ISSN: 2683-3247

HUMANITAS

REVISTA DE TEORÍA, CRÍTICA Y ESTUDIOS LITERARIOS

Vol. 4 Núm. 8 Enero-Junio 2025

UANL

CENTRO DE ESTUDIOS HUMANÍSTICOS

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

Humanitas
Revista de Teoría, Crítica y Estudios Literarios

Paradigmas e Imaginarios Literario-Filosóficos Sobre la Inteligencia Artifical en "Maniac" de Benjamín Labatut.

Literary-Philosophical Paradigms and Imaginaries on Atificial Intelligence in "Maniac" by Benjamín Labatut.

Rubén Gutiérrez Guajardo
Universidad de Monterrey
Monterrey, México
orcid.org/0000-0002-3192-3751

Fecha entrega: 11-07-2024 **Fecha aceptación:** 20-10-2025

Editor: Víctor Barrera Enderle. Universidad Autónoma de Nuevo León, Centro de Estudios Humanísticos, Monterrey, Nuevo León, México.

DOI: https://doi.org/10.29105/revistahumanitas4.8-111
Email: ruben.gutierrezg@udem.edu

Humanitas, vol. 4, núm. 8, 2025

Paradigmas e Imaginarios Literario-Filosóficos Sobre la Inteligencia Artificial en "Maniac" de Benjamín Labatut.

Literary-Philosophical Paradigms and Imaginaries on Artificial Intelligence in "Maniac" by Benjamín Labatut.

Rubén Gutiérrez Guajardo
Universidad de Monterrey
Monterrey, México
ruben.gutierrezg@udem.edu

Sobre Labatut, su obra y el propósito de este artículo

Apenas a un año de su aparición, *Maniac* de Benjamín Labatut (2023) se ha convertido en un éxito literario. Sea debido a la prosa innovadora de su autor o a la influencia que el tópico de las implicaciones éticas de la ciencia y tecnología ha cobrado en la actualidad, resulta evidente que nos encontramos ante una obra no sólo de una extraordinaria calidad literaria, sino también ante una obra clave para pensar las problemáticas actuales acerca de la ciencia y la tecnología, y en especial, de las tecnologías emergentes de Inteligencia Artificial (IA).

Benjamín Labatut nació en Rotterdam, la tierra del humanista Erasmo, en el año de 1980. Según la biografía que la editorial Anagrama (2023) ha adjuntado en sus obras: Pasó su infancia en la Haya y a los catorce se estableció en Santiago de Chile, motivo por el cual ostenta la doble nacionalidad holandesa y chilena. Algunas de sus obras como *La Ant*ártica empieza aquí (2012), *Un verdor terrible (2020), La Piedra de la Locura (2021) y Maniac (2023)* han cobrado relevancia internacional debido a su abordaje literario con implicaciones filosóficas acerca de las relaciones entre la historia, la ciencia y la tecnología.

Es importante señalar que la carrera literaria de Labatut ha sido objeto de numerosos premios y condecoraciones que lo sitúan como una de las promesas de la literatura contemporánea. La crítica literaria internacional se ha desbordado en elogios respecto a su obra, tales como los que se muestran en la más reciente edición de su obra en la editorial Anagrama al respecto de *Maniac*:

> Monstruosamente bueno. Se lee como un oscuro mito fundacional sobre la tecnología moderna, pero con el ritmo de un thriller" (Mark Haddon). "Brillante, peculiar en el mejor sentido de la palabra. Te deja sin aliento" (Sasha Marianna Salzmann). "Labatut es ese fenómeno cada vez menos común. Incluso más que *Un verdor terrible, MANIAC* es una obra de belleza oscura, inquietante y singular" (Becca Rothfield, *The Washington Post*). (Labatut, 2023).

El objetivo de este artículo es presentar cómo la creación literaria de Labatut y sus intuiciones en torno al tema de los alcances y límites de la IA se encuentran enmarcadas en las disputas antropológicas y de filosofía de la mente que se han dado desde la modernidad en Occidente. En el caso de nuestro autor, encontramos un posicionamiento de clara demarcación entre el pensamiento

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

humano y el de los sistemas operativos de IA desde una perspectiva literaria.

El análisis filosófico-literario que se propone en el siguiente artículo se sitúa de manera especial en el último apartado de la novela, titulado *Lee o los delirios de la inteligencia artificial,* el cual hace referencia al acontecimiento real del encuentro entre el jugador profesional de Go[1] Lee Sedol y el sistema operativo AlphaGo, llevada a cabo entre el 9 y el 15 de marzo de 2016,[2] y que supuso un acontecimiento trascendente para comprender las revoluciones informáticas de nuestra época.

La pretensión del artículo es la de establecer algunos paralelismos y vinculaciones entre la narrativa creada por Labatut en *Lee o los delirios de la inteligencia artificial* con algunas de las problemáticas en torno al problema mente-cerebro, la consciencia y la inteligencia de las máquinas, tal como han sido planteadas en las disputas de la filosofía de la mente contemporánea. Con apoyo del análisis de las teorías de algunos filósofos de la mente como Putnam y Searle, se pretende establecer la tesis según la cual los planteamientos narrativos de Labatut ilustran a la perfección los paradigmas actuales acerca de los alcances y límites de la IA desde una perspectiva literaria.

[1] Según la Enciclopedia Britannica (2024), el famoso juego chino *Go,* se remonta a milenios de antigüedad. Es un juego de estrategia de gran complejidad, mayor que el ajedrez, en el que los jugadores deben conquistar el mayor territorio posible colocando piedras blancas y negras sobre un tablero. Por corresponderse con la tradición oriental, no es tan conocido en Occidente.

[2] El encuentro de cinco juegos entre el sistema operativo AlphaGo, desarrollado por Deep Mind, y Lee Sedol fue real, y los resultados del mismo se corresponden fielmente con la narrativa de Labatut. La memoria histórica del encuentro puede encontrarse en la web: https://web.archive.org/web/20160225112928/http://www.deepmind.com/alpha-go.html

El método a seguir es analítico y sintético. Se pretende abordar análisis conceptuales de los términos filosóficos clave en la discusión del texto de Labatut desde la filosofía de la mente, conceptualizar las propuestas hasta su desarrollo actual para luego extraer conclusiones filosóficas relevantes y aplicables sobre los alcances y límites de la IA a la luz de los planteamientos literarios, entendiéndose este proceso como una crítica filosófico-literaria a la razón tecnológica.[3]

¿Pueden pensar los ordenadores? Turing y los antecedentes del funcionalismo computacional

Toda la problemática de interés filosófico relativa a la cuestión moderna de la IA puede rastrearse hasta algunas cuestiones filosóficas clave y de larga tradición intelectual en Occidente; a saber, el antiguo problema mente-cuerpo, el problema de la emergencia de la consciencia y la discusión sobre el concepto y los atributos del término inteligencia, como propiedad exclusiva o no de los seres humanos, así como su eventual desarrollo en las máquinas o sistemas operativos.

[3] En lo que respecta a nuestro conocimiento y alcance, encontramos estudios literarios, mayoritariamente reseñas, sobre la obra literaria de Benjamin Labatut. Entre ellos destacan el ensayo *Las partículas elementales de Benjamin Labatut* de Pedro Pablo Guerrero, publicado en la revista Latin American Literature Today; la reseña *Mano de Dios. La visión infernal de la física teórica de Benjamin Labatut* de Richard Lea, publicado en la revista The Times Literary Supplement; y el artículo: *The intrusion of Gaia in Un verdor terrible (2020) by Benjamín Labatut* del profesor Aníbal Gabriel Carrasco de la Universidad de Concepción, Chile. Sobre *Maniac*, objeto de análisis de este artículo, destacan la reseña *Los delirios de la razón* de Julio José Ordovás, publicada en la revista *Letras Libres,* y el artículo *The other Physicist, Review of The Maniac by Benjamin Labatut* de Guy Stevenson, publicado en la revista *Literary Review.*

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

Estas problemáticas pertenecen por derecho propio a la llamada *filosofía de la mente* y se abordan desde esta perspectiva como los ejes vectores del análisis propuesto. Pensar en ellas implica adentrarse en los acercamientos que, al menos desde la Modernidad, se han dado a partir de Descartes, pasando luego por los empirismos, materialismos, funcionalismos y otras corrientes que han buscado explicar la aparición o emergencia de la consciencia mediante el desarrollo de procesos físicos, orgánicos o funcionales, sin recurrir a instancias externas de explicación como las de las tradiciones antiguas.[4]

El abordaje de este antiguo problema cartesiano y las respuestas que se han dado al mismo cobran especial relevancia para cualquier análisis relacionado con las cuestiones en torno a la IA. Esto se debe a que la supuesta eventual emergencia o no de un "pensamiento", "inteligencia" o "consciencia" autónoma en las máquinas o sistemas operativos depende en gran medida de las concepciones que se tengan sobre la naturaleza de la inteligencia o la consciencia humanas.

En lo respectivo a la cuestión de la IA, objeto de análisis de este artículo debido a su vinculación con la obra de Labatut, esta antigua problemática mente-cuerpo, transformada en mente-cerebro, supone una reflexión más moderna que surge con los escritos e investigaciones de los informáticos, tecnólogos y

[4] Respecto al problema mente-cuerpo, para los empirismos y materialismos de cualquier tipo resulta evidente que ya no podemos hablar de la mente como algo ligado al "espíritu", como lo habían hecho las antiguas tradiciones que, desde el orfismo, el pitagorismo y el cristianismo, concebían este concepto como una entidad separada del mundo físico y perteneciente, por tanto, a un mundo "espiritual". La filosofía contemporánea de la mente descarta de antemano cualquier explicación de este tipo.

pensadores como Alan Turing (1912-1954), John McCarthy (1927-2011), Marvin Minsky (1927-2016), Allen Newell (1927-1992) y Herbert Simon (1916-2001), entre otros. La apuesta teórica de estos pensadores por la emergencia de sistemas operativos de tal potencia que podrían algún día rivalizar con las capacidades humanas de pensamiento autónomo y creativo deviene relevante para nuestro análisis, tal como se explicará a continuación.

Alan Turing, informático y matemático británico, se destaca como paradigmático, considerado uno de los fundadores de la informática moderna. En su célebre artículo *Computing Machinery and Intelligence,* publicado en la revista *Mind,* plantea la célebre cuestión: ¿Pueden pensar las *máquinas?*" (Turing, 1950). Para responder a esta interrogante, Turing propone en su artículo el ya conocido juego de la imitación, en el cual "Un juez humano mantiene una conversación con un humano y una máquina. Si el juez no puede distinguir cuál es cuál, entonces la máquina pasa la prueba" (Crepeau, 2022).

Esta cuestión planteada por Turing acerca del pensamiento de las máquinas detonará todo el pensamiento posterior y marcará el derrotero a seguir por los filósofos de la mente posteriores, especialmente los funcionalistas computacionales, de los que se hablará más adelante. La razón de esto estriba en que la máquina de Turing,[5] planteada por él mismo como "una computadora

[5] De Mol (2021) define a las máquinas de Turing como: "dispositivos computacionales abstractos simples destinados a ayudar a investigar el alcance y las limitaciones de lo que se puede calcular. Las "máquinas automáticas" de Turing, como las denominó en 1936, fueron concebidas específicamente para el cálculo de números reales. Hoy en día, se las considera uno de los modelos fundamentales de la computabilidad y la ciencia informática teórica". No hay nada más acerca de la máquina de Turing que las series infinitas de estados, *inputs, outputs* y la *tabla de la máquina* que relaciona causalmente los elementos anteriores."

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

universal, capaz de hacer lo que cualquier computadora posible puede hacer, todo cálculo posible" (Rodríguez, 2006, p. 55), no está condicionada para su operación al sustrato material de la máquina, sino sólo a la operatoria lógica de la misma. Tal y como lo explica Rodríguez (2006):

> La descripción lógica de una máquina de Turing, o sea la especificación de las instrucciones que integran su "tabla", no da entrada bajo ningún concepto a la precisión de la naturaleza física de los estados computacionales cuya sucesión controla el programa (…) Una máquina de Turing es una máquina abstracta, que puede realizarse en una variedad de modos prácticamente infinita: se abre así la posibilidad a una funcionalización de la mente humana. (55).

Siguiendo esta línea de pensamiento inaugurada por Turing, John McCarthy, programador informático estadounidense a quien se adjudica la creación del término "Inteligencia Artificial" en la Conferencia de Darmouth de 1956, y, por tanto, considerado uno de sus padres fundadores, afirmaba en su artículo *Ascribing mental qualities to machines*: "Atribuir ciertas creencias, conocimientos, libre albedrío, intenciones, conciencia, habilidades o deseos a una máquina o programa de computadora es legítimo cuando dicha adscripción expresa la misma información sobre la máquina que expresa sobre una persona". (McCarthy, 1979: 1).

Esta afirmación de McCarthy es clave, pues ilustra a la perfección la pretensión de estos padres fundadores de alcanzar una IA fuerte, la cual fue descrita por Searle en su artículo *Minds, brains and programs,* como aquella en la que "el computador no es una mera herramienta en el estudio de la mente; más bien, un computador programado apropiadamente es realmente una mente, en el sentido que se puede

decir que los computadores con los programas apropiados pueden literalmente comprender y tener otros estados cognitivos" (Searle, 1980:1).[6] En este sentido, McCarthy es consciente que estas cuestiones se enmarcan en terrenos filosóficos, especialmente en relación con la comprensión de la noción de inteligencia: "Muchos de los problemas filosóficos de la mente toman una forma concreta cuando uno toma en serio la idea de hacer que las máquinas se comporten de manera inteligente […] cuestiones que hasta ahora sólo se habían considerado en relación con las personas." (McCarthy, 1979: 2).

Por su parte, Marvin Minsky, otro de los considerados padres de la IA, en su artículo *Step Toward Artificial Intelligence,* propone un análisis sobre las posibles capacidades que estos sistemas operativos podrían llegar a alcanzar: "Estoy seguro de que tarde o temprano podremos reunir programas de gran capacidad de resolución de problemas a partir de combinaciones complejas de dispositivos heurísticos, múltiples optimizadores, trucos de reconocimiento de patrones, álgebras de planificación, procedimientos de administración recursivos, y similares". (Minsky, 1961: 27). A su vez, continúa afirmando que, aunque sea difícil encontrar en estos programas de principio el "asiento de la inteligencia", no es razón para descartar otras maneras más complejas de entender la inteligencia que, eventualmente, puedan corresponderse con los ordenadores.[7]

[6] La contraparte de la IA fuerte, es decir, la IA débil, será definida como aquella en la cual, "el valor fundamental del computador en el estudio de la mente radica en que nos brinda una herramienta muy poderosa. Por ejemplo, nos permite formular y poner a prueba hipótesis de manera más rigurosa y precisa que antes" (Searle, 1980: 1). En consecuencia, la IA fuerte será postulada como equivalente a una IA general que puede abarcarlo todo y la IA débil como específica de algunas tareas.

[7] En este sentido es fundamental la apreciación de Minsky: "Pero no

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

Finalmente, Allen Newell y Herbert Simon resultan claves en la comprensión de las disputas en torno a la capacidad de alcanzar máquinas pensantes. En su artículo *Computer science as empirical inquiry: Symbols and search* de 1976, postularán su hipótesis del sistema de símbolos físico, la cual sostiene, siguiendo a Martínez-Freire (2001), que "un sistema de símbolos físico tiene los medios necesarios y suficientes para la acción inteligente general" (90). De manera que la consecuencia al respecto de los ordenadores, que también operan con símbolos, puede vislumbrarse lógicamente.

Por ser de capital importancia, conviene aclarar a qué se refieren Newell y Simon siguiendo a López de Mántaras (2018):

> Un sistema de símbolos físicos consiste en un conjunto de entidades llamadas símbolos que, mediante relaciones, pueden ser combinados para formar estructuras mayores y que pueden ser transformados aplicando un conjunto de procedimientos. Estos procedimientos pueden crear nuevos símbolos, crear y modificar relaciones entre estos […] Estos símbolos son físicos en tanto que tienen un sustrato físico-electrónico (en el caso de los ordenadores) o físico-biológico (en el caso de los seres humanos). (45).

Por lo tanto, de acuerdo con esta hipótesis, ya sea que en el caso de los ordenadores, los símbolos se realicen mediante circuitos electrónicos digitales o, en el caso de los humanos, mediante redes de neuronas, lo importante no es la naturaleza de este sustrato (chips o neuronas), sino el procesamiento de los mismos, o como lo llamará

---

debemos permitir que nuestra incapacidad para discernir un lugar de inteligencia nos lleve a concluir que, por tanto, las computadoras programadas no pueden pensar. Porque puede ocurrir lo mismo con el hombre, como con la máquina, que, cuando finalmente comprendamos la estructura y el programa, el sentimiento de misterio (y de autoaprobación) se debilitará." (Minsky, 1961: 7).

más adelante el primer Putnam[8], la función a manera de algoritmo del procesamiento de la información.[9]

Así pues, con estos presupuestos tanto de Turing como de los demás padres de la IA, podemos comprender mejor la tesis del funcionalismo computacional, que emerge como clave en las disputas de la filosofía de la mente contemporánea. De la respuesta que se le dé a esta tesis, depende la cuestión misma de la eventual emergencia de inteligencia en las máquinas.

Desarrollo posterior de la cuestión de Turing y los padres de la IA: del funcionalismo al emergentismo

Los planteamientos de Turing y los padres de la IA, especialmente los de Newell y Simon, acerca de la inteligencia o pensamiento de las máquinas detonaron el desarrollo posterior de los estudios no sólo en la informática y ciencias computacionales, sino también en la filosofía de la mente. Uno de los planteamientos más importantes es el del funcionalismo computacional, que se presenta a continuación y constituye un elemento clave del análisis propuesto.

De acuerdo con Rescorla (2020): "la teoría computacional de la mente sostiene que la mente es literalmente un sistema

[8] Decimos "Primer Putnam" porque las posiciones filosóficas de este autor varían a lo largo del tiempo.

[9] Como bien lo explica Martínez-Freire (2001): "Por otra parte, resulta claro que la hipótesis de símbolos físicos se aplica por igual a la psicología cognitiva y a la inteligencia artificial. En efecto, un sistema de símbolos es una descripción abstracta tanto de un sujeto humano o animal como de un computador. En todos estos casos cabe hablar de unidad de entrada (receptores), unidad de control con sus operadores, memoria y unidad de salida (motores). Y también en todos estos casos cabe hablar por igual de sujetos procesadores de información". (92).

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

informático" en el que las diferencias de sustratos físicos no representan un obstáculo desde una perspectiva que prima la "función" y no la "materia" del pensamiento:

> Por supuesto, los sistemas informáticos artificiales están hechos de chips de silicio, mientras que el cuerpo de carne y hueso. Pero la teoría sostiene que esta diferencia disfraza una similitud más fundamental, que podemos capturar a través de un modelo computacional de estilo Turing. Al ofrecer un modelo de este tipo, prescindimos de detalles físicos. (5).

Esta teoría del funcionalismo computacional tendrá uno de sus primeros representantes en el mundo filosófico con Hilary Putnam (1926-2016), quien en su artículo *Minds and machines,* postula en síntesis que "los estados internos de conciencia son estados abstractos, que se definirían por funciones análogas a las *tablas de máquina* o programas de una máquina universal de Turing" (Putnam, 1960: 384). Así pues, como señala Ruiz (2020), "el funcionalismo queda manifiesto en que los procesos mentales devienen de naturaleza computacional ya que constituyen funciones mediadoras entre las entradas sensoriales (inputs) y las salidas motoras (outputs)". Todo esto, con independencia de la naturaleza de los sustratos materiales, ya que se prima únicamente la función.

La terminología empleada por Putnam referente al "hardware" cobra especial relevancia y utilidad para comprender mejor la teoría del funcionalismo computacional, la cual se entiende bien mediante el binomio hardware-software. Como explica García (2001): "La mente es como un programa de ordenador y un programa puede ser ejecutado en cualquier hardware lo suficientemente potente. Se puede estudiar la mente (el programa o software) con independencia del hardware, en este caso el cerebro." (280).

Como puede observarse, este primer Putnam sigue el mismo hilo argumentativo planteado por Turing. Lo fundamental es la función del cerebro, es decir, el software, de manera que las máquinas podrían llegar a ser capaces de realizar actividades que los cerebros o softwares realizan si fueran suficientemente avanzadas. Así pues, tanto para Turing como para el primer Putnam y el funcionalismo computacional, deberíamos enfocarnos más en las entradas y salidas de datos, que, en la sangre y los nervios, o el cableado y los transistores internos; es decir, en el software y no en el hardware.

Putnam continuó sosteniendo su teoría del funcionalismo computacional durante varios años. Todavía en su texto *Reason, Truth and Histoy,* lo vemos sostener su célebre argumento antichovinista, según el cual, una vez supuesta que la única diferencia reside en la organización funcional, no podemos descartar la inteligencia de las máquinas: "Ahora bien, como no seas un "chovinista del carbono y el hidrógeno" que piense que el carbono y el hidrógeno son intrínsecamente más conscientes, ¿por qué no podrías decir que el robot es una persona cuyo cerebro ocurre que tiene más metal y menos hidrógeno y carbono?" (Putnam, 1981: 96).

Queda así conceptualizada por Putnam la "realizabilidad múltiple", que postula que "una única clase mental (propiedad, estado, evento) puede ser realizada por muchas clases físicas distintas" (Bickle, 2020).[10] Con esto queda consumado teóricamente el anhelo

[10] En este sentido, como afirma Rodríguez (2006): "Si la consciencia suponemos que no queda atrapada en la organización funcional, idéntica en el robot y en mí, entonces queda automáticamente remitida a la constitución física concreta. Pero como la realización física es abierta, múltiple, la consciencia dejaría de tener ningún sentido especificable, se tornaría completamente absurda. Luego de algún modo debe estar recogida en la organización

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

de los padres fundadores de la IA, ya que en este funcionalismo computacional nos encontramos ante una especie de reduccionismo, que al prescindir del "misterio de la consciencia" se decanta por una comprensión acerca de la plausibilidad de "dispositivos mecánicos que reciben *inputs* (argumento de la función), los elaboran aplicándoles reglas algorítmicas, los "computan", y emiten *outputs* verbales y conductuales (valor de la función). (Rodríguez, 2006. 57).

Antes de explicar las razones por las cuales Putnam abandonó esta ortodoxia funcionalista que se ha descrita anteriormente, es preciso apuntar que el funcionalismo computacional, según varios autores, puede ser visto como una especie de reduccionismo fisicalista al tratar de reducir lo psicológico a lo funcional. En otras palabras, las funciones que describen los estados mentales ignoran la complejidad de la interacción de la mente con la realidad de la historia, el lenguaje y la cultura. Siendo precisamente por esta razón que Putnam no mantuvo su postura original durante mucho tiempo.

Para finales de la década de los ochenta, en paralelo a sus desarrollos en filosofía del lenguaje, experimentó una retractación de su postura funcionalista original cuando llegó a la conclusión de que la realizabilidad múltiple pasaba por alto un elemento clave para comprender las relaciones de la mente con la realidad, el de la *referencia* a un marco de interpretación más complejo que no puede reducirse a meras funciones abstractas separadas del contexto. Como explica Ruiz (2020), "la clave está en que el concepto de *referencia* trasciende los sistemas físicos concretos y solo puede ser entendido en términos culturales o relacionales más amplios, lo que invalidaría el diseño de estructuras computacionales individuales" (p. 11). En este sentido, explica Rodríguez (2006):

---

funcional" (57).

> Jamás podremos llegar a poseer el Algoritmo Maestro para la interpretación porque es de todo punto imposible llevar a cabo la "valoración" de todos los posibles modos de conceptualización de los seres humanos en todos sus lenguajes y todas sus culturas, y en todas las formas de fijación de creencias. Por no decir nada, evidentemente, de las posibilidades inimaginables de los lenguajes y culturas no humanos. (63).

Así pues, este abandono del funcionalismo, se sustentará sobre la premisa de la insuficiencia del mismo debido a la reducción fisicalista que lleva aparejada, sin que esto implique, a juicio de Rodríguez (2006) "ninguna metafísica de interés más allá de los conceptos de *superveniencia* y *emergentismo*" (60). Precisamente, serán estos últimos conceptos los que las nuevas corrientes de filosofía de la mente adoptarán, especialmente en las contribuciones de John Searle (1932-).

Finalmente, esta crítica de Putnam al programa funcionalista tendrá amplias repercusiones en torno al tema de la IA, ya que, al descartar la posibilidad de comprender la mente meramente como un programador informático, debido a las razones mencionadas anteriormente, también se descarta la empresa de desarrollar una IA fuerte que pudiera igualar las características humanas. [11] En este sentido, si alguien sostuviera que un eventual descubrimiento de

[11] Como él mismo lo explica en *The Threefold Cord. Mind, Body and World*: Los materialistas tienen razón al insistir en nuestra naturaleza incorporada (embodied), tienen razón al insistir en que la conexión de la mente y el cuerpo es demasiado íntima como para que tenga algún sentido hablar de "espíritus desencarnados" (…) Pero se equivocan cuando su cientificismo los lleva a afirmar que sólo podemos pensar nuestra mente como algo que actúa en y a través de nuestros cuerpos con la condición de reducir los términos de la psicología natural a los términos de la física, la química, la física, la neurología, la ciencia computacional (Putnam, 1999: 149).

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

todas las funciones de la mente (lo cual Putnam considera imposible por la complejidad de la referencialidad) y, por consiguiente, la replicación en una IA fuerte mediante una reducción fisicalista, se encontraría con esta respuesta del filósofo: "Decir que la ciencia tal vez algún día encuentre el modo de reducir la conciencia (o la referencia, o lo que sea) a la física, es aquí y ahora, lo mismo que decir que tal vez la ciencia haga algún día no sabemos qué, no sabemos cómo" (Putnam, 1999: 173).

De la misma manera que el último Putnam, el profesor John Searle, también se opone a la reducción funcionalista de la inteligencia y su pretensión de equiparar la conciencia con un sistema computacional, expresada como la creencia de que "si logramos identificar los *inputs* y los *outputs* explícitos de determinado estado funcional sería posible reproducirlo en una máquina" (Santamaría & Sánchez, 2017: 450). Searle desarrolla esta postura a través de su teoría del naturalismo biológico postulada en su obra *El Redescubrimiento de la Mente,* según la cual:

> Los fenómenos mentales están causados por eventos neuropsicológicos del cerebro y son a su vez rasgos del cerebro (…) Los eventos y procesos mentales son parte de nuestra historia biológica en la misma medida en que lo son la digestión, la mitosis, la meiosis o la secreción de enzimas. (Searle, 1996: 15).

En este sentido planteado por Searle, podemos hablar de un "emergentismo" donde "el cerebro causa los fenómenos mentales conscientes a manera de propiedad emergente de las funciones superiores del cerebro (en especial lo correspondiente a la neo-corteza" (Santamaría & Sánchez, 2017, p. 451). Searle postula además que la conciencia es propiedad de la emergencia de lo mental, siendo una causa superior del holismo cerebral:

> La existencia de la conciencia puede ser explicada por las interacciones causales entre el cerebro a micro-nivel, pero la conciencia misma no puede ser deducida o calculada a partir de la mera estructura física de las neuronas. (Searle, 1996: 122).

Por otra parte, no podemos pasar por alto que, en su crítica al programa funcionalista, Searle retoma también la cuestión del significado y el lenguaje, al precisar que, los ordenadores serían como alguien encerrado en la Habitación China[12], el famoso experimento mental que pretende probar la invalidez del test de Turing al demostrar teóricamente que los ordenadores carecen de inteligencia real ya que su capacidad se limita a procesar o manipular cadenas de símbolos siguiendo directrices incorporadas en su software, lo que les permite comprender la sintaxis, pero no la semántica.

A pesar de las críticas al funcionalismo computacional que hemos descrito anteriormente, es importante no perder de vista que, en el contexto de las discusiones históricas sobre filosofía de la mente y los intentos de responder al problema cartesiano, el funcionalismo, según Santamaría y Sánchez (2017), "era la teoría que terminaba de una vez por todas con la misteriosa mente al hacerla equivalente a procesos algorítmicos de cómputo" (p. 460) y a su vez por supuesto, de posibilitar toda la discusión acerca de simular o hablar de pensamiento inteligente en las máquinas que se explicaría solamente por los procesos algorítmicos mediadores entre *inputs* y *outputs*.

[12] Sobre las implicaciones filosóficas que tiene el experimento de la habitación china puede consultarse el artículo de la Enciclopedia de Filosofía de Stanford: Cole, David, "The Chinese Room Argument", The Stanford Encyclopedia of Philosophy (Summer 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/sum2023/entries/chinese-room/

DOI: https://doi.org/10.29105/revistahumanitas4.8-111

De Literatura y Filosofía en *Maniac* de Labatut

Este artículo sostiene la tesis de que Labatut esboza claramente, desde la literatura, un marco epistémico claro acerca de los alcances y límites de la IA. Este marco epistémico, que ha permanecido invariable a pesar de las revoluciones tecnológicas de las últimas décadas, sugiere la imposibilidad, al menos por el momento, de que las máquinas alcancen una verdadera inteligencia según el paradigma humano. No obstante, Labatut, deja abierta la posibilidad de nuevas interpretaciones de la inteligencia, susceptibles de ser abordadas de forma radicalmente distinta a la propuesta por la antropología convencional.

Como se ha señalado al inicio de este artículo, el análisis literario de la obra *Maniac* se enmarca en la última parte de la obra titulada *Lee o Los delirios de la inteligencia artificial.* Esta sección comienza con un prólogo en el que se narra la historia del famoso juego oriental Go, equivalente al ajedrez occidental, y donde Labatut se hace eco del mito de "el legendario emperador Yao", inventor del juego (Labatut, 2023, p. 299). Posteriormente, se detallan aspectos biográficos de Lee Sedol, apodado *La piedra fuerte,* quien es reconocido como "maestro del Go 9.º dan, el jugador más creativo de su generación, y el único ser humano que ha vencido a un sistema avanzado de inteligencia artificial durante un torneo profesional…" (303) Es importante destacar que en este capítulo se destacan los méritos y reconocimientos que Lee Sedol ha alcanzado como jugador más destacado de Go en la época contemporánea.

En cuanto a estos méritos, encontramos algunos muy significativos, como el hecho de que desde los trece años practicaba durante doce horas al día en la Academia Internacional de Go de

Corea, el ganar el 12º Campeonato Nacional de Go Infantil, siendo el ganador más joven de dicho torneo a los ocho años, competición en la que "demostró el estilo salvaje, violento e impredecible que lo volvería famoso" (304). Además, convertirse en el jugador más joven en alcanzar el 9º dan, el nivel más alto posible (307), entre otros logros, destacando siempre su habilidad eminentemente racional de concebir las innumerables combinatorias probabilísticas del juego del Go:

> La principal fortaleza de Lee Sedol (…) [eran] sus jugadas únicas, fruto de la habilidad que Lee Sedol desarrolló tras pasar tanto tiempo como pudo practicando su don de leer el tablero completamente vacío, mirando hacia el futuro para imaginar los múltiples senderos que se bifurcan a partir de los movimientos más humildes y sencillos. (306).

Es importante señalar que Labatut, al retratar los rasgos biográficos de Lee Sedol, hace hincapié no sólo en los rasgos profesionales que lo destacan como jugador del Go, sino también en aspectos que reflejan su condición humana, con todos sus anhelos, esperanzas e ilusiones. Labatut describe así las aspiraciones de Sedol: "Quiero pasar a la historia como una leyenda viva. Quiero ser la primera persona que la gente asocie con el Go. Quiero que mis partidas perduren, se estudien y admiren como obras de arte" (307).

Por otra parte, son notables las referencias al juego practicado por Sedol como una creación "única" de la raza humana, con toda la carga no sólo lógica y racional, sino también estética:

> En juegos como el ajedrez y el shogi se empieza con todas las piezas sobre el tablero, pero en el Go se empieza con el vacío, se empieza con la nada, y luego los jugadores van añadiendo blanco y negro sobre el tablero, y crean una obra de arte. La infinita complejidad del Go, toda su belleza, brota de la nada. (308).

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

En el mismo sentido, la perfección y grandeza del Go como creación humana se compara con el orden cosmológico, al que en la filosofía clásica se le atribuyen características teleológicas de orden y perfección:

> Lee respondió que el Go era, ante todo, una forma de entender el mundo: su infinita complejidad era el mejor espejo de cómo funcionaba nuestra mente, mientras que sus acertijos y laberintos, aparentemente insondables, lo convertían en la única creación humana capaz de rivalizar con el orden, la belleza y el caos de nuestro universo. (310).

Inclusive estas comparativas entre dimensiones estéticas y cosmológicas culminan en la narrativa de Sedol con la afirmación y asociación del juego del Go con la "mente de Dios":

> Si alguien fuese capaz de comprender el Go totalmente- y con eso no me refiero solo a las posiciones de las piedras y la forma en que se relacionan entre sí, sino también a los patrones ocultos, prácticamente imperceptibles, que surgen por debajo de esas formaciones cambiantes-, creo que sería lo mismo que entrar en la mente de Dios. (310).

Como queda de manifiesto en la presentación del Go realizada en la obra, al destacar la complejidad y belleza del juego, Labatut lo posiciona como arquetipo de creación humana que engloba todo su potencial de racionalidad y creatividad. Al mismo tiempo, lleva al lector a una comparativa ineludible con las aspiraciones de los sistemas operativos de IA, que, aún con toda su complejidad, devienen carentes de estas dimensiones.

Posterior a esta descripción del Go como la creación "suprema" del ingenio humano, Labatut procede a describir a Alpha

Go, el sistema operativo que será el contrincante de Lee Sedol en la legendaria batalla narrada en la obra. Es importante señalar desde ahora que la creación de AlphaGo tiene como protagonista a Demis Hassabis, "un niño prodigio del norte de Londres" (312), quien, motivado por su afición al ajedrez y al darse cuenta de su propia inteligencia y capacidad, emprende la creación de dicho sistema operativo.

Resulta de particular interés el hecho de que, para Labatut, todas las reflexiones de Hassabis en torno a la creación de esta IA están vinculadas a un intento de trascender los límites de la inteligencia humana. Vemos cómo, después de preguntarse por su propia capacidad e inteligencia, así como por su habilidad de cálculo y anticipación, características esencialmente humanas (317), Hassabis plantea la siguiente reflexión sobre el futuro de la tecnología:

> La ciencia del siglo XXI- joya de nuestra corona- progresaba tan rápido que nos empujaba hacia un precipicio, creando un mundo nuevo para el cual no estábamos preparados (…) Pronto alcanzaríamos un punto de quiebre. Nuestros cerebros humanos nos habían llevado tan lejos como podían. Necesitábamos algo radicalmente distinto. Una mente capaz de ver más allá de nuestras limitaciones y penetrar las sombras que nuestros propios ojos proyectan sobre el mundo. (318).

Esta reflexión será la base de su motivación para la creación de la IA, que, apenas en semilla en ese momento, se convertirá en la Deep Mind, el sistema operativo contra el que se enfrentará el protagonista del capítulo, Lee Sedol. Hassabis expresa su intención de la siguiente manera: "Ya no deseaba ser el campeón mundial del ajedrez. Quería algo distinto y mucho más importante: crear una nueva mente, más fuerte, más rápida y más extraña que todo lo

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

conocido. IAG: inteligencia artificial general. El verdadero hijo del hombre." (318).

Es importante señalar que, según la narrativa de Labatut, la preparación de Hassabis para crear Deep Mind se basa en los estudios que el joven realizó sobre varias disciplinas ligadas a la neurociencia cognitiva, especialmente en el estudio de las facultades humanas de la memoria y la imaginación, con la intención de replicar estos mecanismos en un sistema operativo. Encontramos así una vinculación expresa con las teorías del funcionalismo computacional que se han abordado en nuestro marco teórico:

> La investigación de Hassabis demostró que las facultades de la memoria y la imaginación comparten un mecanismo común arraigado en el hipocampo. "Mi trabajo investigaba la imaginación como proceso. Quería saber cómo nosotros, los seres humanos, visualizamos el futuro, y luego ver qué es lo que los computadores venideros podrán conjurar", señaló después de publicar sus resultados. (321).

Al narrar el legendario episodio entre Garry Kasparov y la Deep Blue de 1997, como precedente del episodio de Lee Sedol, Labatut presenta una narrativa que respalda la tesis de la capacidad limitada de los sistemas operativos para desarrollar autonomía de pensamiento, permaneciendo siempre ligados a la operatoria y al cálculo de probabilidades, por más poderosos que puedan parecer. En este sentido, en relación con los juegos de entrenamiento de Deep Blue, se menciona: "Después de todo, IBM había podido estudiar miles de partidas de Kasparov utilizando su poder informático casi ilimitado para analizar sus estrategias, aperturas y movimientos preferidos" (326). Así, tras narrar el episodio de Kasaparov, se

concluye: "Estos programas no juegan al ajedrez como lo hacemos nosotros. No dependen de la creatividad ni de la imaginación, sino que eligen los mejores movimientos utilizando fuerza de cálculo pura y dura..." (327).

Sobre el cálculo y la probabilística como límite de la IA en *Maniac*

En *Lee o los delirios de la inteligencia artificial,* encontramos la aparición de las teorías de filosofía de la mente que, al vincularse con el funcionalismo computacional y sus presupuestos fisicalistas, posibilitan una hermenéutica en la que el materialismo y la IA aparecen estrechamente ligados si se supone un mundo lógico-matemático que daría razón de los procesos que hasta ahora se han considerado exclusivamente humanos:

> Mi padre quería conocer la lógica interna del cerebro. El "lenguaje" que utiliza para funcionar. Quería saber si ese idioma se parecía a la lógica matemática, su método preferido de pensamiento. "Cuando hablamos de matemáticas puede que estemos hablando de un lenguaje secundario construido sobre el lenguaje primordial que utiliza el sistema nervioso. (266).

La comprensión tanto de mente humana como de la IA en un marco hermenéutico centrado en la matemática y el cálculo de probabilidades como único horizonte explicativo aparece en varios pasajes del capítulo. Por ejemplo, cuando Demis Hassabis, creador de AlphaGo, se pregunta: "¿Cuál era el origen de su extraordinaria inteligencia? ¿Por qué podía aprender de forma tan rápida? ¿Por qué tenía esa relación tan cercana con los números?" (317).

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

Encontramos, sin embargo, una oposición al reduccionismo fisicalista de la mente humana y, por consecuencia, de sus creaciones, como el juego del Go, cuando se narra que Sedol en la conferencia de prensa inaugural declara:

> Hay una belleza particular en el Go, y no creo que las máquinas puedan entender esa belleza. Creo que la intuición humana es demasiado avanzada para que la inteligencia artificial la haya alcanzado aún, así que no estoy pensando si voy a ganar o no. Lo que me preocupa es si voy a ganar cinco a cero o cuatro a uno. (336).

Se establece así un marco hermenéutico acerca de los alcances de los sistemas operativos de IA, ligados inexorablemente al cálculo mecánico de probabilidades, por más poderoso que dicho cálculo pueda llegar a ser. Este marco, de la misma manera que en el caso de Deep Blue permanece siempre en la narrativa sobre el proceder de *AlphaGo:*

> Cada vez que su oponente coloca una pieza en uno de los casilleros del tablero, el programa construye un árbol de búsqueda: sus ramas son los posibles futuros que surgen de esa configuración particular de piezas; el árbol crece hasta llegar al final de la partida, y el programa simplemente elige una de las ramas como el resultado que considera más ventajoso (Labatut, 2023: 327).

En este sentido, podemos constatar cómo la obra de Labatut posibilita en todo momento un marco epistémico de reducción a probabilística y mecánica con respecto a la inteligencia de los programas operativos, ya sean de la índole del ajedrez o del Go lo cual parece reflejar el paradigma antropológico clásico que no apuesta por una verdadera inteligencia autónoma en las máquinas:

> Mientras que un ser humano utiliza su memoria, experiencia, intuición, razonamiento abstracto y capacidad de detectar patrones para interiorizar el tablero en su mente y adquirir una comprensión del juego, los programas de ajedrez, en cambio, no necesitan "entender" sino que usan su potencia para calcular, y luego optan por una jugada siguiendo un conjunto de reglas establecidas por sus programadores. (327).

La probabilística como límite de la IA se mantiene invariable, incluso en el supuesto caso del movimiento 37, que aunque a primera vista podría interpretarse como un destello de "originalidad" y, por ende, de "superioridad" de AlphaGo sobre el ser humano[13] sigue siendo entendido dentro de un marco de programación basado en las experiencias humanas con las que el programador humano la había "entrenado".

Labatut es siempre consecuente al dejar claro que incluso en los supuestos golpes "magistrales" del sistema operativo, que parecieran emular lo humano, lo que subyace sigue siendo la probabilística, la programación y el cálculo, como se demuestra incluso en la creación de IA con redes neuronales: "Hassabis y su equipo creían que la única forma de derrotar a un profesional del más alto nivel era emular la manera- algo misteriosa y profundamente intuitiva- en que los humanos jugaban al Go" (353). Por ello, "Para lograrlo, crearon una base de datos con cerca de ciento cincuenta mil partidas de los mejores jugadores y la introdujeron en una red neuronal artificial, un complejo modelo matemático que imita a las neuronas de nuestro cerebro…" (353).

[13] Al respecto de este famoso movimiento se narra: "Cuando los historiadores del futuro observen nuestra época y traten de encontrar el primer destello de la inteligencia artificial, es muy posible que lo hallen en una jugada de la segunda partida entre Lee Sedol y AlphaGo, que tuvo lugar el 10 de marzo de 2016: el movimiento 37". (340).

DOI: https://doi.org/10.29105/revistahumanitas4.8-111

Respecto a esta forma de operación de AlphaGo entrenada con redes neuronales, Labatut describe un aspecto interesante que bien posibilita una reflexión sobre el acercamiento de los sistemas de IA al desarrollo de una inteligencia particular. Dicho aspecto es el referente al "aprendizaje profundo",[14] conocido en informática como una de las tendencias en los desarrollos de la IA. Al respecto, Labatut comenta cómo se entrena AlphaGo:

> A través del ensayo y el error se volvió cada vez más fuerte (…) A lo largo de esas innumerables partidas (de entrenamiento), su modelo matemático experimentó miles de millones de ajustes, mejorando por razones que ningún ser humano podría llegar a entender, ya que el funcionamiento de una red neuronal artificial es algo casi completamente opaco para nosotros, porque somos incapaces de seguir o de comprender los efectos que surgen de la infinidad de pequeñas modificaciones que el algoritmo realiza a sus parámetros a medida que se acerca lentamente al objetivo deseado (354- 355).

En este mismo sentido, encontramos un pasaje sobre una aparente superación de las capacidades cognitivas del ser humano. Este pasaje mencionado hace referencia a la "red de valor" desarrollada por el sistema operativo, según la cual: "analizaba cualquier configuración de piedras y miraba hacia el futuro, proyectando el juego hasta el final, para estimar si estaba ganando o no, y por cuánto" (355). Al respecto Labatut menciona:

[14] Tal y como se explica en otro pasaje: Demis Hassabis lo había explicado antes del campeonato: "Aunque hemos programado esta máquina, no tenemos idea de que movimientos va a inventar. Son fenómenos emergentes, algo que surge de su aprendizaje. Nosotros solo creamos los conjuntos de datos y los algoritmos de entrenamiento. Pero las jugadas que AphaGo imagina no están en nuestras manos y son mucho mejores de lo que podríamos hacer nosotros …" (345).

> Esta capacidad era algo que ningún jugador humano poseía, por talentoso que fuera, ya que, gracias a esta segunda red, AlphaGo podía asignar un valor numérico a algo que los seres humanos solo pueden calcular mediante la nebulosa certeza que nos otorga nuestra intuición. (355).

No obstante, encontramos que esta superación, aunque resulta evidente en un sentido cuantitativo, a la manera, por ejemplo, de la capacidad de cálculo de los sistemas informáticos, sigue estando limitada sólo al cálculo, por muy avanzado que este sea:

> [Los movimientos de AlphaGo] estaban basados en el cálculo puro: cada una de esas piedras perezosas representaba una ganancia- minúscula, casi imperceptible- hacia el resultado, pero su verdadero valor no se apreciaría hasta el final de la partida, cuando todas funcionaran en conjunto. (358).

Encontramos nuevamente este límite del cálculo en otros pasajes, como cuando después de una jugada magistral de la inteligencia de Sedol, el sistema operativo responde en este mismo marco de cálculo y probabilística: "En la sala de control de Deep Mind, el programador principal, David Silver, notó que el algoritmo había proyectado hacia el futuro más de noventa y cinco movimientos posibles tras la espectacular jugada de Lee Sedol, desarrollando las interminables probabilidades que se ramificaban a partir de cada uno de ellos." (367). Lo mismo ocurre en la última partida: "AlphaGo podía realizar algo de lo que ningún ser humano era capaz: calcular, con una precisión absoluta e infalible cuánto territorio necesitaba para vencer" (377).

Sobre la cuestión de "El dedo de Dios"

La narrativa de Labatut en torno al desarrollo de las partidas de Lee Sedol contra el sistema operativo AlphaGo se torna fascinante,

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

pues pareciera representar una historia épica de combate entre el ser humano y la máquina. Bajo esta perspectiva, podemos ver desarrollarse el capítulo titulado *El dedo de Dios,* en el cual encontramos el cuarto juego en el que Sedol logra vencer al sistema operativo, lo que implica una victoria de la inteligencia humana sobre la máquina.

En este pasaje, presenciamos una magistral operación de la inteligencia de Sedol, descrita en el capítulo como "el dedo de Dios", "una jugada de los dioses" (365). En un momento en que, aunque el algoritmo "estimaba su probabilidad de ganar en más del 70 por ciento" y "nadie podía imaginar una forma de romper la granítica fortaleza que la computadora había levantado", la inteligencia de Sedol lo lleva a realizar una jugada magistral: "Como un rayo, la piedra 78 de Lee fulminó las fortificaciones de la máquina, penetrando el centro del tablero con un movimiento de cuña como nadie había visto antes" (365). Movimiento que desencadena una serie de respuestas sin sentido por parte de AlphaGo, su posterior renuncia y la victoria de Sedol en dicha partida.

Esta hazaña de Sedol llevó a que la gente "enloqueciera de emoción" y que reconociera que "era un movimiento impensable, una jugada que nadie salvo él habría considerado". Esta jugada supuso, por supuesto, un momento épico para la inteligencia humana:

> Fuera, en los pasillos del hotel y las calles de Seúl, hombres y mujeres que no se conocían de antes y que habían estado viendo la partida se abrazaban y besaban […] como si Lee Sedol no hubiese logrado esa victoria para sí mismo, sino para todos los miembros de nuestra especie. (370).

Encontramos también una interesante observación sobre la capacidad metacognitiva del ser humano, de la que evidentemente

carece la máquina, cuando después de que Sedol ganara la partida se sumerge en una serie de reflexiones sobre su propio pensamiento y actuación: "Pero el hombre que había vencido a la máquina no se movió de su asiento, sino que continuó analizando las piedras en el tablero, pensando en alternativas posibles, caminos y senderos que podría haber recorrido, al igual que lo había hecho en las tres partidas anteriores" (370)..

Del mismo modo, sobre esta victoria para la inteligencia humana que se deriva del pasaje del dedo de Dios, encontramos también una reflexión literaria de sumo interés que posibilita hablar de una superioridad perenne de la inteligencia humana: "Y esa victoria significó que aún podíamos defendernos, dar batalla. Con el paso del tiempo, será más y más difícil vencer a la inteligencia artificial. Pero ganar esa única partida… fue suficiente. Una vez fue suficiente" (371).

Es importante tener en cuenta que, aunque en varios pasajes del capítulo se dejan entrever algunas dudas sobre la capacidad exclusiva o no de ciertos rasgos de la naturaleza humana, como cuando "Hui llegó a estar convencido de que el algoritmo era capaz de manifestar una creatividad real, uno de los sellos distintivos de la inteligencia humana". (347), posterior al pasaje del dedo de Dios, encontramos esta característica de la creatividad referida explícitamente a Sedol y, por tanto, al ser humano:

> Incluso la gente de Deep Mind no lograban comprender cómo Lee había podido crear algo de la nada. ¿Cómo era posible que un hombre, sin importar cuán inteligente fuera, derrotase a una máquina capaz de calcular doscientos millones de posiciones en un segundo? Era una hazaña que pasaría a la historia, la mejor demostración del genio creativo de Lee Sedol, y algo que toda la humanidad podía festejar. (372).

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

Al cierre del capítulo, encontramos nuevamente esta misma narrativa sobre la genialidad y creatividad de la mente humana que, en consecuencia, supera a la de la máquina. Se describe cómo, después de que el equipo de DeepMind analizara nuevamente la partida, se busca que el sistema operativo evalúe la jugada del dedo de Dios mediante las redes de valor y de políticas con base a sus algoritmos profundos. La narración al respecto es fascinante:

> "¿AlphaGo habría jugado allí?", preguntó Silver mientras encendían el sistema y lo veían ejercer su vasto poder computacional para desentrañar las infinitas hebras de probabilidad con que se teje el futuro. "¿Cuál es la probabilidad que le otorga a esa jugada en particular?" "Cero punto cero cero cero uno", respondió uno de los investigadores más jóvenes del equipo. (374).

Para terminar con una descripción apoteósica de las capacidades cognitivas humanas: "la jugada de Sedol, había sido realmente divina, un roce de los dedos de Dios, algo que solo uno entre diez mil jugadores humanos habría podido imaginar", y concluye afirmando: "era algo que se alejaba demasiado de la experiencia humana, algo que superaba incluso la capacidad, aparentemente ilimitada, de la inteligencia artificial" (374).

Cuando Labatut narra las apreciaciones y reflexiones de Lee Sedol al respecto de su experiencia contra AlphaGo, encontramos un punto sumamente interesante que refleja la tesis defendida en este artículo y que, a lo largo del marco conceptual, encontramos como fundamental: la capacidad cognitiva humana y su intencionalidad como límite para la IA, especialmente sobre aquellos aspectos relacionados con la experiencia humana y la creatividad, que aún se le escapan:

> "No creo que AlphaGo sea necesariamente superior a mí", dijo. "Creo que aún hay mucho que los seres humanos pueden hacer

> contra la inteligencia artificial (…) Porque, ya seas un principiante o un profesional, el Go es un juego que se disfruta. El goce es la esencia del Go. Y AlphaGo es muy fuerte, pero no puede conocer esa esencia. (378).

Nuevas fronteras de la razón tecnológica

En el epílogo de la obra titulado *El Dios del Go,* Labatut narra la aparición de un sistema operativo renovado y con una potencialidad nunca antes vista en cuanto a la destreza para el juego del Go. Al respecto, se menciona "Bajo el apodo de "el maestro", empezó a acumular una victoria tras otra. Aparentemente imbatible, ganó cincuenta partidas de forma consecutiva contra los mejores jugadores del mundo" (385). Posteriormente, se describe una partida del nuevo sistema operativo contra un jugador chino llamado Ke Jie, quien, después de perder la partida, declara: "Para mí, el maestro es un dios del Go […] Él puede percibir el universo completo del Go, mientras que yo solo veo un pequeño espacio a mi alrededor" (386).

La aparición de este sistema operativo (Alpha Zero) representa aparentemente una forma revolucionaria de concebir la IA de aprendizaje automático, al grado de parecer un verdadero desafío en la demarcación sus límites. Al respecto, el mismo Ke Jie declara: "¿Cuánto más podría mejorar ese programa a través del autoaprendizaje? Es difícil imaginar cuáles son sus límites. Creo que el futuro pertenece a la inteligencia artificial" (386).

Posteriormente, encontramos que este nuevo sistema operativo responde al deseo de sus creadores- en especial de Hassabis, el creador de Deep Mind, de alcanzar la IA general, que, como se ha comentado anteriormente, equivale a la IA "fuerte":

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

> Hassabis había alcanzado un hito fundamental en su cruzada para dar vida a la inteligencia artificial general, pero la pregunta que se había hecho Ke Jie (¿Cuánto más podría evolucionar el sistema a través del autoaprendizaje?) le roía la conciencia (…) ¿Hasta dónde podría llegar el algoritmo realmente? (387).

La descripción de los avances de este nuevo sistema algorítmico es fascinante y sugestiva, ya que describe cómo dicho sistema, mediando únicamente un algoritmo de autoaprendizaje, "logró transformarse en la entidad más poderosa que el mundo haya conocido en Go, sino en el ajedrez y el shogi". Esto lo logra únicamente a través de su capacidad algorítmica: "Para todos estos juegos, esa nueva inteligencia artificial no consideró ninguna experiencia humana: simplemente le dieron las reglas y la dejaron jugar contra sí misma […] Su nombre es Alpha Zero" (388).

Podemos afirmar que esta última narrativa acerca de las capacidades imbatibles de AlphaZero evoca en el lector del texto no solo el imaginario de qué tan cerca estaríamos de la IA "general" y "fuerte", sino también un replanteamiento sobre la posibilidad de traspasar los límites que, hasta ahora, suponemos que tiene cualquier sistema operativo de IA.

A manera de conclusión: sobre la perennidad de lo humano

La obra *Maniac* de Labatut cobra toda su vigencia y actualidad al plantear cuestiones de primer orden en la reflexión filosófica contemporánea. Ya sea sobre la perennidad o provisionalidad de las características atribuibles al ser humano, los alcances y límites de la comprensión humana acerca de la inteligencia misma, o la temática de la hipotética singularidad tecnológica, temáticas todas

que se tornan estimulantes para las Humanidades, ya que plantean las interrogantes más profundas sobre la esencia del ser humano.

Después de la caída del funcionalismo computacional, las posturas contemporáneas en filosofía de la mente han favorecido el emergentismo como la explicación más plausible para abordar el problema de la consciencia. Este enfoque al establecer la filogénesis biológica de Searle como un factor indispensable para la emergencia de la consciencia, excluye de antemano a las máquinas o los sistemas operativos. En este sentido, los presupuestos filosóficos en la narrativa de Labatut parecen estar conscientes de ello, aunque sin cerrarse por completo, como lo refleja la narrativa final acerca de AlphaZero como una posible alteridad tecnológica.

En esta época de auge de la IA y su manifestación omnipresente en diversos ámbitos de la vida humana, se plantea que estamos más cerca que nunca de alcanzar la singularidad tecnológica. Sin embargo, las disputas acerca de la filosofía de la mente nos recuerdan que la consciencia humana sigue siendo, hasta el momento, algo no objetivable. La intencionalidad de la misma sigue revelándose como algo que trasciende al funcionalismo computacional, aún y con toda la fuerza de los sistemas operativos actuales de IA, tal como Lee Sedol "ganó la única partida que importaba" contra AlphaGo. Tal y como afirma López de Mántaras (2018): "Efectivamente el éxito de sistemas como AlphaGo, Watson y los avances en vehículos autónomos han sido posibles gracias a esta capacidad de analizar grandes cantidades de datos. No obstante, no hemos avanzado en nada hacia la consecución de IA general" (49).

En su artículo *La Inteligencia Artificial y la Realidad Restringida,* el filósofo español González Quirós hace referencia a lo que él llama "las estrecheces metafísicas de la tecnología" y plantea que "defender

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

la idea de que la mente pueda ser algo como un s*oftware* o una *máquina virtual,* olvidando que el cerebro, o más en general, el cuerpo que la *incorpora*, es una realidad biológica que experimenta de manera continuada diversas alteraciones, algo vedado a cualquier sistema que procese un programa es una inconsecuencia" (González, 2019: 129). Algo que parece estar siempre muy en claro en la narrativa de Labatut.

Sin embargo, no olvidemos que los planteamientos finales de Labatut en torno a la potencialidad de AlphaZero dejan abierta la cuestión sobre nuevas formas y maneras de entender la inteligencia, quizás desde otros marcos epistémicos, incluso el mismo González (2019) sugiere que estos nuevos sistemas podrían constituirse como eventuales alteridades tecnológicas. Citando a Steven Strogatz, matemático de Cornell, González plantea: "Tras los éxitos de AlphaZero, el algoritmo de aprendizaje profundo de DeepMind (...) estaríamos ante formas de inteligencia, y no de mero cálculo, que podrían superar la capacidad de nuestra conciencia, de nuestra inteligencia consciente" (134).

Es de suma importancia considerar que el aporte de la literatura deviene esencial para la Filosofía, ya que las intuiciones literarias de todos los grandes escritores plantean siempre formas sugestivas de analizar la realidad. Desde esta perspectiva, el proceder, las dudas, temores y reflexiones del personaje Lee Sedol en su encuentro con la IA plantean una dimensión antropológica que parece tener carácter de perennidad. Tal y como afirma González (2019):

> Puede que lleguemos a poseer un conjunto extraordinariamente atractivo y poderoso de nuevas tecnologías habilitantes, pero nada de eso nos dirá algo interesante acerca del sentido y valor de la vida, de la personal y la colectiva, ni existe la menor base para creer que podamos hacer conscientes a realidades que carecen de las extraordinarias cualidades de lo vivo. (144).

Sin embargo, al final, cabe la pregunta: ¿llegaremos algún día a establecer otro marco antropológico y epistémico que permita pensar racionalmente y con fundamentos claros la posibilidad de entender de otro modo la inteligencia y la consciencia? De momento, parece ser que los fundamentos son inamovibles, pero quizás algún día mediando el cambio de estos esquemas, podamos hablar de otro tipo de alteridades tecnológicas. Sin embargo, de momento, la filosofía de la mente contemporánea y la literatura del chileno, nos permiten y seguirán permitiendo hablar sobre la perennidad de lo humano.

Referencias

Bickle, J. (2020). "Multiple Realizability", *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/sum2020/entries/multiple-realizability/

Britannica, T. Editors of Encyclopaedia (2024). Go. *Encyclopedia Britannica*. https://www.britannica.com/topic/go-game

Crepeau, B. (2022). Alan Turing. *Salem Press Biographical Encyclopedia.*

De Mol, L. (2021). "Turing Machines", *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2021/entries/turing-machine/

García, E. (2001). *Mente y Cerebro.* Madrid: Síntesis.

González, J. (2019). La inteligencia artificial y la realidad restringida: las estrecheces metafísicas de la tecnología. *Naturaleza Y Libertad. Revista De Estudios Interdisciplinares*, (12). https://doi.org/10.24310/NATyLIB.2019.v0i12.6271

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111

Labatut, B. (2023). *Maniac*. Barcelona: Anagrama.

López de Mántaras, R. (2018). Hacia la inteligencia artificial. Progresos, retos y riesgos. *Mètode: Revista de difusión de la Investigación*. No. 99 (máquina y humanos ante el siglo10101), 2018 (Ejemplar dedicado a: Interconectados), págs. 44-51 https://dialnet.unirioja.es/servlet/articulo?codigo=6692706

Martínez-Freire, P. (2001). Base empírica y teoría funcionalista en las ciencias cognitivas, Ágora: Papeles de Filosofía, Vol. 20, N. 1 (2001), 87-104 http://hdl.handle.net/10347/1172

McCarthy, J. (1979). Ascribing mental qualities to machines. In Martin Ringle (ed.), *Philosophical Perspectives in Artificial Intelligence*. Humanities Press. http://jmc.stanford.edu/articles/ascribing.html

Minsky, M. (1961). Steps toward Artificial Intelligence, in *Proceedings of the IRE*, vol. 49, no. 1, pp. 8-30, Jan. 1961, https://doi.org/10.1109/JRPROC.1961.287775

Putnam, H. (1960). *Minds and machines*. En S. Hook (ed.) *Dimensions of Mind: A Symposium,* New York University Press. pp. 362-385. https://philpapers.org/archive/PUTMAM.pdf

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

Putnam, H. (1999). *The Threefold Cord. Mind, Body, and World.* New York, Columbia University Press.

Rescorla, M, (2020). The Computational Theory of Mind, *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2020/entries/computational-mind/

Rodríguez, M. (2006). Desmontando la máquina: las razones de Putnam contra el funcionalismo. *Logos. Anales Del*

Seminario de Metafísica [Universidad Complutense de Madrid, España] 39:53-76. https://www.researchgate.net/publication/277273378

Ruiz, C. (2020). *Conductismo, Funcionalismo y Eliminativismo.* [Tesis de posgrado] http://www.fisicafundamental.net/doc/mente.pdf

Santamaría-Velasco, F. & Sánchez-Ávila, J. (2017). Pensar la conciencia: mente, intencionalidad y lenguaje. *Escritos*, 25(55), 437-463. https://doi.org/10.18566/escr.v25n55.a05

Searle, J. (1996). *El redescubrimiento de la mente*. Barcelona: Crítica.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. Traducción de Lucía Castillo I., estudiante de Magíster en Estudios Cognitivos, Universidad de Chile. https://doi.org/10.1017/S0140525X00005756

Turing, A. (1950). Computing machinery and intelligence. *Mind*, Volume LIX, Issue 236, October 1950, Pages 433–460, https://doi.org/10.1093/mind/LIX.236.433

 DOI: https://doi.org/10.29105/revistahumanitas4.8-111