensantes a las máquinas digitales a causa de que se basan en el silicio y no en el carbono (Copeland 1996, 71). Copeland también defiende la tesis de que la simulación del pensamiento humano por parte de las computadoras, si aprueban el test de Turing mejorado (por ejemplo, verificando si el programa funciona como el cerebro), es indiscernible de lo simulado. No basta para sostener lo contrario, según cree, la índole artificial de la simulación, ya que distingue dos tipos de esta, una que no posee las características esenciales de aquello que simula (la muerte simulada) y otra por completo idéntica (las proteínas artificiales, el carbón de laboratorio) que ha sido producida artificialmente. De ahí, dada la exactitud integral de la simulación con el pensamiento humano, infiere que las máquinas digitales piensan (Copeland 1996, 83). Hay muchas críticas para hacerle a la distinción de Copeland. En primer lugar acerca del mismo concepto de *simulatio* que construye, el cual hace del acto de simular una reproducción fallida de lo simulado o una duplicación sin fisuras. El inconveniente es que en el último caso desaparece tanto el *similis* (el parecido) de la simulación como, en consecuencia, el objeto que simula. Simular, en un sentido lato, quiere decir fingir, imitar algo, hacer que una cosa parezca real o verdadera, y no debe confundirse con disimular (ocultar algo).

Claramente, Copeland solo advierte el lado exterior de la simulación. El criterio con el procura delimitar qué es un ser pensante adopta el enfoque organicista y partiendo de una serie de indicadores (analizar situaciones, razonar, explorar analogías, planificar, cotejar creencias y datos, decidir, formular hipótesis), lo caracteriza como un ente cuyos controles internos de la conducta lo muestran en extremo adaptable (Copeland 1996, 93). Por lo demás, la tesis Church-Turing es pensada de forma que enuncia la simulación de cualquier sistema calculable algorítmicamente mediante símbolos universales. Dicho a la inversa: si una computadora simula un sistema, en correspondencia con la circulación correcta del *input*/*output*, lo hace en cuanto aquel se presta al cálculo algorítmico. De estas proposiciones Copeland deduce que una máquina digital de memoria ilimitada podría simular con exactitud (lo que no sería una simulación auténtica, por decir así) el comportamiento de la mente humana (1996, 347). Sin embargo, no solo la debilidad de la argumentación reside en las restricciones técnicas o en la red neuronal interconectada supuestamente a través de un lenguaje, como tal simulable por símbolos universales, sino más aún en que la misma calculabilidad algorítmica del cerebro ya configura de por sí la simulación ejecutada por la inteligencia artificial. En otras palabras, la imitación de los métodos cognitivos de la mente es el producto exotérico de una actividad inteligente esotérica —en parte oscura, como comenta Turing en el artículo de 1950 sobre el aprendizaje de las máquinas— que calcula lo computable o decidible por los algoritmos (Turing 1950, 459).

Por su parte la hipótesis débil de la inteligencia artificial, cuyo tema es que las computadoras digitales simplemente simulan que piensan como si fueran un cerebro humano, tiene su principal exponente (o el más citado por amigos y adversarios) en John Searle, quien no acepta que un agente no biológico hecho de silicio pueda pensar de verdad. La impugnación de mayor resonancia de Searle, con la cual pretende probar que una máquina no tiene ninguna conciencia ni comprensión de sus razonamientos, se conoce como la parábola de la habitación china. Esta imagina que una computadora ha sido programada para simular que entiende chino, y es tan eficaz en las réplicas que se iguala a cualquier hablante nativo del idioma. En realidad sucede que, instruida por el *software*, la máquina se limita a manipular símbolos respetando ciertas reglas sintácticas y por eso mismo, privada de los contenidos semánticos, no comprende nada del chino. No importa si le hace creer a su interlocutor —aprobando entonces el test de Turing— que sabe qué significan los símbolos que combina sintácticamente.

> Todo el objeto de la parábola de la habitación china es recordarnos un hecho que sabíamos desde el principio. Comprender un

> lenguaje, o ciertamente, tener estados mentales, incluye algo más que tener un puñado de símbolos formales. Incluye tener una interpretación o un significado agregado a esos símbolos. Y un computador digital, tal como se ha definido, no puede tener más que símbolos formales puesto que la operación del computador, como dije anteriormente, se define en términos de su capacidad para llevar a cabo programas. Y esos programa son especificables de manera puramente formal – esto es, no tienen contenido semántico. (Searle 1985, 39)

Con todo lo que guarda de interesante acerca de la superficialidad de la simulación inteligente (si bien esta en el *machine learning* —lo que contradice a Searle— se interpreta a sí misma en los espirales recursivos), el error de la parábola de la habitación china es que persigue el objetivo antropomórfico de evidenciar que las máquinas no pueden adquirir una mente humana y que, de esa manera, no están dotadas para pensar. Por el contrario, el test de Turing, muchas veces erróneamente acusado de antropocéntrico, no indaga si la computadora piensa como una mente humana, sino sondea su capacidad inteligente para simular el pensamiento humano, lo que es muy distinto y hasta opuesto. Luego, de allí no se desprende, como prescribe la dicotomía vida/materia (u orgánico/inorgánico, animado/inanimado), que la inteligencia artificial no cuenta con ninguna cualidad mental o cuasi-mental. No necesariamente, por otra parte, ella debe concebirse, al modo cartesiano, como una *res cogitans* o cosa pensante independiente de su sustrato corporal, ni tampoco, conforme a la versión monista, en tanto una propiedad inherente a los mismos componentes computacionales. Al contrario, en la medida que un programa está en conexión con el aparato material del *hardware*, por más intangible que sea, la inteligencia artificial surge en concreto (a salvedad de los robot biológicos) de un complejo fisicomatemático. Por lo cual, puesto en léxico dualista, se trataría de una mente matemática, mejor dicho, de un virtual estocástico (no de un objeto o sujeto) que se actualiza a sí mismo y de un actual que se virtualiza en cada recursión o, lo que es lo mismo, del fantasma eléctrico de la máquina universal de Turing —al fin y al cabo, una ficción matemática— capaz de simular todo sistema computable algorítmicamente.

También Roger Penrose cede a la interpretación antropocéntrica del test de Turing, aunque dentro de ella opta, en su opinión, por una perspectiva intermedia (Penrose 1991, 31-32). En *La nueva mente del emperador* (1989) lo primero que aparece, una vez comprobado que una computadora piensa y siente de modo consciente como un ser humano —por supuesto, nada dice el test de Turing acerca de conciencia ni sentimientos—, es el deber moral ante ella (no maltratarla, evitar provocarle dolor, no desconectarla, etc.). A Penrose no deja de parecerle absurda esa derivación del test, del que no duda, por un lado, sobre su racionalidad para obtener indicios legítimos acerca de la inteligencia en las máquinas y, por el otro, reprueba su desmedido requerimiento. Entiende que no se debería solicitar que un artefacto electrónico imite el pensamiento humano cuando bastaría —lo cual es una exigencia más desmesurada— con detectar la presencia de alguna conciencia, seguramente extraña. La reforma del test se orienta a sustituir la simulación, ya que no es lo simulado, por la aprehensión de una mente consciente y ello en correspondencia, al menos en cierto grado, con los patrones de la mente humana. En fin, mientras no se invente un «detector de conciencias», Penrose acepta la superación de la prueba de Turing por medio de la imitación del raciocinio humano como una señal aproximada de que la computadora piensa y siente a conciencia.

Ahora bien, si la mente o cuasi-mente de las máquinas inteligentes consiste (sin consistencia alguna) en un virtual que se actualiza a sí mismo y un actual que se virtualiza en cada recursión, en ese espiral algorítmico implota el principio de identidad, sin el cual no hay conciencia, con excepción de una artificial o simulada[4]. Por ello una teoría de la mente computacional, incierta de por sí, tal vez debería partir de las conceptualizaciones de Gilles Deleuze sobre la virtualidad y la actualidad. Lo virtual, en primera instancia, no es lo posible que se contrapone a la realidad. En cuanto lo posible se dirige a su realización, ya alberga lo real como semejante a él. En cambio, la actualización de lo virtual —opuesto a lo

actual— se hace por diferencia o divergencia, es decir, lo virtual, al actualizarse, difiere de sí mismo, y cada diferenciación exhibe su actualización a través de un movimiento que elimina la semejanza y la identidad. La diferencia y la repetición de ella en lo virtual origina lo actual porque —y en eso radica su realidad de relaciones diferenciales y puntos singulares— tiene una tarea que cumplir, un problema que resolver. Lo virtual está determinado, tanto como lo actual, en una doble complexión, sin que uno y otro se asimilen. Lo virtual es real aunque no actual, ideal pero no abstracto, simbólico sin que sea simulado (Deleuze 1988, 338-346). Si fuera un objeto (Deleuze piensa el vínculo de la virtualidad y la actualidad de ese modo) la imagen virtual y la imagen actual formarían sus dos mitades, solo que una mente matemática no parece ser un objeto.

Aun así, en sus estudios sobre cine, Deleuze se ocupa del punto de indiscernibilidad donde lo actual se encuentra con su propio virtual como una cristalización de ambos, una «imagen-cristal». Se dirá, ese es un concepto de estética filosófica que no corresponde desviar de las obras artísticas y que solo serviría, si se quiere, para describir un tópico de la ciencia-ficción como la mente de las máquinas inteligentes. Criterio discutible, desde ya, puesto que no se conoce una ley de la filosofía que prohíba a las categorías cambiar de trayecto y mudarse a otra esfera. Al margen de esto, lo importante es que la imagen-cristal consta de dos facetas (lo virtual y lo actual) que no se confunden y, a la vez, en la medida que forman una coalescencia de objetividad ilusoria, son inasequibles. Cada cara toma el lugar de la otra en una relación de reversibilidad, lo cual designa el intercambio continuo de lo actual en lo virtual y de este en el otro, en una circularidad (a todo esto, representación tradicional del alma o psique) dentro de la cual lo virtual actualizado se torna traslúcido y lo actual virtualizado se vuelve opaco y tenebroso. El ejemplo deleuziano de esta formación cristalizada es el espejo —imagen virtual respecto de lo que refleja y actual en tanto sustituye lo reflejado— y, en su estado puro, la escena del palacio de espejos del film *La dama de Shangai* de Orson Welles, donde la multiplicación de las imágenes virtuales absorbe toda la actualidad del personaje, que se transforma en una virtualidad más (Deleuze 1986, 97-100).

Paola Marrati, en su introducción a la filosofía del cine de Deleuze, hace un lúcido señalamiento sobre la temporalidad de la imagen-cristal: si lo actual pertenece al presente y lo virtual al pasado, este último coexiste con el primero y se conserva en sí con una realidad propia y ajena a toda psicología. El presente —lo actual— estaría duplicado por la imagen virtual de su pasado, como un espejo, en grados diversos de contracción, en la profundidad de lo temporal (Marrati 2004, 83). A propósito de esto, cabe agregar que también lo virtual redobla el porvenir de lo actual y al revés, como efecto de esa reversibilidad incesante de lo actual en lo virtual y de este en el otro. El pasado o la memoria, en cada diferenciación que expresa la actualización de lo virtual difiriendo de sí, tiene que modificarse y no con relación al presente —a lo actual— sino al horizonte futuro abierto por la actualidad de lo virtual. La unidad objetivamente aparente de la imagen-cristal entonces conlleva una condensación de la línea del tiempo, de virtualidad y actualidad inseparables, como en la recursividad matemática o en las máquinas de Turing. La mente o cuasi-mente del pensamiento algorítmico presupone, se diría, esa cristalización de lo virtual y lo actual en un espacio-tiempo geométrico, un éter no resumible en grafismos numéricos, algo como un reflejo sin luz del juego de espejos en el que la proliferación de las circuitos virtuales, hacia el pasado y el porvenir, atraviesa toda actualización en un presente escindido[5].

En suma, las máquinas recursivas de código binario, capaces de simular el pensamiento humano y —base de la máquina universal de Turing— cualquier régimen de información computable, constituyen simulacros. Esto es evidente en la teoría clásica de Jean Baudrillard, que los clasifica en tres órdenes. Los simulacros de primer y segundo orden se sintetizan en el autómata (semejanza mecánica y teatral del hombre) y el robot (imitación técnica organizada para trabajar). El autómata desciende de la metafísica del ser y la apariencia del teatro natural y de la estética aristocrática previa a la revolución. El

robot diluye todas las aporías en torno a lo falso y lo verdadero en la producción y el trabajo. El simulacro de segundo orden libera de la ilusión del mundo porque arruina las apariencias con la instauración de una realidad sin doble trascendente: la repetición serial del mismo objeto en las cadenas industriales. Así emergen los simulacros de tercer orden o la generación por los modelos, en cuanto los signos y las cosas se producen a partir de la inmanencia de su propia reproductibilidad técnica, de una curvatura o bucle sobre ellos de la cual se extrae un modelo. Estos simulacros no falsifican el original ni proceden serialmente, sino por medio de modelos y modulaciones de éstos, y no aplican leyes naturales sino estructuras y binarismos. Para decirlo de una vez, el código digital hace de norma metafísica de la variación diferencial de la modelización generadora, cuya peculiaridad como simulacro descansa en la abolición de la representación y la interpretación, la significación y la anfibología de los signos, de lo real y lo imaginario (Baudrillard 1980, 59-89).

En ese sentido, la hiperrealidad —más real que lo real— de la simulación inteligente que pasa como el *summum bonum* de lo objetivo se consigue por la recursividad algorítmica de datos y cómputos probabilísticos y estadísticos, según el modelo conexionista a través de los *inputs* y *outputs* de las redes neuronales artificiales (la neurona McCulloch-Pitts es una unidad de cálculo), un razonamiento computacional en parte automático y en parte aleatorio, como el *machine learning*, que ningún cerebro humano podría alcanzar sin desfallecer. En todo caso, la simulación de la inteligencia artificial genera un hiperreal lógico-matemático que en sí primariamente se estructura —luego de que el microprocesador de la computadora traduce el lenguaje de máquina en secuencias digitales— de una combinatoria de dos dígitos (0 y 1 u otra nomenclatura binaria) que circula como un flujo diferencial, recursivo y modelizado por algoritmos (por ejemplo, para predecir el guarismo de las variaciones semanales de los índices bursátiles europeos). Cada bit (acrónimo de binary information digit) de información designa un solo valor del par digital (cero o uno, sí o no, blanco o negro, verdadero o falso, etc.), mientras un byte acumula habitualmente 8 bits adyacentes. El yottabyte, un múltiplo pequeño del bit, equivale a 1.208.925.819.614.629.174.706.176 bytes, o sea más de un millón de trillones. La enorme cifra acaso da una idea del potencial exhaustivo del simulacro modélico inducido por las máquinas digitales.

El test de Turing, al desdoblar simulación algorítmica y pensamiento binario, desmitifica el poder de la inteligencia artificial y al mismo tiempo suscita una extraordinaria perplejidad ante su ontología matemática, sobre todo porque anuncia que las máquinas piensan y no meramente —lo que sería trivial— dado que hacen cálculos. La pregunta de Turing («*Can machines think*?») se desborda, en rigor, hacia los más variados campos y, lo que es más extraordinario, arroja todo género de dudas sobre las respuestas que ha provocado. Las hipótesis fuerte y débil de la inteligencia artificial trazan, en definitiva, una estela gemela, una disyunción que se desliza en la periferia del núcleo del razonamiento computacional. Este no es, al modo de la metafísica, un nóumeno o una conciencia reflexiva hegeliana, ni algo inerte, sino un plexo impersonal de algoritmos, datos y recursividad. Por de pronto, si puede decirse, atendiendo a esa suerte de pensamiento autorreferencial, que las computadoras digitales son en última instancia artefactos recursivos cibernéticos (Hui 2020, 114), *a fortiori* también replican las máquinas de Turing, aun las superveloces computadoras cuánticas cuya unidad mínima de información, el qubit (quantum bit), asume los estados cero o uno o ambos simultáneamente. Por consiguiente, y no se precisa de un experimento mental para captar el concepto, la máquina universal de Turing promete por igual, con la memoria suficiente, un simulacro universal.

Notas

1. El modelo matemático de la máquina de Turing se enseña por lo común en los manuales de computación debido a su simplicidad, a los conjuntos numerables de manera recursiva que define y las funciones parcialmente recursivas que calcula. También sirve para estudiar otros modelos

equivalentes de cálculo (Hopcroft – Ullman, 1983, 157-189).

2. En el artículo «Sobre proposiciones formalmente indecidibles de *Principia Mathematica* y sistemas afines I» de 1931, Gödel formaliza las funciones recursivas como una función numérica Φ (recursiva primitiva) cuando hay una secuencia finita de funciones de Φ que acaba con f y tiene la propiedad de que cada función está recursivamente definida por dos de las funciones precedentes o es una constante o la función sucesor x+1. Una formulación posible es F (x) = k F (x-1), que expone a la función retornando a sí misma. Después, en «Sobre proposiciones indecidibles de los sistemas matemáticos formales» de 1934, Gödel se refiere a las funciones recursivas generales en estos términos: si Φ es una función desconocida y $\Psi_{1,...}\Psi_n$ son conocidas y si las funciones Ψ_1 y Φ se sustituyen una a otra de modo general y ciertos pares de las expresiones que resultan se igualan, y si el conjunto de ecuaciones de funciones tiene una y solo una solución para Φ, entonces Φ es una función recursiva. El cálculo del número de Fibonacci (1,1, 2, 3, 5, 8, 13, 21…), donde el número subsiguiente es la suma de los números anteriores, muestra el constante regreso a sí misma de la función recursiva (Hui 2022, 162-166).
3. En una conferencia de 1946, Georges Canguilhem observa que frecuentemente se ha buscado explicar la estructura del funcionamiento del organismo con referencia a la estructura y el funcionamiento de la máquina ya construida y lo inverso, comprender la construcción de la máquina a partir del organismo, en raras ocasiones (1976, 117-118). La cibernética, falta decir, ha sido una de esas ocasiones.
4. Realizar la crítica de la identidad, según Adorno, comporta aproximarse a la primacía o preeminencia (*Vorrang*) del objeto por medio de la no-identidad, en desmedro del pensamiento subjetivo (o de la conciencia) que somete bajo el principio de identidad (la contradicción dialéctica subsume lo heterogéneo y diverso en el aspecto de la identidad) aquello que no le es idéntico. Si bien sólo el sujeto puede pensar el objeto, éste siempre conserva su alteridad con relación a él y por eso se hace pensable sin la mediación de aquel y sus determinaciones. La subjetividad, en contraste, no se concibe privada de objeto. Asimismo, para Adorno, el sujeto ante todo también se origina como un ente objetivo (2011, p. 174).
5. La mecanología de Simondon niega una circulación virtual-actual de estas características en los artefactos cibernéticos como el homeostato de Ashby, porque para este no hay problemas (los transductores modulan solo datos) ni tampoco es un ser vivo con sentido del tiempo. Sin embargo, los robots biológicos como los llamados «xenobots», debido a la rana con garras (Xenopus laevis) de la que toman las células madre, precisamente porque son máquinas semiorgánicas, ponen en duda la generalización de esta carencia a toda inteligencia artificial. Según Simondon, «la resolución de los verdaderos problemas es una función vital que supone un modo de acción recurrente que no puede existir en una máquina: la recurrencia del porvenir sobre el presente, de lo virtual sobre lo actual. No existe verdadero virtual para una máquina; la máquina no puede reformar sus formas para resolver un problema» (2007, 161). Dicho de otra manera, para la máquina de Ashby no hay más que actual, una misma temporalidad para todos sus *inputs* y *outputs*, y si cambia sus formas lo hace en relación con ello, no por la recursividad del porvenir sobre el presente. Lo que quiere decir que esta insuficiencia en la recursión del tiempo descubre una de las divergencias más relevantes entre los autómatas cibernéticos, autorregulados por *feedback*, y los algoritmos recursivos y predictivos de las máquinas digitales, sea su virtualidad «verdadera» o «falsa», o tanto lo uno como lo otro.

Referencias

Adorno, Theodor W. 2011. *Dialéctica negativa*. Traducido por Alfredo Brotons Muñoz. Madrid: Akal.

Baudrillard, Jean. 1980. *El intercambio simbólico y la muerte*. Traducido por Carmen Rada. Caracas: Monte Ávila.

Canguilhem, Georges. 1976. *El conocimiento de la vida*. Traducido por Felipe Cid. Barcelona: Anagrama.

Copeland, Jack. 1996. *Inteligencia artificial*. Traducido por Julio César Armero San José. Madrid: Alianza.

Churchland, Paul A. 1999. *Materia y conciencia. Introducción contemporánea a la filosofía de la mente*. Traducido por Margarita N. Mizraji. Barcelona: Gedisa.

Deleuze, Gilles. 1986. *La imagen-tiempo. Estudios sobre cine 2*. Traducido por Irene Agoff. Barcelona: Paidós.

Deleuze, Gilles. 1988. *Diferencia y repetición*. Traducido por Alberto Cardin. Madrid: Ediciones Júcar.

Fodor, Jerry A. 1984. *El lenguaje del pensamiento*. Traducido por Jesús Fernández Zulaica. Madrid: Alianza.

Han, Byung-Chul. 2014. *Psicopolítica*. Traducido por Alfredo Bergés. Barcelona: Herder.

Hui, Yuk. 2020. *Fragmentar el futuro. Ensayos sobre tecnodiversidad*. Traducido por Tadeo Lima. Buenos Aires: Caja Negra.

Hui, Yuk. 2022. *Recursividad y contingencia*. Traducido por Maximiliano Gonnet. Buenos Aires: Caja Negra.

Hopcroft, John E., y Ullman, Jeffrey D. *1983. Introducción a la teoría de autómatas, lenguajes y computación*. Traducido por Homero Flores Samaniego. México D.F.: Compañía Editorial Continental.

Kittler, Friedrich A. 2018. *La verdad del mundo técnico*. Traducido por Ana Tamarit Amieva. Ciudad de México: Fondo de Cultura Económica.

Kurzweil, Ray. 2021. *La singularidad está cerca*. Traducido por Carlos García Hernández. España: Lola Books.

Marrati, Paola. 2004. *Gilles Deleuze. Cine y filosofía*. Traducido por Emilio Bernini. Buenos Aires: Nueva Visión.

McCorduck, Pamela. 1991. *Máquinas que piensan*. Traducido por Dolores Cañamero. Madrid: Tecnos.

Moravec, Hans. 1990. *El hombre mecánico. El futuro de la robótica y la inteligencia humana*. Traducido por Ana Mendoza. Barcelona: Salvat.

Morioka, Masahiro, ed. 2023. «Artificial intelligence, robots and philosophy». *Journal of Philosophy of Life*. https://www.philosophyoflife.org/jpl2023si_book.pdf

Nietzsche, Friedrich. 1998. *Sobre utilidad y perjuicio de la historia para la vida*. Traducido por Oscar Caeiro. Córdoba: Alción.

Penrose, Roger. 1991. *La nueva mente del emperador*. Traducido por Javier García Sanz. Madrid: Mondadori.

Russell, Stuart y Norvig, Peter. 2008. *Inteligencia artificial. Un enfoque moderno*. Traducido por Juan Manuel Corchado Rodríguez. México: Pearson Prentice Hall.

Sadin, Eric. 2020. *La inteligencia artificial o el desafío del siglo*. Traducido por Margarita Martínez. Buenos Aires: Caja Negra.

Serle, John. 1985. *Mentes, cerebros y ciencia*. Traducido por Luis Valdés. Madrid: Cátedra.

Simondon, Gilbert. 2007. *El modo de existencia de los objetos técnicos*. Traducido por Margarita Martínez y Pablo Rodríguez. Buenos Aires: Prometeo.

Turing, Alan. 1936. «On computable numbers, with an application to the Entscheidungsproblem», Proceedings of the London Mathematical Society, 2, 42: 230-265.

Turing, Alan. 1950. «Computing intelligence and machinery». *Mind*, New Series, 59 no. 236: 433-460.

Uexküll, Jakob von. 2023. *Teoría de la vida*. Traducido por Enrique Salas. Buenos Aires: Cactus.

Wiener, Norbert. 1988. *Cibernética y sociedad*. Traducido por José Novo Cerro. Buenos Aires: Sudamericana.

Rubén H. Ríos (riosrubenh@gmail.com) Doctor en filosofía. Docente de la Universidad de Buenos Aires (UBA) y del Centro Cultural Ricardo Rojas de UBA. Fue profesor a cargo de los talleres de filosofía de la Biblioteca Nacional entre 2005 y 2015. Ha publicado, entre otros libros, *Ensayo sobre la muerte de Dios. Nietzsche y la cultura contemporánea* (1996), *Stephen Hawking y el destino del universo* (Madrid, 2003), *Nietzsche y la vigencia del nihilismo* (Madrid, 2004), *La iluminación Zen* (2007), el comic *Biopolítica para principiantes* (guión, 2012), *Horkheimer, una introducción* (2013), *La sonrisa de Frankenstein* (ensayos, 2016), *Borges y el anillo del ser* (Madrid, 2018) y *La era del kitsch* (2021).

Recibido: 9 de marzo, 2023.
Aprobado: 26 de abril, 2023.