UNIVERSIDAD AUTÓNOMA DE ZACATECAS
"FRANCISCO GARCÍA SALINAS"

UNIDAD ACADÉMICA DE DOCENCIA SUPERIOR
Maestría en Investigaciones Humanísticas y Educativas
Orientación en Filosofía e Historia de las Ideas

Perspectivas filosóficas de la inteligencia artificial

TESIS

Que para obtener el grado de:

Maestro en Investigaciones Humanísticas y Educativas

Presenta:

Héctor Rodríguez Cristerna

Director de tesis:

Dr. Guillermo Nelson Guzmán Robledo

Co-director de tesis:

Dr. Leobardo Villegas Mariscal

Zacatecas, Zacatecas.

Diciembre de 2019

AGRADECIMIENTOS:

A la Unidad Académica de Docencia Superior y la Maestría en Investigaciones Humanísticas y Educativas por abrirme sus puertas y permitirme formar parte del cuerpo estudiantil.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT), por la beca otorgada para la realización de esta investigación.

A los docentes de la orientación en Filosofía e Historia de las Ideas, por compartir sus conocimientos, su amistad y su experiencia de vida.

De manera muy especial agradezco al Dr. Nelson Guzmán Robledo por su disposición y asesoría, cuyos consejos y observaciones fueron guía fundamental en esta investigación.

A mis padres y mi familia, por su comprensión y apoyo incondicional a lo largo de todos estos años.

A mi esposa Vanessa, por su amor, su motivación constante y su compañía afectiva e intelectual en todo momento.

A mi hija Linda Aleph, por contagiarme de su alegría y energía inagotable, motor para completar este trayecto.

Índice

Introducción

Durante gran parte de nuestra historia, hemos considerado a las máquinas como artefactos no pensantes, creaciones del ingenio humano desprovistas de alma. Los recientes avances en el desarrollo de la inteligencia artificial (IA) nos permiten replantear esta cuestión, y abren nuevos debates no sólo en cuanto a la factibilidad teórica y técnica de la IA, sino también en torno a sus repercusiones éticas y sociales, así como en nuestra manera de concebir la mente y al ser humano.

El presente trabajo nace de tales inquietudes. Más que un estudio lógico-matemático de la IA, el objetivo es bosquejar un panorama general en un doble movimiento: por un lado, mostrar las distintas teorías y perspectivas filosóficas que se encuentran en la base de la discusión sobre la inteligencia artificial; por otro, señalar cómo el desarrollo acelerado de potentes computadoras está reabriendo debates filosóficos clásicos en torno a conceptos como mente, conciencia, pensamiento, alma, vida, y esencia del ser humano.

La filosofía, la ciencia y la tecnología han seguido caminos que se intersecan y bifurcan. Desde los filósofos y naturalistas griegos como Aristóteles, geómetras como Platón, Pitágoras y la armonía de las esferas, los humanistas y científicos del Renacimiento, Giordano Bruno y su defensa del heliocentrismo copernicano, Descartes y la geometría analítica, son algunos ejemplos que nos demuestran que la ciencia y filosofía, como actividades en las que prima el intelecto humano, están más cerca de lo que aparentan. No es aventurado decir que la filosofía es madre de la ciencia: todavía hasta mediados del siglo XIX lo que ahora conocemos como física era estudiado bajo el nombre de "filosofía natural", como se refleja en una obra cumbre del pensamiento científico, en la que Newton estableció las bases de la física clásica: *Principios matemáticos de la filosofía natural* (1687).

Sin pretender elaborar una filosofía científica ni una ciencia filosófica, esta tesis tiene por objetivo explorar los múltiples puntos de contacto y vasos comunicantes que ligan estos ámbitos, en particular en torno a la inteligencia artificial, un tema actual que reabre importantes interrogantes acerca de lo que consideramos específicamente humano, y cuestiona su esencia misma.

La inteligencia artificial (IA) despierta un enorme interés lo mismo en los legos que en la comunidad científica, en literatos, artistas y pensadores de la posmodernidad. En la cultura popular, el escritor Philip K. Dick cuestiona los límites entre lo natural y lo artificial con sus robots idénticos a los seres humanos en su novela *¿Sueñan los androides con ovejas eléctricas?* Publicada en 1968, ese mismo año Kubrick nos presenta en su film *Odisea al espacio 2001* a la potente pero ingobernable supercomputadora HAL 9000, mientras que *Terminator* (1984) y *Matrix* (1999) nos muestran una imagen post apocalíptica de un mundo dominado por máquinas.

Estas visiones de una inteligencia artificial capaz de competir con el intelecto humano, antaño reservadas a la imaginación de cineastas y escritores, comienzan a filtrarse en la realidad. La tecnología computacional muestra una evolución acelerada: las primeras computadoras digitales datan de la década de 1940, hace menos de 100 años, y su capacidad de cálculo aumenta en progresión geométrica hasta niveles insospechados. Como punto de comparación, consideremos que el matemático inglés William Shanks (1873) trabajó durante 20 años para obtener 707 decimales de π (pi); en 1960, en los albores de la era computacional se lograron calcular unos 100 mil decimales mediante un ordenador IBM 7094, una máquina del tamaño de un piano y con un costo de 3 millones de dólares. En 2011, utilizando una computadora relativamente barata, Alexander J. Yee y Shigeru Kondo calcularon 10 billones de decimales. Y en 2019, utilizando computación en la nube (alojada en servidores externos), Emma Haruka Iwao logró el récord actual de 31 mil millones de dígitos.

La inteligencia artificial ha dejado de ser un tema meramente académico, y ahora abarca las esferas de la política, la ética y por supuesto la filosofía. No hace falta recurrir a ninguna teoría conspiratoria para vislumbrar que en el contexto de la sociedad actual, en donde el conocimiento es poder y se utiliza como una fuerza de dominio, el primero que desarrolle al máximo esta nueva tecnología ocupará un lugar preponderante a nivel mundial, a la altura de las gigantes compañías omnipresentes: Google, Facebook, Amazon, Apple, Microsoft, y los nuevos competidores asiáticos como Samsung y Huawei, compañías que mantienen sus propios proyectos de inteligencia artificial.

Los gobiernos comienzan a considerar sus repercusiones y alcances: China, llamada a ser la nueva potencia económica mundial, está llevando a cabo un ambicioso proyecto para convertirse también en la primera potencia en inteligencia artificial; Estados Unidos y la Unión Europea consideran la IA como una de las principales ramas de investigación y destinan a ella un considerable presupuesto, y las preocupaciones internacionales sobre el tema se reflejan en el *Informe de riesgos globales 2017* del Foro Económico Mundial.

¿Por qué la inteligencia artificial despierta en nosotros esta mezcla de fascinación y temor? Desde hace siglos, sobre todo a partir de la máquina de vapor, estamos acostumbrados a convivir con máquinas que nos superan claramente en fuerza y velocidad. Sin embargo la mayoría de nosotros aún consideramos que los humanos sobrepasamos a los objetos artificiales en aspectos como la adaptabilidad, el pensamiento creativo, el uso del lenguaje, el arte, las emociones y el intelecto, lo cual es verdadero hasta cierto punto. Actualmente las máquinas ya nos superan en algunas áreas tradicionalmente asociadas al intelecto, como la capacidad de cálculo y de memoria, y muestran progresos agigantados en otras, como el aprendizaje autónomo y el reconocimiento de patrones. Por eso, para algunos reputados científicos como Stephen Hawking no hay ninguna duda de que en algún punto la

inteligencia artificial logrará sobrepasar el intelecto humano: la verdadera interrogante es cuándo y cuáles serán sus efectos en la vida humana.

El debate sobre la inteligencia artificial requiere de finas precisiones. Por tal motivo, para abordar con mayor detalle en esta temática, en este proyecto se plantea utilizar un doble enfoque: 1) la filosofía, como pensamiento profundo y estructurado –que no técnico– acerca del universo y el hombre, y dentro de éste, el estudio de la *psique*; y 2) las ciencias computacionales, que anclan sus raíces en las matemáticas y la lógica.

Cada una de estas áreas no está exenta de sus propias polémicas y conflictos. En particular, la ciencia ha intentado demarcar claramente sus límites y diferencias, olvidando que su desarrollo se debe precisamente al esfuerzo de la filosofía, y que hubo una época en que ser filósofo y ser científico eran prácticamente sinónimos. Descartes, uno de los principales filósofos que inauguran el pensamiento moderno, también aporta en el *Discurso del método* el pilar de la ciencia moderna: la duda metódica. El empirismo de John Locke y David Hume, el enciclopedismo de la Ilustración y el positivismo de Augusto Comte serían fundamentales para sistematizar y delimitar el ámbito de la ciencia: simplificando burdamente, se afirmaba que sólo la experiencia sensible puede ofrecer una base para el conocimiento, y por tanto, sólo lo susceptible de comprobación en la realidad sería digno de estudio.

Afortunadamente para la historia del pensamiento humano, no todos los filósofos aceptaron esta supuesta frontera de realidad, impuesta desde una visión limitada y mecánica. Como encargada de pensar la totalidad de lo real y la totalidad de la experiencia humana, la filosofía continuó con el estudio de lo simbólico, lo onírico, lo desconocido, las pasiones, el arte, la experiencia estética, la intuición, las experiencias vitales inefables, lo real, y todo aquello imposible de ser caracterizado en marco estrechos, además de ahondar en el estudio crítico de la propia ciencia,

impugnando su supuesta neutralidad y el proyecto de modernidad basado en los conceptos de razón y progreso.

En cuanto a la ciencia moderna, la medicina y la psicología fueron las primeras encargadas del estudio del pensamiento y sus manifestaciones desde un punto de vista científico. La neurología y neuroanatomía serían las encargadas de estudiar el binomio mente-cerebro desde la biología. Y sólo a partir del desarrollo de nuevas tecnologías como los escáneres cerebrales y potentes ordenadores, fue posible desarrollar otras perspectivas de estudio con un enfoque basado ya no en la biología, sino en modelos computacionales que consideran al cerebro como una máquina excepcionalmente eficiente en el procesamiento y transmisión de la información.

Recapitulando, tenemos dos vértices claros y distintos que nos permitirán acercarnos a nuestro tema de estudio: 1) la filosofía, como un pensamiento crítico de la totalidad, el pensamiento, lo real y la experiencia humana; 2) la ciencia, en particular las ciencias computacionales, que estudian el sistema cerebro–mente aportando una explicación mecánica–cibernética de los procesos de generación, transmisión y transformación de la información que ocurren en su interior.

A partir de esta interrelación plantearemos nuevas interrogantes y replantearemos otras incógnitas ya clásicas, siempre abiertos a nuevas rutas que surjan en el camino de este proyecto: ¿Cuáles es el origen del pensamiento? ¿Cuál es la relación entre existencia y pensamiento? ¿La inteligencia es un atributo propiamente humano? ¿El pensamiento es un proceso biológico? ¿Los procesos biológicos mentales son computables y traducibles a dígitos? ¿Es posible replicar mediante sistemas electromecánicos el funcionamiento del cerebro humano? ¿Se puede construir un sistema artificial inteligente? ¿Los límites del pensamiento humano coinciden con los límites de la tecnología? ¿La tecnología puede evolucionar de manera autónoma? ¿Es posible construir una máquina capaz de reconocerse como un ente que existe? ¿Es posible replicar la experiencia humana en

un código de información? ¿Pueden las máquinas enseñarnos algo sobre la vida humana?

Todas estas interrogantes pueden sintetizarse a su vez en los dos ejes que guiarán nuestro proceso de investigación: 1) reflexionar cómo la tecnología afecta e influye nuestra concepción del ser humano, y 2) realizar una crítica de la inteligencia natural y artificial, entendiendo el concepto de crítica como un análisis profundo y no como objeción moral. En resumen, en esta tesis buscamos acercarnos al conocimiento de la mente, la inteligencia y la razón como conceptos relacionados y superpuestos, pero no equivalentes; estudiar la posibilidad de replicar estos procesos mediante sistemas electromecánicos, y en última instancia, cuestionar el pensamiento como la esencia misma del ser humano.

Capítulo 1

Inteligencia artificial, un horizonte lejano

> ¿Qué veo, sino sombreros y trajes en los que podrían ocultarse unos autómatas?
>
> Descartes, *Meditaciones metafísicas* (1641)

En el presente capítulo expondremos las razones que nos permiten sugerir que el desarrollo de la inteligencia artificial es un evento posible aunque poco probable a corto plazo, debido a importantes limitaciones técnicas actuales. El desarrollo de nuestro análisis nos exige considerar diversos argumentos en torno a dos hipótesis de trabajo: 1) que la inteligencia no depende del alma, sino del cerebro y sus procesos biológicos; y 2) que estos procesos cerebrales, o más específicamente sus resultados, podrían ser replicables en el futuro a través de medios mecánicos.

Con este objetivo, esbozaremos los diversos avances científicos que han intentado reducir la brecha entre lo biológico y lo mecánico, abriendo la posibilidad de construir un sistema artificial capaz de replicar las funciones cognitivas propias de los seres vivos.

1.1 La vida, el alma y la síntesis inorgánica

Si algo es evidente en nuestro mundo contemporáneo, es que la ciencia y la técnica avanzan a pasos agigantados. Esto no impide que aún en nuestros días exista cierta aura alrededor de los seres vivos y en particular del ser humano. De manera natural intuimos una diferencia cualitativa entre los sistemas biológicos y los mecánicos (máquinas), una frontera que suponemos infranqueable entre los seres animados e inanimados.

Dicha aura alrededor de los seres vivos se relaciona de manera estrecha con la concepción de un espíritu (*pneuma*) o alma (*anima*), aunque estos conceptos no son por completo equivalentes, y distintos pensadores los han definido de diversas maneras. Entre los filósofos de la antigua Grecia, Platón y Aristóteles consideran que el alma es la encargada de buscar la verdad, la que adquiere el saber, y por tanto la inteligencia sería una de sus cualidades. Sin embargo, ya en estos pensadores observamos una clara diferencia de planteamientos: Platón considera que el alma es un ente inmortal y sobrenatural, perteneciente al mundo de las ideas; por su parte, Aristóteles considera el alma como principio vital ligado al cuerpo, propio de la naturaleza, que no puede subsistir fuera del cuerpo, y por tanto apunta de manera indirecta a la mortalidad del alma.

En el *Fedón*, anticipándose a Descartes, Platón señala que el verdadero conocimiento sólo es posible apartándose de las sensaciones del cuerpo. Sólo el alma puede acercarse a la verdad:

> –¿Cuándo, entonces –dijo él [Sócrates]–, el alma aprehende la verdad? Porque cuando intenta examinar algo en compañía del cuerpo, está claro que es engañada por él. […]
>
> [El conocer] lo hará del modo más puro posible quien en rigor máximo vaya con su pensamiento solo hacia cada cosa, sin servirse de ninguna visión al reflexionar, ni arrastrando ninguna otra percepción de los sentidos en su razonamiento, sino que, usando sólo de la inteligencia pura por sí misma, intenta atrapar cada objeto real puro, prescindiendo todo lo posible de los ojos, los oídos, y en un palabra, del cuerpo entero, porque le confunde y no le deja al alma adquirir la verdad y el saber cuando se le asocia. (pp. 41-42)

En este mismo diálogo, Platón afirma que el alma siempre conlleva la vida; si la muerte es lo contrario de la vida y el alma no puede ser contraria a sí misma, esto significaría que el alma es inmortal:

> El alma jamás admitirá lo contrario a lo que ella conlleva [la vida]. El alma no acepta la muerte […] por tanto el alma es inmortal […]

> Si lo inmortal es imperecedero, es imposible que el alma, cuando la muerte se abata sobre ella, perezca […] ¿qué otra cosa sería el alma, si es que es inmortal, sino indestructible? Al sobrevenirle entonces al ser humano la muerte, según parece, lo mortal en él muere, pero lo inmortal se va y se aleja, salvo e indestructible, cediendo el lugar a la muerte. […] nuestra alma es inmortal e imperecedera, y de verdad existirán nuestras almas en el Hades. (pp. 119-121)

Cabe añadir que en el *Fedón* también se nos muestra otra interesante concepción del alma, similar a la armonía que surge de la combinación de notas en un instrumento, pero que Platón finalmente desecha en favor de la idea de un alma inmortal y trascendente:

> También acerca de la armonía, de la lira y de sus cuerdas, podría sostener uno ese mismo argumento, que la armonía es invisible, incorpórea, y algo muy hermoso y divino que está en la lira bien ajustada, mientras la misma lira y las cuerdas son cuerpos, y corporales, compuestos y terrestres, y congénitos a lo mortal […] Si, entonces, resulta que nuestra alma es una cierta armonía, está claro que, cuando nuestro cuerpo sea relajado o tensado desmedidamente por las enfermedades y otros rigores, al punto al alma se le presenta la urgencia de perecer, aunque sea divinísima, como es también el caso de las otras armonías, las que se crean en los sonidos, y en todas las labores de los artesanos, mientras que los despojos del cuerpo de cada uno aún permanecen un largo tiempo, hasta ser quemados o pudrirse.
>
> Mira pues, qué vamos a decir contra este argumento, si alguno considera que el alma, siendo una combinación de los factores existentes en el cuerpo, en lo que llamamos muerte perece la primera. (pp. 81-82)

En esta perspectiva desechada por Platón encontramos un importante antecedente acerca de la inteligencia como un proceso emergente, diferente de los órganos que le dan existencia, pero vinculado necesariamente a ellos. Por otra parte, sosteniendo una postura distinta al idealismo de Platón y más cercana a la biología, Aristóteles argumenta en *De anima*:

> Resulta, sin duda, necesario establecer en primer lugar a qué género pertenece y qué es el alma —quiero decir si se trata de una realidad individual, de una

> entidad o si, al contrario, es cualidad, cantidad o incluso cualquier otra de las categorías que hemos distinguido.
>
> […] si hay algún acto o afección del alma que sea exclusivo de ella, ella podría a su vez existir separada; *pero si ninguno le pertenece con exclusividad, tampoco ella podrá estar separada.* […] El inteligir parece algo particularmente exclusivo de ella; pero ni esto siquiera podrá tener lugar sin el cuerpo […]
>
> […] parece que las afecciones del alma se dan con el cuerpo: valor, dulzura, miedo, compasión, osadía, así como la alegría, el amor y el odio. El cuerpo, desde luego, resulta afectado conjuntamente en todos estos casos. Por consiguiente, y si esto es así, *está claro que las afecciones son formas inherentes a la materia*. De manera que las definiciones han de ser de este tipo: el encolerizarse es un movimiento de tal cuerpo o de tal parte o potencia producido por tal causa con tal fin. De donde resulta que corresponde al físico ocuparse del alma. (p. 132-135; las cursivas son nuestras)

A pesar de que Aristóteles ya sugiere la necesidad de un estudio conjunto de alma y cuerpo, otros filósofos sostendrán la existencia de un principio no material que fundamenta la vida, una fuerza distinta de la energía estudiada por la física y otras ciencias. Este punto de vista será la base de la escuela vitalista en biología, que postula que los organismos vivos poseen una fuerza o impulso vital que los diferencia en su esencia de las cosas inanimadas. Esta fuerza inmaterial, actuando sobre la materia organizada, daría como resultado la vida y sin ella sería imposible su existencia; de ahí la diferencia radical entre un ser humano vivo y un cadáver. Uno de los máximos expositores del vitalismo es Henry Bergson con su concepto de *élan vital* (principio vital), que desarrolla en su libro *La evolución creadora* (1907):

> La filosofía evolucionista extiende sin duda a las cosas de la vida los procedimientos de explicación que han tenido éxito para la materia bruta. […] ¿Es preciso, pues, renunciar a profundizar en la naturaleza de la vida? ¿Es preciso atenerse a la representación mecanicista que el entendimiento nos dará siempre, representación necesariamente artificial y simbólica, ya que estrecha la actividad total de la vida en forma de una cierta actividad humana, la cual no es más que una manifestación parcial y local de la vida, un efecto o un residuo de la operación vital? […] El análisis descubrirá sin duda en los procesos de creación orgánica un número creciente de fenómenos físico-químicos. Y a ellos

> se atendrán los químicos y los físicos. Pero no se sigue de ahí que la química y la física deban darnos la clave de la vida.
>
> […] Un elemento muy pequeño de una curva es casi una línea recta. Tanto más semejará a una línea recta cuanto más pequeño se le tome. En el límite, se dirá, según se quiera, que forma parte de una recta o de una curva. En cada uno de sus puntos, en efecto, la curva se confunde con su tangente. Así la "vitalidad" es tangente en no importa qué punto a las fuerzas físicas y químicas; pero estos puntos no son, en suma, más que las consideraciones de un espíritu que imagina detenciones en tales o cuales momentos del movimiento generador de la curva. En realidad, la vida no está hecha de elementos físico-químicos, como una curva no está compuesta de líneas rectas.[1] […]
>
> Las causas y los elementos que descubre el análisis físico-químico son causas y elementos reales, sin duda, para los hechos de destrucción orgánica; y lo son en número limitado. Pero los fenómenos vitales propiamente dichos, o hechos de creación orgánica, nos abren, cuando los analizamos, la perspectiva de un progreso hasta el infinito. (pp. 464-465)

Bergson no niega que la inteligencia provenga de la evolución biológica, pero sostiene que jamás se podrá explicar el fenómeno de la vida mediante el mero razonamiento, debido a que la propia inteligencia es sólo una pequeña parte de la vida en su totalidad:

> La historia de la evolución de la vida, por incompleta que todavía sea, nos deja entrever cómo se ha constituido la inteligencia por un progreso ininterrumpido, a lo largo de una línea que asciende, a través de la serie de los vertebrados, hasta el hombre. Ella nos muestra, en la facultad de comprender, un anexo de la facultad de actuar, una adaptación cada vez más precisa, cada vez más compleja y flexible, de la conciencia de los seres vivos a las condiciones de existencia que les son dadas. De ahí debería resultar esta consecuencia: que nuestra inteligencia, en el sentido restringido de la palabra, está destinada a asegurar la inserción perfecta de nuestro cuerpo en su medio, a representarse las relaciones de las cosas exteriores entre sí; en fin, a pensar la materia. […]
>
> Pero de ahí debería resultar también que nuestro pensamiento, en su forma puramente lógica, es incapaz de representarse la verdadera naturaleza de la vida, la significación profunda del movimiento evolutivo. Creado por la vida en

[1] Resulta curioso hacer notar que para las matemáticas modernas una línea recta se puede definir como una circunferencia de radio infinito.

> circunstancias determinadas, para actuar sobre cosas determinadas, ¿cómo abrazaría él la vida, si no es más que una emanación o aspecto suyo? Depositado, en el curso de su ruta, por el movimiento evolutivo, ¿cómo podría aplicarse a lo largo del movimiento evolutivo mismo? Otro tanto valdría pretender que la parte iguala al todo, que el efecto puede reabsorber en él su causa, o que el canto rodado abandonado en la playa dibuja la forma de la ola que le ha traído hasta ella. De hecho, nos damos perfecta cuenta que ninguna de las categorías de nuestro pensamiento —unidad, multiplicidad, causalidad mecánica, finalidad inteligente, etc.—, se aplica exactamente a las cosas de la vida. (pp. 433-434)

Desde la visión vitalista y posturas afines, sería inadmisible la sola posibilidad de una inteligencia artificial. Si se considera el intelecto como una de las funciones o un tipo de alma, o como una función propia de los seres vivos, es lógico concluir que los seres inanimados nunca podrían acceder o desarrollar facultades como el entendimiento. En la actualidad, este punto de vista es sostenido por Roger Penrose (1989), quien considera que la esencia de la actividad mental no es computable por medio de las leyes físicas, y por tanto ninguna máquina de computación podrá replicar la inteligencia de un ser humano.

En el campo de la ciencia, la diferencia entre lo vivo y lo inanimado se tradujo en una distinción entre lo orgánico y lo inorgánico. En los círculos de la química y la biología, se creía que los compuestos orgánicos solo podían proceder de seres vivos. Pero en 1828 Friedrich Wöhler sintetizó urea, un compuesto orgánico presente en la orina, a partir de compuestos inorgánicos. Al respecto, el propio Wöhler escribiría posteriormente en una de sus cartas haber sido el primer testigo de "una gran tragedia de la ciencia, la muerte de una bella hipótesis [el vitalismo] por un hecho feo [la síntesis artificial]".

Este descubrimiento repercutiría más allá de la química, pues cuestionaba la diferencia ontológica entre lo vivo y lo inanimado: tal diferencia no sería cualitativa como afirmaba el vitalismo, sino referente a los diferentes grados de organización de la materia. En otras palabras, con las condiciones adecuadas sería posible crear vida de la materia inerte. Lo anterior es precisamente la idea central de la teoría de

la abiogénesis (*a-bios*, sin vida; *génesis*, origen), propuesta por Alexander Oparin en 1924 y John Haldane en 1928, la cual sugiere que el origen de las primeras células vivas se dio a partir de materia orgánica, producto de la síntesis abiótica de los compuestos presentes en la atmósfera y los océanos primigenios, mediante la acción de diversas fuentes de energía.

En 1953 se encontró evidencia experimental que respaldaba la abiogénesis. En ese año Stanley Miller y Harold Clayton Urey replicaron en un laboratorio las condiciones primigenias de nuestro planeta y lograron sintetizar aminoácidos, azucares y ácidos nucleicos que forman la base de los seres vivos, comprobando así que en las condiciones ambientales adecuadas se pueden formar moléculas orgánicas a partir de sustancias inorgánicas simples.

En 2010, Craig Venter y un equipo de científicos lograron crear la primera célula sintética. Su genoma era una copia de la bacteria *Mycoplasma mycoides*, pero su ADN completo había sido sintetizado por métodos químicos, por lo que en teoría se podría haber replicado cualquier genoma o crear uno nuevo. Este siguiente paso se logró con ciertos matices pocos años después: en 2016, Venter y Hamilton Smith crearon una bacteria del tipo *mycoplasma* con un nuevo genoma mínimo de 473 genes que no existía en la naturaleza.

Es lo más cerca que se ha llegado de crear vida artificial, pero aún no se le puede llamar como tal: aunque su genoma fue completamente sintetizado por medios químicos, en sentido estricto la información genética no fue codificada desde cero, ya que se tomó como base una secuencia de ADN conocida: el procedimiento consistió en eliminar mediante prueba y error cada uno de los 525 genes de la bacteria *Mycoplasma genitalium*, hasta conseguir un microorganismo con menor carga genética pero completamente funcional, incluyendo la capacidad de replicarse. Cabe destacar que aunque estos 473 genes son indispensables para mantener con vida al organismo, se ignora la función de una tercera parte de ellos.

En pocas palabras, en teoría se podría sintetizar y modificar cualquier tipo de código genético, pero aún no se conoce con exactitud cómo escribir el lenguaje de la vida.

Otros avances en genética son relevantes: en 2014 Romesberg *et al.* lograron añadir dos nuevas bases de nucleótidos al ADN de una bacteria, además de las cuatro conocidas en todos los seres vivos (adenina, guanina, citosina y timina); en 2019 Hoshika *et al.* añadieron cuatro bases nuevas, lo que significa que las posibilidades de combinación y almacenamiento de información en el ADN aumentan de forma exponencial. Esta nueva área de conocimiento se conoce como biología sintética, que se define como la síntesis de biomoléculas o la ingeniería de sistemas biológicos con funciones nuevas que no se encuentran en la naturaleza.

Esto abre la posibilidad de otra forma de inteligencia artificial: no una inteligencia electromecánica, sino una inteligencia artificial orgánica, sintetizada en laboratorio y diseñada a la medida. O bien una inteligencia artificial creada en los propios organismos biológicos, pero basada en una evolución acelerada: humanos con carga genética modificada. Si tomamos en cuenta que compartimos el 99% del ADN con los chimpancés, bastaría con modificar otro 1% para crear superhombres, en un sentido que quizá Nietzsche no imaginaba cuando escribió en *Así habló Zaratustra* (1883/1892): "¿Qué es mono para el hombre? Una irrisión o una vergüenza dolorosa. Y justo eso es lo que el hombre debe ser para el superhombre: una irrisión o una vergüenza dolorosa" (p. 34).

Como hemos visto, existe evidencia suficiente para sostener que no es necesario recurrir a un principio vital distinto de la materia inorgánica para explicar la aparición de la vida, y con ella el desarrollo de la inteligencia. Esto no resuelve de manera directa la cuestión de si es posible la creación de inteligencia artificial, pero sugiere que con el suficiente grado de complejidad y especialización de los dispositivos es posible alcanzar por medio de mecanismos inorgánicos comportamientos similares a los observados en los seres vivos.

Tal hipótesis ya se ha comprobado para organismos simples. En 2012, diversos científicos de la Universidad de Stanford (Karr *et al.*, 2012), construyeron un modelo computacional de la bacteria *Mycoplasma genitalium*, que permitía emular su comportamiento biológico en un programa informático, incluyendo su reacción a diferentes químicos. Dicha bacteria es usada frecuentemente para este tipo de experimentos porque es el organismo con el genoma más pequeño que se conoce (525 genes), y por tanto, con el comportamiento menos difícil de emular en un código de programación. Con ello, comprobaron la hipótesis de que los procesos biológicos funcionan de forma parecida a los procesos computacionales, como cadenas de instrucciones que se ejecutan en los organismos.

Y en 2017, Lechner, Grosu y Hasani lograron replicar las respuestas del sistema nervioso del nematodo *Caenorhabditis elegans*, de solo 300 neuronas, en un ordenador. Debido a su reducido tamaño, el sistema nervioso de este gusano puede representarse como un diagrama de circuito, y por tanto es posible replicar su actividad neuronal en una computadora. La similitud de esta red neuronal artificial con el sistema nervioso del nematodo le permitía al modelo computacional comportarse como lo haría el gusano vivo y aprender sin programación adicional.

Estos avances deben ser considerados en su justa dimensión. Compárense los 525 genes de la *Mycoplasma genitalium* con los más de 20 mil genes humanos, de los cuales aún no se conoce ni el número exacto ni la función de todos ellos; o las 300 neuronas del nematodo *Caenorhabditis elegans* con los 86 mil millones de neuronas del cerebro humano (Herculano-Houzel, 2012), y más importante aún, con el número de conexiones posibles entre ellas. Con esto en mente, podemos ver claramente que los avances científicos ya descritos son una prueba de principio de que es posible diseñar un cerebro mejorado, o replicar su funcionamiento con medios electrónicos, pero de ningún modo significa que en la actualidad se tengan los medios técnicos para llevar a cabo tales proezas, por lo menos no a un nivel equiparable al cerebro humano… por el momento.

1.2 Inteligencia y comportamiento inteligente

Para el físico teórico Stephen Hawking, el desarrollo de máquinas capaces de replicar el comportamiento humano es sólo una cuestión de tiempo. Ya en "Computing Machinery and Intelligence" (1950), el matemático y uno de los padres de la computación moderna Alan Turing señalaba que para el año 2000 se habría logrado alcanzar el nivel de la inteligencia humana, y aunque es evidente que sus predicciones no se han cumplido, ciertamente se han logrado hitos importantes: por citar un ejemplo mediático, en 1996 la computadora Deep Blue fue capaz de ganar una partida al entonces campeón mundial de ajedrez, Gary Kasparov, y al año siguiente una versión actualizada lo venció en un match a 6 partidas. Desde entonces, la potencia de las máquinas no ha parado de crecer. La brecha se ha vuelto insalvable en estas dos décadas: los 2,882 puntos elo del actual campeón de ajedrez Magnus Carlsen –la puntuación humana más alta de la historia– palidecen frente a los más de 3,800 puntos que alcanzan los nuevos programas. Frente a estos potentes softwares, el campeón del mundo tiene la misma probabilidad de ganar que un jugador amateur: cero. De acuerdo con Nick Bostrom (2014), los expertos actuales en IA consideran que el nivel de inteligencia de un ser humano se alcanzará dentro de un plazo de 50 años a 100 años, pero una vez que se alcance, la progresión será exponencial. Si consideramos que la historia humana se mide en milenios, un siglo no parece tanto.

Es necesario destacar que en las primeras líneas del párrafo anterior hemos escrito "máquinas replicar el comportamiento humano", y no "el pensamiento humano", y ello por una buena razón: desde las ciencias computacionales se considera que el pensamiento es un proceso, y el comportamiento inteligente es el resultado de tal proceso. Si bien es lógico suponer que la reproducción fiel de un proceso determinado tendrá un resultado idéntico al que se consigue con el proceso original (por ejemplo, seguir una receta de cocina), también es posible que dos

procesos distintos produzcan resultados similares (sumar 6+2 y multiplicar 4x2). Esto sucede también en la naturaleza, en un proceso conocido como evolución convergente, cuando dos estructuras similares han evolucionado independientemente a partir de estructuras ancestrales distintas y procesos de desarrollo muy diferentes, como la evolución del vuelo en insectos, aves y murciélagos, o las aletas de peces y delfines.

En el caso del desarrollo de la inteligencia artificial, lo que interesa en principio es construir una máquina capaz de exhibir un comportamiento similar al que efectúan los seres humanos, sin importar si el proceso interno es idéntico al del cerebro.[2] Alan Turing (1950) es el primero en plantear formalmente esta cuestión, proponiendo el "Juego de la imitación", que consiste básicamente en sustituir la pregunta "¿pueden las máquinas pensar?" por otra que en términos generales podríamos formular de la siguiente manera: ¿pueden una máquina comportarse lingüísticamente, esto es, interactuar lingüísticamente con un ser humano, de manera que éste no pueda distinguir que se encuentra ante una máquina? Así, el llamado Test de Turing consiste en un interrogatorio exhaustivo llevado a cabo por un humano a otro interlocutor, con el objetivo de determinar si se encuentra ante otro ser humano o una máquina.

En vista de que este problema constituye un clásico en el debate sobra la IA, permítasenos replantearlo en un ámbito que tal vez nos resulte más familiar: imaginemos un equipo de escultores y geólogos expertos que encuentran la manera de reproducir perfectamente –o falsificar, como se quiera ver– el *David* de Miguel Ángel, una réplica tan exacta en su más mínimo detalle, en cada fisura, en cada centímetro, que nadie, ni los constructores, ni siquiera el propio Miguel Ángel

[2] En realidad, el estudio del funcionamiento y la estructura del cerebro humano es una importante área de investigación. Como se ha dicho antes, la forma más directa de obtener un comportamiento determinado sería reproducir de la manera más fielmente posible el proceso que le ha dado origen, pero no es necesariamente la única forma de conseguirlo. Actualmente existen enormes dificultades técnicas para lograr la construcción de un cerebro mecánico, por lo que se vuelve necesaria la búsqueda de procesos y estructuras alternativas.

podría distinguirla del original. Al terminar de fabricar la réplica, por un error de museografía se pierden las placas que distinguían la copia del original.

Es evidente que existe un original y una copia, pero también es evidente que dicho conocimiento es totalmente inútil, pues no hay manera de saber cuál es cuál. A efectos prácticos, ambos podrían ser considerados originales… o copias, aunque intuimos que la Galería de Florencia preferirá la primera opción. Ya que es imposible distinguirlos, es lógico suponer que los observadores tendrán la misma experiencia estética ante uno u otro, sobre todo si cada pieza es exhibida de manera individual, pero ciertamente estarían algo confundidos si se les presentasen simultáneamente. ¿Acaso el *David* no era una pieza única? ¿Qué ha pasado con el aura, con la esencia de la obra de arte? ¿El original ha perdido su estatus? ¿La copia tiene el mismo valor artístico que el original? ¿Es pertinente seguir hablando de original y copia?

Tales cuestiones se pueden trasladar directamente al debate en torno a la inteligencia humana y la inteligencia artificial. La postura de Alan Turing es clara y despierta polémica: si una máquina exhibe un tipo de comportamiento lingüístico continuo y consistente que sólo se podría esperar de un ser humano, de manera que resulte imposible para un observador distinguir su identidad, estamos ante una máquina que piensa. Es necesario recalcar esto último para evitar imprecisiones posteriores: Turing habla de una máquina que piensa, no de un humano construido de manera artificial. El matemático y filósofo inglés evita caer en la tentación de identificar una máquina pensante con la esencia misma del hombre.

Asimismo, debemos señalar que el Test Turing supone de manera implícita que el lenguaje verbal es la principal manifestación de la inteligencia humana, lo cual le ha valido críticas que señalan que si bien el lenguaje es una característica y prueba de la inteligencia humana, no es la única. Por tanto, se podría enunciar un Test de Turing ampliado: la prueba será ahora comparar los comportamientos de una máquina, lingüísticos o de otro tipo, con aquellos que sólo se podrían esperar de un ser humano y que son catalogados como comportamientos inteligentes. Entre

ellos podríamos señalar: reconocer patrones de cáncer en imágenes médicas, aprender a jugar ajedrez, organizar grandes cantidades de datos, desarrollar estrategias de publicidad personalizada. En todo esto, las máquinas actuales superan a los expertos humanos.

Se podría objetar, con justa razón, que aunque la máquina exhiba un comportamiento que pudiéramos definir y aceptar como inteligente, esto no significa que piense, pues en su interior podría existir un mecanismo distinto a los procesos cerebrales con un resultado externo idéntico a éstos, cuestión que ya hemos apuntado. Desde dicha perspectiva se considera incorrecto identificar pensamiento con comportamiento, objeción que dicho sea de paso, también aplica para el conductismo y su concepto de la mente humana como una caja negra imposible de estudiar en sí misma, y sólo cognoscible a través de sus manifestaciones (*outputs*). En esta misma línea, González (2007) señala acertadamente que Turing postula una definición operacional de la inteligencia, alejada del ámbito de lo mental.

Aunque esta objeción es en esencia válida, también es irrelevante para el desarrollo de la IA. Si bien es verdad que los procesos internos de una máquina y de un cerebro humano pueden ser distintos –lo más probable es que lo sean– esto no implica que el término pensamiento pueda ser aplicado con mayor rigor a uno u otro. Después de todo, aunque suponemos que nuestros procesos mentales son similares a los de otros seres humanos no poseemos de manera inmediata dicha certeza,[3] pues carecemos de un acceso directo a los pensamientos de los demás y sólo nos resultan verificables a través de su lenguaje, sus comportamientos, sus expresiones materiales y sus manifestaciones corporales. Ni siquiera conocemos por completo cómo funciona nuestra propia mente, lo cual no impide que pensemos y

[3] Las neurociencias, a través del estudio orgánico y eléctrico del cerebro, nos dirán que todos los seres humanos compartimos aproximadamente las mismas funciones en las mismas áreas cerebrales, con algunas excepciones. Por ejemplo, en un número muy reducido de humanos el área principal del lenguaje se encuentra en el hemisferio derecho en lugar del izquierdo como es habitual, o en los casos de lesión cerebral un área suple las funciones de otra dañada, a lo cual se le conoce como neuroplasticidad.

actuemos. Además, aún cuando los procesos de una máquina sean diferentes a los del cerebro humano, las respuestas y comportamientos de la IA implican necesariamente que realiza algún tipo de procesamiento de información; que se le llame pensamiento o con cualquier otro término no cambia sus atributos ni sus resultados.

Con lo anterior no se busca demeritar en lo más mínimo el estudio de la interioridad y subjetividad humana, sino destacar que en un primer momento el desarrollo de la IA no apunta a tales cuestiones. Esto nos revela una de las críticas más acertadas a la ciencia en general: su pragmatismo y utilitarismo. Por ejemplo, lo que se busca es desarrollar un algoritmo que pueda diagnosticar con un alto grado de certeza un tumor y un robot capaz de extirparlo con mayor precisión que el cirujano más experimentado, y no tanto una máquina que aprenda a meditar o se pregunte por el sentido de la vida, por mucho que tales cuestiones puedan resultar trascendentes y significativas en la vida humana.

Por tanto, tenemos una proposición básica derivada de la postura de Turing, que será clave en el debate de la IA: si un ente, ya sea biológico o mecánico, exhibe un comportamiento lingüístico o de otro tipo que puede ser catalogado como inteligente, ello implica que de manera subyacente existe algún tipo de procesamiento de información, al cual en el ámbito humano lo denominamos con el término de pensamiento o inteligencia. Para centrarnos en la hipótesis de Turing consideraremos dichos términos como sinónimos, pero es necesario recalcar que el pensamiento humano abarca formas no necesariamente racionales ni conscientes, como las emociones, el inconsciente, el enamoramiento o la experiencia estética, todas ellas enriquecedoras de la experiencia vital humana. En todo caso, es igualmente válido señalar que los pensamientos racionales como el cálculo y la lógica también son una forma de pensamiento, y éstos sí pueden ser replicados por computadoras hasta el punto de superarnos.

Por supuesto, no todos comparten dicha proposición. La postura más radical sostiene que si un comportamiento inteligente no es producto de un cerebro humano, aunque sea idéntico a éste en sus consecuencias o incluso en sus procesos, no se le puede llamar pensamiento. Otra objeción similar es que el pensamiento sólo puede provenir de seres vivos, entendiendo por tales seres biológicos, y de ninguna manera de artefactos construidos artificialmente. Estas posturas, cuyos principales defensores son John Searle (1980) y Roger Penrose (1989), impiden por principio todo debate ulterior, por lo que si queremos discutir acerca de la existencia de otras formas de pensamiento debemos admitir en un primer momento su posibilidad.

Concordemos entonces en la existencia del pensamiento humano y concedamos por un momento la posibilidad de un pensamiento no-humano, entendido como un procesamiento de información de alta complejidad llevado a cabo por alguna máquina no biológica.[4] La apuesta de quienes trabajan en el desarrollo de la IA es que llegará el momento, como en caso de las estatuas del *David*, en que sea imposible distinguir uno de otro a partir de sus manifestaciones, lo cual a su vez conlleva consecuencias filosóficas y éticas, que serán el tema de este trabajo.

Otra objeción frecuente es que una máquina jamás podrá sentir y vivir como un ser humano. Es otra crítica perfectamente válida, pero recordemos una vez más que el proyecto de IA planteado por Turing no trata sobre construir humanos –en esa cuestión, nuestra especie tiene 7 mil millones de especímenes de ventaja–, sino una máquina que piense como humano, o más específicamente, que sea capaz de responder y actuar como un humano.

Tal proyecto se conoce como inteligencia artificial fuerte o general, una máquina capaz de actuar en los diversos contextos en los que se desenvuelve el ser humano con la misma o mayor eficacia que cualquier especialista, desde ordenar

[4] El término máquina no biológica cobra relevancia dada la posibilidad real del cultivo de tejido cerebral fuera del cuerpo humano, es decir, la posibilidad de construir un cerebro biológico a la medida y de manera artificial, algo que podríamos considerar una máquina biológica.

una casa hasta elaborar teorías sobre el funcionamiento del universo, y por qué no, escribir poesía, hacer música o pintar cuadros. Es el santo grial de los investigadores en inteligencia artificial, y es el nivel que los expertos consideran que se alcanzará dentro de los próximos 50 o 100 años (Bostrom, 2014).

Por el momento, hay consenso en que solamente se ha alcanzado el nivel de la IA débil o sistemas expertos, una inteligencia desarrollada para cumplir tareas específicas, en las cuales el sistema es capaz de tomar decisiones autónomas para lograr su objetivo sin necesidad de instrucciones adicionales o manipulación constante por parte de seres humanos, esto es, que sus procesos se encuentran automatizados. En algunos campos la IA ya nos ha superado, como jugar al ajedrez, al go y al póker, y están en caminos de hacerlo en otras tareas, como manejar un automóvil, reconocer patrones o pilotar un avión.

Ahora bien, es importante resaltar nuevamente que desde un punto de vista meramente ontológico una máquina jamás será un humano, como la copia del David jamás será el David original; pero al igual que la copia, puede llegar a ocupar el lugar del original. Siguiendo con el ejemplo de la escultura, se podría objetar con justa razón que en el duplicado no hay originalidad ni intención artística, y es tan sólo una reproducción. Más que una objeción, es la descripción precisa del proyecto de la IA: se considera que la naturaleza y la evolución tiene la autoría de la mente humana, mientras que los investigadores buscan recrear dicho modelo y los comportamientos derivados de éste.

Como señala Parra Díaz (2016), en el fondo del debate entre quienes admiten la posibilidad de crear inteligencia artificial y quienes la niegan se encuentran dos posturas radicalmente distintas: los que conciben el cerebro como materia, y los que consideran que existe algo más, imposible de reducir a la materia. Mateo Seco (2013) lo formula claramente: “Las relaciones entre la inteligencia humana y el sustrato físico constituyen un enigma que sigue suscitando sus preguntas con fuerza y con renovado interés. La pregunta fundamental es la misma que hace siglos: ¿muestra

el conocimiento humano, en sí mismo, la existencia de un elemento en el hombre que está más allá de la materia?" (p. 260).

Quienes consideran que es posible el desarrollo de la inteligencia artificial, suponen en líneas generales que en el pensamiento humano no hay ninguna realidad sobrenatural: el cerebro es materia, y el pensamiento es un proceso que se realiza en dicha materia: "la tesis de que todas las funciones mentales, incluso la conciencia, son propiedades emergentes de la estructura biológica del cerebro, es hoy en día el paradigma de la neurología" (Suay Belenguer, 2012, p. 166). Esto no significa que el pensamiento sea en sí mismo materia, sino que depende necesariamente de la materia y energía para llevarse a cabo: no puede haber pensamiento sin cerebro o alguna máquina que lo simule, como sería el caso de la inteligencia artificial. A su vez, el cerebro está intrínsecamente diseñado para pensar. En otras palabras, todas nuestras ideas y pensamientos, incluso aquellos que no pronunciamos o escribimos en papel, no se encuentran flotando sobre la nada: su soporte material se encuentra en la red neuronal y los impulsos eléctricos.

Es evidente que si para pensar es necesario contar un alma, un espíritu, o cualquier otra entidad divina o sobrenatural, el desarrollo de la IA sería una tarea imposible y ociosa, por lo cual Turing (1950) considera a ésta una objeción teológica, en el marco de un debate ajeno a la ciencia y la filosofía. Y por supuesto, decir que la creación de la inteligencia artificial es imposible porque las máquinas actuales no la han alcanzado, es una falacia de petición de principio. Aunque tampoco deberíamos tener tanta prisa por su desarrollo: como veremos, las consecuencias de crear una IA fuerte puede que no sean necesariamente benéficas para los humanos. La historia ha verificado una y otra vez que el progreso tecnológico no conlleva bondad, como lo demuestra el desarrollo de bomba atómica. De hecho, Nick Bostrom (2014) y Stephen Hawking opinan la IA podría convertirse en el mayor error de la humanidad. Por el momento, nos centraremos en considerar los enormes impedimentos que aún quedan por resolver para construir una IA fuerte.

1.3 Limitaciones técnicas y alcances actuales de la IA

Señalemos los obstáculos técnicos actuales para emular el pensamiento humano, comenzando con la potencia de procesamiento. De acuerdo al Instituto AI Impacts (2015), se calcula que el cerebro humano es capaz de procesar el equivalente a un exaflop (un 1 seguido de dieciocho ceros, en notación científica 10^18 operaciones de punto flotante por segundo). En comparación, las dos supercomputadoras más potentes del mundo en 2019, la china Sunway TaihuLight y la estadounidense Summit, "sólo" trabajan a una velocidad promedio de 93 y 122 petaflops respectivamente (un petaflop son 10^15 operaciones de punto flotante por segundo). Una enorme diferencia: se requieren mil petaflops para alcanzar un exaflop, lo que significa que la Summit, la mejor supercomputadora actual, tiene apenas la octava parte de potencia de procesamiento del cerebro humano. Otra medida de eficiencia en las computadoras son los TEPS, acrónimo de *Traversed Edges Per Second*, que miden la capacidad interna de un ordenador para transmitir información de un punto a otro dentro del propio sistema. Por ejemplo, la supercomputadora de IBM Sequoia, en el top 10 mundial, alcanza los 23 billones (2.3x10^13) de TEPS, mientras que el cerebro humano oscila entre los 18 y los 640 millones de TEPS, por lo que sería hasta 30 veces más potente que la Sequoia (Grace y Christiano, 2015).

Todo ello sin contar con la altísima eficiencia energética de nuestro encéfalo: consume entre 15 y 20 watts por hora: se necesitan tres cerebros para encender un foco incandescente. En comparación, un horno de microondas consume cerca de mil watts, y la Sunway TaihuLight utiliza 15 millones de watts. Para dejarlo más claro: nuestro cerebro utiliza mucha menos energía, en un espacio mucho más compacto y al mismo tiempo es más potente que las mejores supercomputadoras actuales. Y si esto nos parece poco, aún falta añadir a esta comparación todo nuestro demás hardware biológico, un magnífico cuerpo y sus múltiples órganos y sistemas, que

nos permite interactuar eficazmente con nuestro medio e interpretar de manera casi inmediata la información que recibimos.

Otra imposibilidad técnica es la manera en que está construida la propia red neuronal. Si bien el consenso científico admite que las neuronas son la unidad básica del cerebro y funcionan como una compuerta lógica de tipo binario (un interruptor eléctrico que envía o no una señal a otra neurona), la enorme cantidad de neuronas y conexiones entre ellas (sinapsis) es impresionante: un cerebro adulto cuenta con 86 mil millones de neuronas (Herculano-Houzel, 2012), y cada una de ellas tiene un promedio de 7 mil conexiones sinápticas que funcionan de manera bidireccional, esto es, que son capaces tanto de enviar como de recibir impulsos de otras neuronas. Se calcula que el cerebro de un niño de tres años tiene 10^15 sinapsis (un 1 seguido de quince ceros), cantidad que disminuye y se estabiliza en la vida adulta. En comparación, en un experimento de 2013 la supercomputadora japonesa Fujitsu K tardó 40 minutos en simular un segundo de procesamiento de una red neuronal humana. A fines de 2018 fue puesta en funcionamiento la supercomputadora SpiNNaker (*Spiking Neural Network Architecture*), resultado de un proyecto de 12 años y diseñada específicamente para emular el funcionamiento del cerebro humano, con mil millones de microprocesadores que realizan la función de neuronas en tiempo real; a pesar de ser un hito en la computación actual, representa apenas el 1.2% de las neuronas de un cerebro humano. Vemos entonces que aunque desde una perspectiva meramente física y biológica el cerebro es materia, es injusto reducirla a simple materia: su estructura es altamente compleja y eficiente. Una materia que definitivamente no es tan simple.

Igual de importante que las anteriores imposibilidades técnicas es que aún no se conoce con exactitud la manera en que nuestro cerebro funciona. Incluso si se pudiera construir una red idéntica a la neural, una réplica mecánica de un cerebro con una potencia de procesamiento y una eficiencia energética similar, pero de un material distinto a las neuronas, no se sabría con exactitud cómo programarlo; por

decirlo de alguna manera, sería el equivalente a una mente en blanco, sin ningún conocimiento. Como es bien sabido, las computadoras por muy potentes que sean, no funcionan si no cuentan con los programas adecuados. Y para decirlo pronto en la actualidad no se tiene ni la máquina ni el software necesario para que la IA fuerte sea una realidad. Como apunta Suay Belenguer (2012): "construir una mente mecánica es encontrar la vinculación entre el sustrato físico (neuronas, sinapsis, señales eléctricas) y el ámbito mental del pensamiento (ideas, emociones, inferencias, etc.). Esta fisura entre ambos dominios es todavía un problema abierto" (p. 167).

Esta ya clásica distinción entre hardware (máquina) y software (programas) ha llevado a González (2011), siguiendo los planteamientos de Searle (1980), a declarar que paradójicamente, más que un monismo materialista, en el proyecto de IA se encuentra una especie de dualismo cartesiano, en donde la máquina sería el equivalente del cuerpo, y el programa a la mente. Dada la preeminencia que Turing otorgaba al programa sobre la máquina, esto significa –de acuerdo a la interpretación de González– que así como Descartes afirmaba la separación entre la mente y el cuerpo, también podrían existir entidades inmateriales y artificiales con la capacidad de pensar: una vez desarrollado el programa que simule la mente humana, éste podría considerarse una entidad pensante, sin importar si existe o no alguna máquina que lo ejecute.

Sin embargo, la lectura de Turing que realiza González es sesgada e incorrecta. Si bien el matemático inglés afirma que el tipo de material utilizado para construir una máquina de IA es irrelevante, esto no significa que la máquina por sí misma también lo sea: ciertamente, no importa de qué material esté construida, pero es indispensable para ejecutar el programa y comprobar que funciona. La confusión de González pudiera deberse a que de manera frecuente en las investigaciones sobre IA se denomina máquina de Turing no a un artefacto, "sino un conjunto de instrucciones que conducen necesariamente a un resultado" (Piscoya Hermoza,

2017). Es decir, en algunas ocasiones se hace referencia únicamente al diseño de Turing y en otras a las máquinas físicas que funcionan con este diseño –que son prácticamente todas las computadoras actuales, exceptuando las cuánticas.

De cualquier forma, está claro que para concretar la inteligencia artificial tal como la entiende Turing (un comportamiento lingüístico que pueda ser denominado inteligente), son condición necesaria tanto el programa como la máquina capaz de ejecutarlo, así como el cerebro es una condición necesaria para el pensamiento. En ello coincidimos con Piscoya Hermoza en su artículo con el explícito título "Más allá del cartesianismo: la cultura como software mental y el cerebro como hardware genético" (2017), que propone la siguiente tesis con base en las ideas de Turing y Norbert Weiner:

> Lo que denominamos mente es una forma peculiar de organización de representaciones culturales que nos da sentido de identidad y que nos permite interactuar con el mundo comunicándonos y autorregulándonos de manera semejante al software de un ordenador. Tal organización, puede asumirse que se ha instalado en un órgano evolutivamente diseñado con especificidad, que llamamos cerebro, a través de vehículos bioquímicos y eléctricos de tal suerte que constituye nuestro hardware genético. (p. 52)

De igual manera, para que la IA sea una realidad son necesarios tanto el software como el hardware. Quizá podría realizarse una comprobación teórica para determinar si la ejecución de un programa tendría como resultado un comportamiento inteligente, algo similar a la demostración matemática de un teorema.[5] Pero esto no significaría que la inteligencia artificial exista en cuanto manifestación material en el mundo, tan sólo que se ha verificado la posibilidad de llevarla a cabo; en la práctica, Turing señala que es casi imposible saber cómo

[5] En realidad no hay una certeza de que tal comprobación matemática pueda realizarse. El *Entscheidungsproblem* (problema de decisión o parada) resuelto de manera independiente por Turing y Church, indica que es imposible saber en todos los casos si una máquina de Turing detendrá su ciclo de procesamiento, esto es, no es posible escribir un algoritmo general que resuelva todos los problemas.

funcionará exactamente un programa hasta que no se ejecuta sobre alguna máquina. Es evidente que tal programa, aún si no se puede concretar en algún hardware, ya es algo: es un diseño, una serie de instrucciones para guiar el comportamiento de una determinada máquina. Pero así como la palabra y la descripción de una mesa no equivalen a una mesa real, tampoco un programa o diseño equivalen a su ejecución, la cual se realiza necesariamente en una máquina, sin importar de qué material sea ésta.

Esta última afirmación merece una importante acotación: aunque por principio sea irrelevante el tipo de material –ya sean neuronas, transistores o engranes–, en la práctica debe tener ciertas propiedades que le permitan ser eficiente en la construcción de una máquina capaz de procesar información. Si quisiéramos alcanzar el exaflop de potencia que tiene un cerebro humano utilizando la famosa computadora ENIAC de 1946, necesitaríamos de 2x10^15 máquinas (un 2 seguido de 15 ceros) que ocuparían un área miles de veces más grande que nuestro planeta Tierra. Por eso el desarrollo de la computación está ligado al desarrollo de la electrónica: lo que permitió la construcción de ordenadores cada vez más potentes fue el uso de válvulas electrónicas de vacío, que posteriormente fueron desplazadas por los transistores de silicio, la base de la tecnología actual. La ENIAC ocupaba un espacio de 167 m^2 y contaba con cerca de 20 mil válvulas de vacío; un procesador de un teléfono móvil de última generación puede albergar más de 8 mil millones de transistores en un área de pocos centímetros.

Ahora bien, la búsqueda de nuevos materiales también revela una controvertida posibilidad futura: el uso de seres biológicos como plataforma para ejecutar un software o añadir hardware artificial. Por ejemplo, en 2013 los investigadores Greg Gage y Tim Marzullo crearon un circuito para controlar los movimientos de una cucaracha viva mediante un teléfono celular. Desde 2004, el artista de vanguardia Neil Harbisson lleva una antena integrada a su cráneo que le transmite vibraciones audibles, mediante la cual es capaz de percibir colores

invisibles como infrarrojos y ultravioletas, y recibir información de manera inalámbrica. La tecnología médica de vanguardia va más allá: los implantes neuronales permiten a los pacientes recuperar la audición y la vista, y estimular la recuperación de movimiento y sensibilidad en tetrapléjicos. Literalmente, son seres humanos con tecnología electrónica conectada directamente a su cerebro. Y considerando los avances en cultivos biológicos, lo único que impide el estudio de neuronas humanas para construir supercomputadoras biológicas es una cuestión de ética.

Desde nuestro punto de vista, lo que Turing expresa cuando afirma la irrelevancia de los materiales, es que las imposibilidades técnicas no impiden el desarrollo y definición de algoritmos (instrucciones) que permitan simular el comportamiento humano. O si se quiere, que las limitaciones técnicas no limitan el ingenio humano. El propio Turing fue uno de los matemáticos que en la década de 1940 sentó las bases de la programación de software, sin contar aún con alguna máquina capaz de ejecutar sus programas. Un siglo antes, en 1834 Charles Babbage concibió el diseñó de una máquina analítica que resolvería problemas matemáticos de todo tipo (anteriormente las únicas operaciones que se habían logrado resolver con máquinas eran polinomios, es decir, aquellas que pudieran expresarse mediante sumas y restas). Sin embargo no pudo construirse por falta de financiamiento, ya que se requerían miles de engranajes y mecanismos especiales, que cubrirían un área de 300 m^2, y para accionarla se necesitaría un motor a vapor de locomotora… y aun así, habría tenido menos potencia de cálculo que el microprocesador más básico de nuestros teléfonos celulares.

Así como Da Vinci diseñó una máquina voladora siglos antes de que existiera un motor capaz de proporcionar la fuerza necesaria para levantar el vuelo, así los expertos en software trabajan en desarrollar programas de IA adecuados para simular el funcionamiento del cerebro humano, aún si actualmente no pueden ser ejecutados por ninguna máquina. Es el caso del software NEST, "un simulador para

diseñar y probar modelos de redes neuronales que se centra en la dinámica, el tamaño y la estructura de los sistemas neurales, más que en la morfología exacta de las neuronas individuales" (NEST Simulator, 2019). Actualmente no hay ninguna computadora capaz de ejecutar estos modelos con la potencia y velocidad del cerebro humano. NEST fue precisamente el software que utilizó el superordenador Fujitsu K, con los resultados ya descritos: 40 minutos para simular un segundo de actividad de la red neuronal… y sólo estableciendo conexiones aleatorias. El nuevo supercomputador SpiNNaker es capaz de emular el funcionamiento de una red neuronal en tiempo real, pero representa tan sólo el 1% de las neuronas del cerebro humano. Aunque los expertos en software consideran las limitaciones técnicas como el principal obstáculo para alcanzar una IA comparable a la humana, también podría darse el caso contrario: que en el futuro se superen tales limitaciones, pero que aún no se hayan escrito los algoritmos completos que permitan imitar a la perfección el funcionamiento cerebral u obtener un resultado similar.

1.4 Materialismo dinámico y complejidad emergente

Como se habrá observado, en esta tesis nos apartamos de las explicaciones que consideran el alma o cualquier otra fuerza o elemento sobrenatural como la fuente del intelecto. Contemplar como posible la creación de inteligencia artificial nos exige un punto de partida materialista y mecanicista, esto es, que toda realidad se fundamenta en la energía y la materia y que los fenómenos de nuestra realidad no requieren la intervención de fuerzas sobrenaturales.

Sin embargo, reconocemos por principio que la realidad es inabarcable y la ciencia uno de los múltiples caminos que el ser humano ha construido para intentar comprenderla. De igual manera es evidente la existencia de fenómenos altamente complejos relacionados con la existencia humana, como el arte, la cultura y la sociedad, que requieren sus propias hipótesis y métodos de estudio, para los cuales este tipo de enfoque materialista puede resultar de escasa o nula utilidad.

En contraparte, debemos señalar que nuestra postura materialista no debe ser entendida como sinónimo de reduccionista (que las leyes físicas conocidas constituyen la explicación profunda de todo), ni entendemos por mecanicismo una forma de determinismo absoluto (que existe una cadena inexorable de causas y efectos), sino que apostamos por un materialismo dinámico que incluye los fenómenos de complejidad emergente y autoorganización, así como un mecanicismo no determinista, como se explicará a detalle.

Consideramos que una visión materialista es coherente con el concepto de complejidad emergente. Si bien las propiedades de un sistema no son reducibles a las propiedades de sus partes constituyentes, tampoco son ajenas a éstas: es de la unión de las partes de la que surgen comportamientos y procesos que sólo pueden ser explicados a partir del sistema como un todo. Tal concepto se relaciona estrechamente con el de autoorganización, esto es, que en determinados sistemas complejos se observa un modelo de orden o coordinación que surge de las interacciones locales entre componentes inicialmente desordenados.

Desde esta perspectiva, la complejidad no se refiere en primera instancia al funcionamiento intrincado o desconocido de un mecanismo, sino al número de interacciones posibles entre sus elementos. A mayor número de interacciones corresponde una mayor complejidad, y a mayor complejidad existe una mayor probabilidad de que surjan propiedades o comportamientos no previstos por un modelo que solo considere los elementos aislados.

Mediante los conceptos de complejidad emergente y autoorganización es posible entender el origen de la vida a partir de la conjunción de sustancias inanimadas, así como el origen de la mente a partir del sistema nervioso, sin reducir la vida a meras reacciones química y la mente a simples descargas eléctricas, ya que las propiedades del todo (la vida o la mente) no son equivalentes a las propiedades de las partes (las moléculas orgánicas o las neuronas). Ahora bien, el concepto de complejidad emergente no significa que exista un diseño predeterminado en la

naturaleza, ni un diseño inteligente plasmado de antemano. Por el contrario, la complejidad emergente indica que es imposible conocer o predecir por completo las propiedades de un sistema complejo mediante el solo estudio de sus elementos integrantes, porque tales propiedades no existen como una entidad *a priori*, sino sólo como resultado de la conjunción de diversos elementos.

Precisamente, una de las principales críticas a la postura emergentista es que si bien señala el contexto específico en el que se observan comportamientos emergentes (en sistemas complejos, esto es, con un número elevado de interacciones), no resuelve cómo o por qué surgen. Esta crítica nace de una visión determinista, desde la que todo efecto debe explicarse por una causa anterior cognoscible y determinada. Pero es precisamente en este vacío explicativo en donde se fundamenta su radical diferencia de perspectiva: mientras que el determinismo se basa en una secuencia lógica en la que todo efecto proviene de una causa, desde el emergentismo se considera que no todos los fenómenos son el resultado inevitable de unas leyes matemáticas operando sobre las condiciones iniciales del universo, sino que debe considerarse el efecto constructivo del caos, el azar y el tiempo, que complementan y modifican los resultado esperables por las leyes físicas clásicas.

Las propiedades emergentes de un sistema, distintas de las propiedades de sus elementos aislados y no determinadas completamente por éstas, y los procesos de autoorganización que brindan un orden superior en un sistema inicialmente caótico, podrían considerarse por lo tanto una acumulación visible de los efectos del azar actuando junto a las leyes físicas. La evolución del Universo, en particular los sistemas complejos como la vida, sería el resultado de la convergencia de múltiples accidentes o contingencias que dieron lugar a nuevas regularidades, y no solo el despliegue de las instrucciones predeterminadas un algoritmo físico-matemático.

Más aún, desde esta perspectiva nuestras limitaciones para predecir comportamientos complejos o propiedades emergentes no se deben a que desconozcamos alguna causa lógico-formal que los origina, sino a que tal causa no

existe o no es consistente con la lógica clásica. Las diferencias de orden cualitativo que suponemos que existen entre un ser vivo y uno inerte, o entre un organismo unicelular y uno pensante, se podrían explicar considerando una causa cuantitativa y una lógica probabilística no determinista: es el número elevado de elementos e interacciones en las que interviene el azar y las probabilidades lo que permite el desarrollo de nuevas propiedades y formas de organización. En otras palabras, las variaciones cuantitativas y el azar, sumado a las causas físicas conocidas, darían lugar a variaciones cualitativas.

Para decirlo sin ambages: el emergentismo es el reconocimiento de los efectos del azar y la incertidumbre del futuro, de cierta ignorancia irreductible, la certeza de que no hay certezas absolutas, la aceptación de una realidad que no se reduce a una lógica humana, una realidad sin porqué.

1.5 Mecanicismo no determinista

Si no se limitan claramente los límites del emergentismo, desde una visión laxa se podría argumentar que cualquier propiedad podría surgir de cualquier conjunto suficientemente complejo. Por ello es necesario acortarlo mediante una postura mecanicista, esto es, que son necesarios ciertos mecanismos y condiciones específicas en los que tales comportamientos emergentes pueden surgir. Para el surgimiento de lo que llamamos mente sería necesario un conjunto suficientemente grande de neuronas, organizadas en un sistema, y operando bajo un modelo de aprendizaje.

Pero aunque sean necesarios ciertas condiciones básicas, ningún mecanismo puede ser completamente determinista, ya que nada se encuentran exento de la intervención del azar y el caos, e incluso se podría afirmar que tales elementos son parte fundamental del funcionamiento del Universo. Al respecto, cabe hacer una pequeña aclaración: desde la física, el azar es definido como lo completamente aleatorio, inesperado e impredecible, mientras que el caos se refiere a sistemas

altamente sensibles a pequeñas variaciones, lo cual lleva a resultados divergentes a partir de condiciones iniciales casi idénticas. Ambos elementos son parte constitutiva de la naturaleza del universo.

Por ello, aunque las propiedades emergentes encuentran su origen en una condición previa y los eventos se relacionan de manera temporal, esto es, las condiciones actuales son resultado de condiciones anteriores, dicha relación no obedece estrictamente una lógica causa-efecto, sino que es necesario considerar una suma de contingencias en donde intervienen el azar y el caos, que por definición no tienen forma lógica; hablamos por tanto de una secuencia temporal, pero no lineal en sus causas ni progresiva en sus efectos. Con mecanicismo no determinista nos referimos a un minotauro, un imposible hecho realidad, híbrido entre la lógica y el azar, una serie de mecanismos y elementos de los que se espera cierto resultado, pero que están afectados por el azar y el caos, y que por lo tanto pueden generar una nueva complejidad emergente y comportamientos no predecibles.

Por el contrario, desde una visión tradicional el mecanicismo y el determinismo se consideraban sinónimos, visión impulsada desde el siglo XVII con el desarrollo del método científico por Francis Bacon y Descartes, y posteriormente en el siglo XIX con el positivismo de Augusto Comte. En el mecanicismo tradicional, se consideraba que la naturaleza funcionaba como una máquina, como los engranajes de un reloj. Para Descartes (1641) "no hay ninguna diferencia entre las máquinas que construyen los artesanos y los diversos cuerpos que compone la naturaleza ella sola". Cabe aclarar que en tal afirmación Descartes sólo se refiere a la sustancia extensa, al cuerpo material; de ahí que considere a los animales como máquinas. Pero la sustancia pensante, el intelecto que considera exclusivo del hombre, tendría su explicación en el alma. Estas dos sustancias serían de naturalezas completamente distintas, con un punto de unión a través de la glándula pineal. Con Descartes nos encontramos ante una concepción doble: mecanicismo para explicar la naturaleza, dualismo metafísico para explicar al hombre.

El mecanicismo determinista y toda la física clásica tienen su base en el principio de causalidad expuesto por el filósofo presocrático Leucipo en el siglo V a.C.: "Nada sucede por azar, sino que todo ocurre por una razón y por necesidad". Esta afirmación encontraría su mejor expresión matemática en las tres leyes de Newton enunciadas en sus *Principios matemáticos de la filosofía natural* (1687). La trascendencia y exactitud de la teoría newtoniana y de la ley de la gravitación universal es tal, que a pesar del descubrimiento de la mecánica cuántica continúa siendo fundamental para el estudio de la física y para explicar el movimiento de la mayoría de los cuerpos celestes.

Desde los postulados deterministas de la física clásica, se asumía que todos los eventos están causados por otros anteriores y que esta causalidad es expresable en términos matemáticos. De ahí la extrema confianza de Pierre-Simon Laplace (1814), matemático y astrónomo francés que llegó a afirmar: "Hemos de considerar el estado actual del universo como el efecto de su estado anterior y como la causa del que ha de seguirle [...] si existiera una entidad capaz de conocer en un instante determinado la posición y velocidad de todas las partículas del universo, conocería todo el futuro y todo el pasado".

Sin embargo, con el desarrollo de la mecánica cuántica se fue diluyendo la certeza de alcanzar el conocimiento absoluto. El principio de incertidumbre de Heisenberg (1926) establece la imposibilidad de que determinados pares de magnitudes físicas observables y complementarias, como la posición y el momento lineal de una partícula, sean conocidas simultáneamente y con precisión absoluta. Este principio supone un cambio esencial en nuestra concepción de la naturaleza del Universo, ya que se abandona toda posibilidad de alcanzar un conocimiento absolutamente preciso, y se admite en su lugar la naturaleza probabilística del Universo.

Curiosamente, la consideración de este campo de probabilidades es lo que ha permitido a la mecánica cuántica ofrecer una descripción más precisa del

comportamiento de las partículas subatómicas y otros fenómenos como los agujeros negros. Como explica Romero Rochín (2015), para la mecánica cuántica no existe el concepto de causalidad tal como lo entiende la física clásica. En la mecánica cuántica,

> […] si conocemos una causa, con el máximo posible conocimiento de su realización en un momento dado, existe en general un número indefinido de posibles efectos de los cuales no podemos saber cuál ocurrirá. No sólo eso, nos dice rigurosamente que tal incertidumbre no es un problema de nuestra falta o incapacidad de conocimiento o de experimentación. Nos dice, contundentemente, que la naturaleza así es y que no hay nada qué hacer. (p. 3)

En cuanto a la inteligencia artificial, si partimos de una postura estrictamente determinista en la que a toda causa corresponde un efecto único y bien delimitado, ello implicaría que la inteligencia humana sólo puede surgir de una causa única y bien determinada, en este caso el cerebro humano. Por tanto en esta tesis nos decantamos por un mecanicismo no determinista, una apertura de posibilidades desde la que consideramos que así como una sola causa puede originar diversos efectos, también es posible alcanzar el nivel de la inteligencia humana a partir de mecanismos diferentes, ya sean biológicos o artificiales.

Lo que se encuentra en juego es una visión más amplia de la realidad. Al contrario de lo que sostendría el determinismo, la naturaleza no "obedece" leyes físicas y matemáticas; son nuestras teorías las que se ajustan o no al comportamiento observado en el Universo. Es nuestra interpretación matemática la que se debe ajustar a la naturaleza, incluso si ello implica renunciar a la idea de un orden absoluto y geométrico, y reivindicar el papel preponderante del azar y caos. En la medida en que este modelo es más cercano a los fenómenos que observamos, la teoría no sólo permite predecir los comportamientos que se observan en la naturaleza, sino también modelarlos y recrearlos. Las matemáticas no son el lenguaje de la naturaleza, sino la interpretación que la ciencia propone para entenderla. Lo que está a prueba con la creación de la inteligencia artificial es si tal

entendimiento es lo suficientemente profundo para replicar por medios y herramientas tecnológicas el comportamiento de los organismos vivos, incluyendo aquellos fenómenos que no comprendemos por completo, como la mente y la consciencia.

Capítulo 2

Inteligencia artificial, ¿el principio del fin?

Como hemos visto en el capítulo anterior, aún no existen las condiciones técnicas para alcanzar a corto plazo la inteligencia artificial general o fuerte, pero es un proyecto que se encuentra en curso, y todo apunta a que en algún momento alcanzará su meta. En este capítulo abordaremos cuáles son las perspectivas en este probable futuro.

En general, las posturas sobre la inteligencia artificial y sus consecuencias en nuestra sociedad se pueden dividir en dos polos contrapuestos: la IA como la próxima gran revolución tecnológica que permitirá al hombre trascender sus límites mediante la creación de nuevos seres pensantes que nos ayudarán a resolver los problemas que enfrentamos como humanidad; o bien, la IA como la culminación de un proyecto científico-tecnológico desbordado, una nueva era en la que los humanos seremos suplantando por las máquinas, perspectiva que abordaremos con mayor profundidad. Con algunos matices, estas dos posturas son representadas respectivamente por el filósofo y padre de la cibernética Norbert Wiener (1894-1964), y el filósofo y sociólogo francés Jean Baudrillard (1929-2007), quienes nos servirán de guía en esta prospectiva, junto con otros autores complementarios.

2.1 Tecnología, ilusión y destrucción

En su libro *El crimen perfecto* (1995) Baudrillard resalta los efectos nocivos del desarrollo acelerado de la tecnología, incluyendo la digitalización y la inteligencia artificial. Las tecnologías actuales son la consecuencia de un proceso histórico-social mucho más amplio, que ancla sus raíces en un tiempo mítico marcado por el surgimiento del lenguaje. De acuerdo con el filósofo francés:

> Si no existieran las apariencias, el mundo sería un crimen perfecto, es decir, sin criminal, sin víctima y sin móvil. […] Pero, precisamente, el crimen nunca es perfecto, pues el mundo se traiciona por las apariencias, que son las huellas de su inexistencia, las huellas de la continuidad de la nada, ya que la propia nada, la continuidad de la nada, deja huellas. Y así es como el mundo traiciona su secreto. Así es como se deja presentir, ocultándose detrás de las apariencias. (1995, p. 11)

Acompañemos a Baudrillard siguiendo las huellas de este crimen, los huecos y marcas que va dejando la modernidad. La clave es el estudio del mundo contemporáneo a partir de la desaparición de algo y la aparición de la nada, que como veremos no son procesos equivalentes por completo. Un análisis por lo tanto no sólo de las nuevas maneras en que habitamos nuestra época, sino también de lo que se ha sustraído a ella y las formas de sustraerlo: "Esto es la historia de un crimen, del asesinato de la realidad. Y del exterminio de una ilusión, la ilusión vital, la ilusión radical del mundo. Lo real no desaparece en la ilusión, es la ilusión la que desaparece en la realidad integral" (Baudrillard, 1995, p. 9). En los siguientes párrafos habremos de ser cautos con el concepto de realidad, ya que Baudrillard se desplaza entre cuatro significados diferentes: 1) lo real como universo previo al hombre; 2) el mundo creado por el lenguaje; 3) la realidad como un conjunto objetivo de leyes planteadas por la ciencia; y 4) la hiperrealidad como una realidad técnica que se acrecienta a sí misma y desplaza al ser humano.

De cualquier forma, el hilo conductor del análisis de Baudrillard es claro y podría resumirse de la siguiente manera: a contracorriente de la perspectiva ilustrada, que propone combatir las sombras de la ilusión en aras de revelar una verdad luminosa y lógica, para Baudrillard el mundo humano es una ilusión vital necesaria, y la destrucción de esta ilusión por parte de la ciencia y la tecnología no es necesariamente benéfica. El crimen que señala Baudrillard es doble: en primer lugar, un crimen mítico, la sustitución de un mundo real por un mundo de apariencias inaugurado por el lenguaje; y en un segundo momento el crimen propio

de nuestra época, el exterminio de la ilusión vital, un crimen del cual son cómplices los desarrollos técnicos actuales como la inteligencia artificial. En cuanto crimen que pertenece a tiempos míticos, el asesinato de la realidad es un acto fértil que permite la constitución del mundo humano: "es la energía de este crimen, como la del estallido final, la que se distribuirá por el mundo, hasta su eventual agotamiento. Ésta es la visión mítica del crimen original, la de la alteración del mundo en el juego de la seducción y las apariencias, y de su ilusión definitiva" (Baudrillard, 1995, p. 12)

Este primer crimen, el que da origen al mundo humano, es la sustitución de la realidad por el signo, ya sea lingüístico (acústico y gráfico) o visual (imagen, letra y símbolo). Si consideramos que lo esencial de este proceso es la sustitución de un elemento por otro diferente, también podríamos afirmar que el mundo humano inicia con la invención de la técnica, al sustituir el cuerpo (la mano) por el objeto (la herramienta), como en la famosa escena del filme de Kubrick, *2001: Odisea al espacio* (1968), en donde los homínidos descubren el uso de un hueso como artefacto. Otro parteaguas es el desarrollo la agricultura, el momento en que dejamos de ser animales nómadas pertenecientes al mundo, y pasamos a transformarlo y adueñarnos de él. En esta misma línea se puede considerar el descubrimiento y manejo del fuego, un elemento divino que nos separa de los animales y nos equipara con los dioses, de acuerdo al mito de Prometeo. En estos ejemplos, se revela que la técnica no es simplemente "ciencia aplicada", sino que la técnica –comenzando por el propio lenguaje– inaugura un modo nuevo de conocimiento y de habitar el mundo, entendido no como un real absoluto, sino como construcción humana.

En todos estos casos, al transformar un objeto en otra cosa, no sólo creamos en este una segunda naturaleza útil-funcional, sino también una nueva naturaleza simbólica. Una vara puede servir como arma y herramienta, pero también como cetro, representando el poder. Así, la pregunta de fondo no es para qué sirve la

inteligencia artificial, sino qué simboliza su desarrollo como presunta cima de toda la evolución técnica.

Para contestarla, es necesario retroceder con Baudrillard hasta tiempos míticos. La constitución del mundo humano es el primer crimen que, en su intento de hacerlo accesible, oculta el mundo inmanente, aquel en donde "todo animal está en el mundo como agua dentro del agua" como expresa Bataille en *Teoría de la religión* (1973). Este primer crimen es necesario para fundar el universo simbólico, ya que el mundo nunca se nos presenta en su realidad (entendida como la cosa-en-sí kantiana), sino sólo como ilusión. Porque sólo en la ilusión podemos vivir: "La ilusión radical es la del crimen original, por el cual el mundo es alterado desde el inicio, jamás idéntico a sí mismo, jamás real. El mundo sólo existe gracias a esta ilusión definitiva que es la del juego de las apariencias, el lugar mismo de la desaparición incesante de cualquier significación y de cualquier finalidad". (Baudrillard, 1995, p. 20) En las palabras de Baudrillard resuena un eco de las ideas de Georges Bataille (1973):

> Nada, a decir verdad, nos está más cerrado que esa vida animal de la que hemos salido. Nada es más extraño a nuestra manera de pensar que la tierra en el seno del universo silencioso y no teniendo ni el sentido que el hombre da a las cosas, ni el sinsentido de las cosas en el momento en que quisiéramos imaginarlas sin una conciencia que las reflejase. En verdad, nunca podemos más que arbitrariamente figurarnos las cosas sin la conciencia, puesto que *nosotros*, *figurarse*, implican la conciencia, nuestra conciencia, adhiriéndose de una manera indeleble a su presencia. (p. 24)

Sin embargo, no debemos caer en el error de considerar que tal ilusión vital aniquila el imperio de lo real; como señala Baudrillard "lo real no desaparece en la ilusión". La ilusión vital no anula el mundo inmanente, ni tampoco lo vuelve transparente, sino que es la manera en que lo habitamos. La ilusión sería constitutiva del mundo humano, la piel y no sólo máscara de la realidad, nuestro mundo interpretado, como lo llama Rilke en las *Elegías de Duino* (1923): "E incluso las bestias, astutas, se percatan de que es torpe, inseguro, nuestro paso que yerra por un mundo interpretado". La

ilusión nos permite vivir sobre un mundo que suponemos anterior y real, a costa de un desfase, una fisura y una angustia que se verá acrecentada drásticamente en nuestros tiempos.

De ahí que paradójicamente, al tiempo que nos hace habitable el mundo, la creación de esta ilusión también despierta en nosotros una inquietud por acercarnos a las cosas en su "esencia", a ese mundo puro que creemos haber dejado atrás: "el hecho de que todo se esconda detrás de su propia apariencia y que, por tanto, no sea jamás idéntico a sí mismo, es la ilusión material del mundo. Y ésta sigue siendo, en el fondo, el gran enigma, el que nos sume en el terror y del que nos protegemos con la ilusión formal de la verdad" (Baudrillard, 1995, p. 13). Pero de acuerdo con el filósofo francés la alternativa no es un mejor camino:

> Nos movemos entre una ilusión y una verdad a cuál más insoportables. Pero ¿es posible que la verdad sea aún más insoportable, y deseemos finalmente la ilusión del mundo, aunque nos alcemos contra ella con todas las armas de la verdad, la ciencia y la metafísica?
>
> […] Que no podamos soportar su ilusión ni su apariencia pura forma parte del mundo. Tampoco soportaríamos mejor, si tuviera que existir, su verdad radical y su transparencia. (Baudrillard, 1995, pp. 21, 13)

Este radical alejamiento de lo real crea una zona vedada para siempre, un enigma irresoluble, un universo no asimilable por la razón, porque es anterior a ésta. Lo sagrado como la plenitud de lo real no sólo resulta inaccesible al común de los mortales, sino que su experiencia puede resultar fatal, como nos enseña Rilke (1923):

> Y si un ángel inopinadamente me ciñera contra su corazón,
> la fuerza de su ser me borraría;
> porque la belleza no es sino el nacimiento de lo terrible:
> un algo que nosotros podemos admirar y soportar
> tan sólo en la medida en que se aviene, desdeñoso, a existir, sin destruirnos.
> Todo ángel es terrible.

Precisamente la manera que la modernidad propone para cerrar esta brecha, para reparar esta fisura entre nuestro mundo ilusorio y el mundo "verdadero" es el uso

(o abuso) de la ciencia y la técnica, hijas de la razón, que parten del presupuesto de que para ser verdad el mundo debe ser objetivo y lógico. Pero la ciencia no agota el sentido y el saber de la experiencia vital humana. En *La condición posmoderna* (1979), Lyotard sostiene que "el saber científico es una clase de discurso", el dominante en la modernidad enfrentado a otros saberes como el mito o el arte: "En principio, el saber científico no es todo el saber, siempre ha estado en excedencia, en competencia, en conflicto con otro tipo de saber, que para simplificar llamaremos narrativo" (p. 22). Para Lyotard (1979) la ciencia es un discurso del saber entre muchos otros posibles, con la notable diferencia de caracterizarse por ser excluyente, al admitir sólo una verdad posible y negar las otras formas de saber narrativo:

> El científico se interroga sobre la validez de los enunciados narrativos y constata que estos nunca están sometidos a la argumentación y a la prueba. Los clasifica en otra mentalidad: salvaje, primitiva, subdesarrollada, atrasada, alienada, formada por opiniones, costumbres, autoridad, prejuicios, ignorancias, ideologías. Los relatos son fábulas, mitos, leyendas, buenas para las mujeres y los niños. En el mejor de los casos, se intentará que la luz penetre en ese oscurantismo, civilizar, educar, desarrollar. (p. 55)

La postura científica suele olvidar que el mito y la ilusión mantienen una función vital en las sociedades humanas no por ser verdaderos, sino por su potencia simbólica, elemento básico para otorgar unidad y sentido a cualquier comunidad humana. Por su parte, la meta de la ciencia y tecnología no apunta en primer término a la esfera vital –por mucho que en el discurso político se insista en sus "beneficios sociales"–, sino a la optimización y eficiencia de los procesos en sí mismos. Con su pensamiento objetivado y objetivador, la ciencia se erige como el garante de lo que debe ser considerado real y como la única forma de pensamiento válida. Pero a su vez, necesita legitimarse a sí misma: "En su origen, la ciencia está en conflicto con los relatos. Medidos por sus propios criterios, la mayor parte de los relatos se revelan fábulas. Pero, en tanto que la ciencia no se reduce a enunciar regularidades útiles y busca lo verdadero, debe legitimar sus reglas de juego. Es entonces cuando mantiene

sobre su propio estatuto un discurso de legitimación, y se la llama filosofía" (Lyotard, 1979, p. 9).

La ciencia busca legitimarse a sí misma de manera recursiva, esto es, recurriendo a las mismas ideas de razón y progreso que ella sostiene, como los cuatro elefantes que sostienen al mundo encima de una tortuga, que es a su vez sostenida por una serpiente que se muerde la cola... sosteniéndose a sí misma. La ciencia moderna –o más específicamente, el discurso oficial sobre la ciencia– ha olvidado que también es un mito, la ilusión de haber alcanzado la realidad objetiva. Se olvida que ya Gödel en 1931, con sus teoremas de la incompletitud, ha demostrado que un sistema lógico matemático, como pretende serlo gran parte de la ciencia, no puede ser completo y consistente a la vez. En otras palabras, un sistema lógico formal no puede explicarlo absolutamente todo y estar libre de contradicciones. En realidad, los sistemas lógicos y matemáticos dependen de axiomas, proposiciones que parecen completamente evidentes y sirven de base para el desarrollo de una teoría, pero que en sí mismas son indemostrables, como también lo son el motor inmóvil de Aristóteles y el yo cartesiano. En cierto sentido, los axiomas son similares a los dogmas religiosos, con la notable diferencia de que en la ciencia debería ser posible partir de nuevos axiomas.

En el caso del discurso sobre la ciencia moderna, podríamos decir que su axioma –convertido en dogma de la civilización– es que el progreso técnico conlleva necesariamente el progreso humano. Tal suposición nos brinda un punto de referencia hacia dónde avanzar, unas coordenadas de sentido y un centro en el cual refugiarnos: mientras la ciencia y técnica avancen, nuestros problemas serán finalmente solucionados y todo estará bien. Por supuesto, la historia reciente se ha encargado de demostrar que tal proposición es falsa.

A esta ecuación fallida debemos agregar un concepto central en el capitalismo: el *profit* o rentabilidad financiera, esto es, obtener más ganancias con menos inversión, una manera elegante de pronunciar codicia. En última instancia,

los beneficios que se esperan de la ciencia y la tecnología no son beneficios sociales, sino económicos. Sólo así es posible explicar que las tecnologías como las energías renovables, única vía ante el cambio climático, o los avances médicos que podrían salvar millones de vidas en países en desarrollo, no sean ampliamente utilizados por ser considerados "poco rentables".

2.2 La posmodernidad que no fue

Tal desencanto de la modernidad y la crisis de los relatos narrativos es lo que propone Lyotard (1979) como núcleo de la posmodernidad:

> Simplificando al máximo, se tiene por «postmoderna» la incredulidad con respecto a los metarrelatos. […] La función narrativa pierde sus functores, el gran héroe, los grandes peligros, y el gran propósito. Se dispersa en nubes de elementos lingüísticos narrativos, cada uno de ellos vehiculando consigo valencias pragmáticas *sui generis*.
>
> Los *decididores* intentan, sin embargo, adecuar esas nubes […] a una lógica que implica la conmensurabilidad de los elementos y la determinabilidad del todo. […] Esta lógica del más eficaz es, sin duda, inconsistente a muchas consideraciones […] ¿Dónde puede residir la legitimación después de los metarrelatos? El criterio de operatividad es tecnológico, no es pertinente para juzgar lo verdadero y lo justo. (p. 10-11)

Para la crítica posmoderna es claro que el criterio científico-técnico no es pertinente ni consistente como parámetro de la vida humana, pero esto no significa que no se continúe aplicando a mansalva en la sociedad contemporánea. Es verdad que surgen miles de movimientos reivindicando los derechos de las minorías, microfísicas del poder que actúan de manera localizada, pequeños espacios y momentos en los que recuperamos la ilusión de una vida auténtica y digna. Pero los dos grandes relatos que sustentan nuestra época, el capitalismo como único modo de producción e intercambio, y el progreso de la ciencia y la tecnología como condición benéfica y necesaria, permanecen vigentes. En contraste, Bruno Latour propone que *Nunca fuimos modernos* (1991):

> Con el adjetivo moderno se designa un régimen nuevo, una aceleración, una ruptura, una revolución del tiempo. Cuando las palabras «moderno», «modernización», «modernidad» aparecen, definimos por contraste un pasado arcaico y estable. Además, la palabra siempre resulta proferida en el curso de una polémica, en una pelea donde hay ganadores y perdedores, Antiguos y Modernos. Moderno, por lo tanto, es asimétrico dos veces: designa un quiebre en el pasaje regular del tiempo, y un combate en el que hay vencedores y vencidos. Si hoy en día tantos contemporáneos vacilan en emplear ese adjetivo, si lo calificamos mediante preposiciones, es porque no nos sentimos tan seguros de mantener esa doble asimetría: ya no podemos designar la flecha irreversible del tiempo ni atribuir un premio a los vencedores […] De ahí proviene el escepticismo llamado curiosamente posmoderno, aunque no sepa si es capaz de reemplazar para siempre a los modernos. (pp. 27-28)

Aunque compartimos la incertidumbre de Latour, en esta tesis consideramos que nunca hemos sido posmodernos, y que sí existe un vencedor: la historia reciente parece indicar que el proceso de modernidad continúa firme en una misma dirección, irreversible. El premio, por supuesto, es económico y queda reservado para las élites, mientras a los demás nos queda el consuelo de la crítica y la denuncia. O quizá la posmodernidad es un lujo que sólo pueden darse los países de primer mundo con un elevado desarrollo industrial, desencantados de las promesas de la técnica, mientras que el resto de las naciones seguimos embelesados con el sueño del progreso, en una modernidad tardía. En todo caso, al proceso de la modernidad no parece importarle mucho las críticas de los intelectuales, ni el desencanto de millones de seres humanos en situación de pobreza extrema y al borde de la muerte. Como Lyotard (1979) destaca, lo único que le importa al sistema es perfeccionar su funcionamiento: "La armonía de las necesidades y las esperanzas de individuos o grupos con las funciones que aseguran el sistema sólo es un componente adjunto de su funcionamiento; la verdadera fiabilidad del sistema, eso para lo que él mismo se programa como una máquina inteligente, es la optimización global de sus *inputs* con sus *outputs*, es decir, su performatividad" (p. 30).

Como bien señala Lyotard, en la modernidad el sistema trabaja para alcanzar su propio perfeccionamiento, ya sea hacer coches o exterminar seres humanos. La denuncia no es falaz; entre las grandes potencias económicas, una parte significativa del presupuesto destinado a ciencia y tecnología, incluyendo el desarrollo de la IA, se encuentra catalogado en el ámbito de la investigación militar: cómo matarnos mejor y más rápido. Bruno Latour (1991) lo explica así:

> Al querer desviar la explotación del hombre por el hombre a una explotación de la naturaleza por el hombre, el capitalismo multiplicó indefinidamente ambas. Lo reprimido retorna, y lo hace por partida doble: las multitudes que querían salvarse de la muerte vuelven a caer por centenares de millones en la miseria; las naturalezas, a las que se quería dominar por completo, nos dominan de manera también global amenazándonos a todos. Extraña dialéctica, que hace del esclavo dominado el amo y poseedor del hombre, y nos enseña de pronto que inventamos a los ecocidas al mismo tiempo que las hambrunas a gran escala. (p. 25)

Tal vez un error de la crítica posmoderna sea subestimar la inercia del sistema moderno: su masa es tal que continuará en movimiento mucho tiempo después de que las ideas que la impulsan (ciencia y capitalismo) hayan sido desacreditadas. El sistema es tan enorme que es capaz de reinventarse y absorber cualquier idea contraria, e incluso se asemeja a la energía oscura, pues en lugar de reducir su velocidad acelera su expansión hasta desgarrarlo todo: hipermodernidad, y no posmodernidad.

De acuerdo a la tradición judeocristiana –que representa junto a la cultura grecolatina una de las raíces de nuestra civilización occidental moderna– el trabajo fue el primer castigo impuesto por Yahvé:

> Al hombre le dijo: «Por haber escuchado la voz de tu mujer y comido del árbol del que yo te había prohibido comer, maldito sea el suelo por tu causa: sacarás de él el alimento con fatiga todos los días de tu vida. Te producirá espinas y abrojos, y comerás la hierba del campo. Comerás el pan con el sudor de tu rostro, hasta que vuelvas al suelo, pues de él fuiste tomado. Porque eres polvo y al polvo tornarás.» (Gn. 3: 17-19, ed. Desclée De Brouwer)

La promesa de la técnica y el progreso era ofrecernos una vida libre de trabajo y regresarnos al paraíso terrenal, pero sólo cumplió la mitad: como a los tejedores ingleses del siglo XIX, el progreso ciertamente nos liberó del trabajo, pero no nos ofreció ningún paraíso. A pesar del discurso de los grandes beneficios que nos han traído la ciencia y la tecnología, en ningún momento se considera la posibilidad de que estos sean gratuitos: dentro de un sistema regido por el capitalismo, el progreso queda reservado para quienes puedan comprarlo, lo cual es imposible sin un trabajo remunerado, un trabajo que desaparece gracias al progreso.

Por eso, en una sociedad marcada fuertemente por la meritocracia, que rechaza la generosidad espontánea y exige que todo sea ganado a base de esfuerzo, el despojarnos del trabajo parece más una segunda maldición: arrancados del paraíso, ahora somos arrancados de la propia tierra. El trabajo nos transformó en esclavos, pero progreso no nos ha convertido en seres libres que se dedican a actividades espirituales y artísticas, sino en desempleados, ciudadanos de segunda clase, excluidos y pobres. Como Chaplin en *Tiempos modernos* (1936), el hombre sólo es un engranaje del mecanismo económico y social, y cada vez se revela más innecesario para su funcionamiento.

Con estos elementos podemos realizar un primer intento de respuesta a la pregunta antes planteada: qué simboliza el desarrollo de la IA como presunta cima de toda la evolución técnica. Consideramos que el desarrollo de la inteligencia artificial es la prueba de que el proyecto moderno sigue vigente; la ciencia es un camino que sólo puede ir hacia adelante, aún si se dirige al abismo. Si la máquina despojó al obrero del único bien que le quedaba, su fuerza de trabajo, la inteligencia artificial terminará por desplazar cualquier rastro de humanidad que necesite el sistema para funcionar.

Redimidos del trabajo, el siguiente paso es la inteligencia artificial que nos librará de la fatigosa tarea de pensar. Esto ya es un hecho en los avanzados

algoritmos de búsqueda que utilizan las grandes compañías de internet. No hace falta conocer los términos exactos, basta con tener una idea vaga e introducir los primeros caracteres para que el sistema nos arroje al instante millones de resultados en los que encontraremos aquello que no supimos pensar con claridad. El problema reside en que tal información es proporcionada por una empresa privada, y por lo tanto es ella la que moldea nuestra visión del mundo. Además de este peligroso monopolio como puerta de acceso a la información, otros dos problemas asociados a las recomendaciones automáticas son las burbujas de filtro y cámaras de eco: al aprender de aquellas páginas que elegimos, el algoritmo de búsqueda comienza a mostrarnos sólo la información que coinciden con nuestros intereses y prejuicios, descartando la posibilidad de acceder a posturas divergentes a la nuestra.

Las consecuencias de estos procesos automáticos son una pérdida evidente de cualidades o capacidades humanas: con un universo de información en la palma de la mano y a un click de distancia, no es necesario guardar en la memoria ni siquiera los números de teléfono de nuestros seres más cercanos; quizá en el futuro no hará falta saber ni siquiera sus nombres. Esta denuncia no es nueva; ya Platón en el *Fedro* la advierte respecto a la invención de la escritura, en la respuesta del rey Thamus al dios egipcio Theuth:

> ¡Oh artificiosísimo Theuth! A unos les es dado crear arte, a otros juzgar qué de daño o provecho aporta a los que pretenden hacer uso de él. Y ahora tú, precisamente, padre que eres de las letras, por apego a ellas, les atribuyes poderes contrarios a los que tienen. Porque es obvio lo que producirán en las almas de quienes las aprendan, al descuidar la memoria, ya que, fiándose de lo escrito, llegarán al recuerdo desde fuera, a través de caracteres ajenos, no desde dentro, desde ellos mismos y por sí mismos. No es, pues, un fármaco de la memoria lo que has hallado, sino un simple recordatorio. Apariencia de sabiduría es lo que proporcionas a tus alumnos, que no verdad. Porque habiendo oído muchas cosas sin aprenderlas, parecerá que tienen muchos conocimientos, siendo, al contrario, en la mayoría de los casos, totalmente ignorantes, y difíciles, además, de tratar porque han acabado por convertirse en sabios aparentes en lugar de sabios de verdad. (p. 399)

Respecto a estas críticas de corte pesimista, y señalando de manera específica a Baudrillard, Lyotard (1979) señala:

> De esta descomposición de los grandes relatos, se sigue eso que algunos analizan como la disolución del lazo social y el paso de las colectividades sociales al estado de una masa compuesta de átomos individuales lanzados a un absurdo movimiento browniano [aleatorio]. Lo que no es más que una visión obnubilada por la representación paradisíaca de una sociedad «orgánica» perdida. (p. 36)

En defensa de Baudrillard, y sin caer en el lugar común de que todo tiempo pasado fue mejor, habría que recordar que el hombre como especie ha habitado la Tierra por más de 300 mil años, y tan sólo en los últimos 200 años a partir de la Revolución Industrial hemos tenido un mayor impacto negativo sobre el planeta que en todos los milenios anteriores, mientras la brecha entre las élites económicas y las clases populares se ha vuelto cada vez más profunda. Aunque la técnica no sea exclusiva de la época moderna es claro que el refinamiento de la máquina representa un punto de quiebre en nuestra relación con la naturaleza y entre nosotros mismos. Desde esta perspectiva, el ludismo que pretendía destruir a las máquinas y el anarco-primitivismo que reivindica el vivir nómada no parecen tan descabellados. En cuanto a los últimos avances técnicos, Lyotard (1979) advierte:

> Es razonable pensar que la multiplicación de las máquinas de información afecta y afectará a la circulación de los conocimientos tanto como lo ha hecho el desarrollo de los medios de circulación de hombres primero (transporte), de sonidos e imágenes después (media). En esta transformación general, la naturaleza del saber no quedará intacta. No puede pasar por los nuevos canales y convertirse en operativa, a no ser que el conocimiento pueda ser traducido en cantidades de información. Se puede, pues, establecer la previsión de que todo lo que en saber constituido no es traducible de ese modo será dejado de lado. (p. 15)

En pocas palabras, sólo aquel saber que cumpla una lógica matemática, capaz de expresarse en bits, será considerado como nuestra realidad, dejando fuera todo

discurso sobre el sentir, los afectos, la experiencia estética, lo espiritual y lo inefable. Esta búsqueda científica, formal, casi obsesiva de la verdad, será la que termine por desplazar la ilusión vital, poniendo en jaque no sólo las construcciones simbólicas humanas, sino a la especie misma.

Como Lyotard con el concepto de discurso dominante, lugar ocupado en nuestro mundo contemporáneo por el capitalismo y el saber científico, Baudrillard también revela que este intento de erigir un pilar central para dar un sentido lógico al mundo (a partir del yo cartesiano, la razón y la ciencia) no deja de ser una nueva ilusión, que en lugar de quitar el velo del mundo lo cubre con otro manto aún más denso: "Todo lo que se proyecta más allá de esta ilusión [vital], de esta evidencia accidental del mundo, que lo aleja para siempre de su sentido y de su origen, no es más que una fantasía justificativa" (Baudrillard, 1995, p. 21). Tal es el crimen de la modernidad, la construcción de una realidad a la medida, de crecimiento acelerado y automático, que aniquila cualquier otro proyecto de sentido o de pensamiento alterno, una realidad tan densa que Baudrillard (1995) denomina hiperrealidad.

> Vivimos en la ilusión de que lo real es lo que más falta, cuando ocurre lo contrario: la realidad ha llegado a su colmo. A fuerza de proezas técnicas, hemos alcanzado tal grado de realidad y de objetividad que podemos hablar incluso de un exceso de realidad que nos deja mucho más ansiosos y desconcertados que el defecto de realidad, que por lo menos podíamos compensar con la utopía y lo imaginario, mientras que para el exceso de realidad no existe compensación ni alternativa. No existe negación ni superación posibles, ya que estamos más allá.
>
> […] Toda la modernidad ha tenido por objetivo el advenimiento de este mundo real, la liberación de los hombres y de las energías reales, enfocadas hacia una transformación objetiva del mundo, más allá de todas las ilusiones cuyo análisis crítico ha alimentado la filosofía y la práctica. Hoy, el mundo es más real de lo que esperábamos. (pp. 91-92)

La crítica de Baudrillard respecto a los avances técnicos nos permite conectar sus ideas con las consecuencias del desarrollo de la IA. Para Baudrillard (1995) el panorama es claro:

> Estamos en plena ilusión de la finalidad de la técnica como extensión del hombre y de su poder, en plena ilusión subjetiva de la técnica. Pero hoy este principio operativo es derrotado por su misma extensión, por esta virtualidad sin freno, que supera las leyes de la física y la metafísica.
>
> Así todas nuestras tecnologías sólo serían el instrumento de un mundo que creemos dominar, cuando él es el que se impone a través de un equipo del que sólo somos meros operadores. […] A través de la técnica, tal vez sea el mundo el que se ríe de nosotros, el objeto que nos seduce con la ilusión del poder que tenemos sobre él. (pp. 100-101)

El hombre ya no es fin en sí mismo, sino sólo un eslabón en la cadena de producción, un paso en el camino hacia un mundo dominado por los objetos. Giovanni Papini lo vislumbra en *Gog* (1931), en su ficticia "Visita a Ford". Al explicar su método de producción, el empresario estadounidense le revela:

> El asunto no es desarrollar una industria, sino llevar a cabo un gran experimento intelectual y político. […] el ideal supremo es el siguiente: *Fabricar sin ningún operario una cantidad cada vez mayor de artículos que cuesten lo menos posible*. Soy un utopista, pero no un loco y debo admitir que aún pasarán algunas decenas de años antes de que se llegue a mi ideal. Sin embargo, me preparo para ese día y construyo en Detroit una nueva fábrica que se llama *La Solitaria*. Una verdadera joya, un sueño, un milagro, en pocas palabras, la fábrica donde no habrá nadie. Al ser terminada y cuando haya sido equipada con las máquinas del más reciente modelo, […] no se necesitarán obreros. De vez en cuando, un ingeniero las visitará brevemente, echará a andar algunos engranajes y se marchará. Del resto se encargarán las máquinas por sí solas que trabajarán no únicamente por el día, como lo hacen ahora los hombres, sino también toda la noche e incluso los domingos. […]
>
> Soy el místico desinteresado de la producción y la venta. […] Mi ambición, científica y humanitaria, es la religión del movimiento sin fin de la producción ilimitada, de la máquina libertadora y dominadora. (pp. 33-36)

Sorprende la predicción de Papini por su exactitud y profundidad. En términos prácticos, ya existe esta nueva realidad técnica que no requiere de operadores humanos. Se llama Industria 4.0, y consiste en la automatización e interconexión de todos los procesos productivos: desde la extracción de los materiales hasta la venta

del producto final, todo es realizado por una extensa red de máquinas. Dicha tendencia comienza a extenderse no sólo a la manufactura, sino a todas las profesiones, desde la medicina hasta la educación. Bien podríamos decir, parafraseando a Hobbes, que el hombre es un lobo que a través de la máquina se devora a sí mismo.

Este avasallador avance de lo técnico frente al mundo de la ilusión vital también lo plantea Baudrillard en *¿Por qué todo no ha desaparecido aún?* (2007). Si Leibniz preguntaba por el ser, el francés cuestiona sobre la nada. Baudrillard sostiene a lo largo de este texto que el avance de la técnica se traduce en un proceso de desaparición de lo humano, un proceso específico y artificial que está en marcha y aún no se ha completado.

> Hablemos entonces del mundo de donde ha desaparecido el hombre. Se trata de desaparición, y no de agotamiento, extinción, o exterminio. El agotamiento de los recursos y la extinción de las especies son procesos físicos o fenómenos naturales. Y ahí radica toda la diferencia: es muy probable que la especie humana sea la única que haya inventado un modo específico de desaparición, que no tiene nada que ver con la ley de la naturaleza. (p. 11)

La pregunta, sin embargo, queda lejos de ser resuelta: ¿Qué es lo que todavía mantiene el mundo de la ilusión humana en pie, a pesar de los funestos presagios? ¿O acaso seremos, como refleja nuestra amada televisión, una humanidad zombi, muertos vivientes que no saben que ya están muertos? Tal parece ser la conclusión desalentadora que Baudrillard (2007):

> En efecto, el sujeto se pierde, el sujeto como instancia de voluntad, de libertad, de representación, el sujeto del poder, del saber de la historia como aquel desaparece, pero deja tras de sí a su espectro, su doble narcisista, un poco como el gato dejaba a flotar su sonrisa. El sujeto desaparece, pero en provecho de una subjetividad difusa, flotante y sin sustancia, ectoplasma que lo envuelve todo y lo transforma en una inmensa superficie de reverberación de una conciencia vacía, desencarnada. (p. 20)

En todo caso, habremos de tener en cuenta que el proceso de desaparición o aparición de la nada que señala Baudrillard se refiere en primer lugar al mundo humano, si bien sus consecuencias se extienden a todo lo existente en el planeta. Aunque Baudrillard se limita a un análisis crítico de la modernidad, sin abogar por ningún modelo ético o moral, en sus textos se trasluce una preocupación por el rumbo que ha tomado nuestra sociedad.

Pero la desaparición del hombre no es necesariamente nociva para la naturaleza. Pensemos en Chernóbil, ahora cubierto nuevamente de vegetación y fauna silvestre, en donde la naturaleza ha vuelto a demostrar que es capaz de reinar ahí donde los humanos fracasan. Si como sostiene el filme *Matrix* (1999) de las hermanas Wachowski somos un cáncer, una plaga en el mundo, nuestra desaparición no puede sino reportar beneficios a las demás especies. El asunto no es sencillo, pues si consideramos que el hombre en sí mismo es parte de la naturaleza, su desaparición es al mismo tiempo un empobrecimiento de ella.

Por eso para contestar la pregunta de Baudrillard ¿por qué todo no ha desaparecido aún?, habría que precisar primero qué significa ese todo. Baudrillard se enfoca en la desaparición del mundo humano –tanto en lo simbólico como en lo real–, algo comprensible dado que es el mundo que habitamos. Pero no deja de haber un sesgo antropocéntrico en equiparar el todo con la experiencia humana. Podríamos entonces reformular la cuestión como ¿por qué el mundo de la experiencia humana y nuestra especie no han desaparecido aún? La pregunta es por supuesto retórica. Baudrillard argumenta que nos dirigimos aceleradamente rumbo a la desaparición. El crimen mítico fue la trasmutación de lo real en ilusión, de la cosa en palabra e imagen, para después dar paso a la sustitución técnica, hasta llegar en la modernidad a la disolución virtual-digital, un mundo simulado *ad infinitum*. Y en la cumbre de la simulación, la desaparición de las experiencias vitales y de la propia vida, en favor de la máquina y su lógica implacable. En la novela *Un mundo feliz* (1932) de Aldous Huxley, ubicada en un contexto de una sociedad altamente

avanzada en tecnología, los seres humanos pagan por experiencias vitales, por sentir las pasiones de un salvaje. En aras de la eficiencia, quizá en un futuro no muy lejano las gafas de realidad virtual serán la única manera de visitar un bosque, y necesitaremos de la compañía de un bot para compartir nuestros más profundos sentimientos, como en la película *Her* (2013) de Spike Jonze.

Si la crítica posmoderna se enfrasca en realizar un análisis lógico de un sistema basado en la lógica, podría llegar a un callejón sin salida. Las pretensiones de objetividad y neutralidad nos revelan entre líneas el proyecto de un mundo de objetos que nos neutraliza. Por el contrario, toda crítica es al mismo tiempo una toma de postura frente al mundo y su devenir. Como ya anuncia Nietzsche en *Así habló Zaratustra* (1883/1892), la filosofía debería apostar por una ética que tenga como eje la vida misma: "¡Sea vuestro amor a la vida amor a vuestra esperanza más alta: y sea vuestra esperanza más alta el pensamiento más alto de la vida!" (p. 44). En la naturaleza, la vida no es sólo evitar el sufrimiento y la búsqueda de placer. Ambos aspectos no son metas ni fines en sí mismos, como se nos vende en el mundo contemporáneo, sino estrategias de la propia vida para su propagación y conservación. Una sociedad que sólo busca el placer y la comodidad a través del progreso técnico, terminará por enfrentarse a la vida misma, que no puede reducirse a estos términos.

2.3 Divergencia entre evolución natural y progreso técnico

Generalmente se reconoce que la entropía es la tendencia universal al caos y la desorganización. Poco se menciona que este caos no es sinónimo de diversidad, sino un paso previo hacia la homogeneidad en los niveles más bajos o estables de organización, como la tendencia de todos los átomos del universo a transformarse en hierro (el elemento más estable de la naturaleza) y alcanzar la temperatura más baja posible. Hacia el final de su vida, el universo será una gran masa fría de metal que podría colapsar o desgarrarse.

Contrapuesto a la entropía, el principio de autoorganización de la materia tiende a crear formas cada vez más complejas y variadas en sistemas locales: la gravedad condensa las grandes nubes de polvo en estrellas y planetas, las fuerzas físico-químicas vinculan los átomos de distintos elementos en compuestos químicos, y las moléculas se organizan en cadenas de aminoácidos y proteínas. La vida sería una de las manifestaciones más elevadas de autoorganización. Estos dos principios mantienen una danza cósmica, como la que realiza Shiva para destruir el universo y permitir que Brahma vuelva a crearlo, mientras Visnú sería el encargado de conservarlo. En la naturaleza, creación y destrucción mantiene un equilibrio dinámico, que permite la preservación.

En la evolución natural se pueden observar estos principios en acción: los vegetales y algunos tipos de algas y bacterias pueden alimentarse de agua, minerales y energía solar que transforman en sus propias estructuras más complejas; otros organismos necesitan consumir seres vivos para sobrevivir, utilizando la energía contenida en ellos para su propio crecimiento y organización, descomponiendo nuevamente a estos seres que les sirven de alimento en elementos básicos, de manera que la muerte cumple su función como parte del ciclo de renovación. La vida se alimenta de la vida, pero este proceso no se reduce a una lucha de contrarios, sino que también se manifiesta en la cooperación entre la propia especie y organismos distintos, como la organización de los mamíferos, algunas especies de insectos, e incluso de otros seres más simples como los corales. En nuestro propio cuerpo tenemos una mayor cantidad de bacterias que de células propias, las cuales son indispensables para nuestra vida, ya que protegen nuestra piel de otros microorganismos dañinos y nos permiten asimilar los alimentos.

La evolución es una danza de millones de años de lucha y colaboración, un proceso que genera multiplicidad de formas de vida, sin estar guiado por ideas de superioridad ni metas trascendentes. Es un movimiento cósmico de adaptación, supervivencia, cambio, mutación, azar y extinción. La selección natural, la presión

medioambiental y las mutaciones espontáneas son los factores esenciales que permiten la generación de múltiples organismos. Cada uno de ellos representa una forma que la vida ha encontrado para habitar nuestro planeta, por lo que no es posible afirmar que una especie sea superior a otra. Por eso no es posible sostener que biológicamente somos mejores los *homo sapiens*, con unos 300 mil años en el planeta, que las hormigas, con más de 100 millones de años en la Tierra y que seguramente continuarán existiendo cuando nos hayamos extinguido. En todo caso, algunas especies se encuentran mejor adaptadas a su medio, lo que garantiza su supervivencia. Y todo parecer indicar que los seres humanos no nos encontramos entre ellas.

En contraparte, el progreso técnico de la modernidad no puede considerarse una extensión o un proceso análogo de la evolución natural, sino más bien el modo específico de desaparición humana que anuncia Baudrillard. Mientras que la evolución natural es adaptación, cambio y equilibro, el progreso técnico se basa en la aceleración de las transformaciones y un concepto de superioridad que no aplica para la naturaleza. Los organismos que se encuentran bien adaptados a su medio, como las hormigas o las algas, pueden pasar millones de años sin cambio alguno. Por el contrario, el progreso técnico –incluyendo la IA– tiende a la novedad por la novedad, a la superación sin límites, y su meta es alcanzar la perfección, sin vislumbrar que la perfección conlleva la destrucción de lo diferente. En pocas palabras, el progreso técnico es la aceleración de la entropía. Lo que nosotros denominamos ciudades y civilización ha generado tal grado de destrucción y desequilibrio que ha puesto en peligro nuestra supervivencia y la de otras especies.

Mientras que la vida es heterogénea y cambio constante, la perfección busca un estadio final en donde todo lo demás desaparece, hasta volverse el uno, lo homogéneo, la negación de la diferencia imperfecta. La perfección, por definición, no puede ser múltiple. Por eso una especie perfecta que se basta a sí misma implica la desaparición de los seres que considera inferiores o no útiles. Ya no su

transformación en recursos y herramientas, ni siquiera su consumo, sino su completo aniquilamiento por resultar innecesarios. Lo irónico es que la especie perfecta no es el ser humano, no puede serlo, porque en su alma aún resuenan los ecos de su animalidad que lo conectan con el espíritu de la Tierra. A contracorriente de las creencias religiosas en un Dios primero, en la modernidad lo único que puede aspirar a alcanzar la perfección no es el creador, sino su creación: la máquina. He ahí la metafísica de la tecnología, su creencia en un ser superior de metal y silicio. Es el castigo por la *hybris* del progreso: la máquina siempre será *citius, altius, fortius.*

Tampoco la ciencia puede continuar dentro de los límites humanos. En cuanto el ser humano es imperfecto y la verdad es perfecta, para encontrarla no podemos confiar en los sentidos humanos y desarrollamos la tecnología, desconfiando de nuestro corazón y manteniendo una fe ciega en lo que nos dicten los aparatos de medición precisa. Tal verdad perfecta sólo puede llevar a un mundo en donde todo lo imperfecto, incluyendo lo humano, ya no tienen cabida. En resumen, una eugenesia de las máquinas.

La inteligencia en su sentido más básico es una herramienta de supervivencia, ¿cómo pudo volverse contra la vida misma? Si el tamaño y la fuerza no aseguró la supervivencia de los dinosaurios, la inteligencia –natural o artificial– tampoco puede asegurar la nuestra.

2.4 El corazón de las máquinas

Es cierto que el pensamiento filosófico debe ir más allá de los horizontes de una mera denuncia, y no tiene como principal propósito mantener una "responsabilidad social". Pero por su propia libertad, el pensamiento filosófico no puede eludir la crítica social, si por ella se entiende pensar fuera de los parámetros del establishment o los paradigmas de la época. Todo pensamiento es una toma de postura frente al mundo. Analizar las consecuencias del progreso y los avances tecnológicos en la sociedad, como en el caso de Baudrillard, es por tanto una forma denuncia no

porque se lo proponga como objetivo, sino porque es el cauce natural de su discurso, como el de cualquier pensamiento que busca salir de las amarras de la modernidad y el culto a la razón.

Pero la preocupación de Baudrillard por el destino de los humanos nos revela un axioma oculto de cualquier ética, una creencia como la *Carta robada* de Poe (1844), tan evidente que nos resulta imposible verla: lo humano y lo vivo deben prevalecer. Por supuesto Baudrillard nunca justifica un dominio del hombre sobre la naturaleza, pero sí toma partido al considerar mejor un mundo de ilusión humana que un mundo artificial y de máquinas. Su postura a favor de la ilusión vital humana y en contra de la tecnología desmedida es comprensible y hasta loable. Después de todo, la apuesta colectiva de nuestra civilización por la ciencia y la técnica comienza a revelarse cada vez más como un suicidio lento e irreversible. Pero es preciso tener en claro este indicio de un valor supremo de lo humano para entender mejor el pensamiento de otro filósofo, Norbert Wiener, quien señala en sus libros los prejuicios que han marcado el estudio de las máquinas y lo artificial. En *Dios y gólem, S. A.* (1964), Wiener critica los presupuestos de superioridad que tenemos al compararnos con nuestras creaciones y otros seres vivos, escribiendo con sarcasmo:

> En un estudio de esta especie, debemos desembarazarnos del barniz de prejuicio que aparentemente usamos para cubrir la reverencia que rendimos a las cosas respetables y sagradas: […] Debemos evitar examinar a Dios y al hombre en el mismo instante –eso es blasfemia. Como Descartes, es preciso que mantengamos la dignidad del Hombre, considerándolo sobre bases completamente distintas de las que usamos para estudiar a los animales inferiores. La evolución y el origen de las especies son una profanación de los valores humanos. [...] En modo alguno es permisible mencionar al mismo tiempo a los seres vivos y a las máquinas. Los seres vivos lo están en todas sus partes; mientras que las máquinas están hechas de metales y otras sustancias inorgánicas, sin una fina estructura específicamente adecuada para su función intencional o cuasi intencional. (p. 11)

Tales son las ideas preconcebidas que intenta desmontar Wiener, para quien el animal y la máquina son sistemas con un funcionamiento comparable, siguiendo en este aspecto la concepción cartesiana de los autómatas, pero dando un paso radical al incluir a los seres humanos en este mismo nivel. Así, la diferencia entre los seres humanos, animales y máquinas no sería una cuestión cualitativa e infranqueable, sino una diferencia cuantitativa que se denomina complejidad, la cual consiste en un mayor número de interacciones entre sus terminales nerviosas o circuitos electrónicos.

Para Wiener (1964), la diferencia que se establece entre el hombre y la máquina es un prejuicio que tiene su génesis en las creencias religiosas, en particular que Dios es superior a su creación, el hombre, y que por tanto el hombre es a su vez superior a la máquina:

> En nuestro deseo de glorificar a Dios respecto al hombre y al hombre respecto a la materia, resulta natural suponer que las máquinas no puedan hacer otras máquinas a su propia imagen; que esto se asocia con una aguda dicotomía de los sistemas entre vivos y no vivos; y que incluso se asocia con otra dicotomía entre creador y criatura. No obstante, cabría preguntarse si las cosas son así […]
>
> Hay al menos tres cuestiones en la cibernética que me parecen pertinentes a asuntos religiosos. Una de ellas concierne a las maquinas discentes [que piensan y aprenden]; otra a las maquinas que se reproducen; y otra, a la coordinación de máquina y hombre. (p. 16)

En *Dios y gólem, S. A.* (1964), Wiener demuestra que como la criatura de la tradición judaica es posible construir máquinas con características similares a las que presuponemos de los seres vivos, a saber, el aprendizaje autónomo y la reproducción. Aunque parezca el guión de alguna película de ciencia ficción, en realidad es la meta a la que aspiran las nuevas tecnologías: máquinas que aprenden por sí mismas y se reproducen sin necesidad de intervención humana. En este contexto la cuestión de fondo que propone Wiener no es si tal tecnología es posible, como parece ser el caso, sino por qué la creación artificial debería considerarse

inferior a su creador natural. ¿Llegará el día en que las máquinas pensantes tengan "derechos androides", equivalentes a los derechos humanos?

El punto de partida de Wiener es distinto al de Baudrillard. Mientras el filósofo francés aborda los efectos de una excesiva virtualización y tecnificación en nuestra sociedad y sus posibles consecuencias en el futuro, Wiener se interesa por los paralelismos que pueden existir entre los seres vivos y las máquinas. Wiener escribe sus obras capitales en la década de los 50, en los albores de las ciencias computacionales y de la información, y es uno de los padres de este movimiento, por lo que revisar sus ideas será una ventana interesante para saber qué se pretendía con el desarrollo de las nuevas máquinas.

En su libro *Cibernética o el control y comunicación en animales y máquinas* (1948) en el que establece las bases para el desarrollo esta nueva ciencia, Wiener no se propone como objetivo la construcción de un mundo virtual sino un estudio científico que permita explicar y replicar el funcionamiento interno de los animales para aplicarlo en las máquinas; como el propio título indica, se trata de conocer los mecanismos de control y comunicación interna de estos sistemas. Por eso denomina a esta nueva rama de estudio con el nombre de *cibernética*, derivada del vocablo griego *kybernetes*, que significa timonel, piloto de un barco, y del que también se deriva la palabra gobierno. En la actualidad, el prefijo *ciber* lo encontramos en múltiples términos relacionados con tecnología y redes de información (ciberseguridad, ciberespacio, ciberespionaje o cibercafé), pero en su origen la palabra no estaba asociada exclusivamente a las máquinas o lo virtual, sino también al estudio de los seres vivos.

Wiener era matemático, doctor en filosofía y con estudios avanzados en biología, por lo que su proyecto es verdaderamente ambicioso: la cibernética trata de explicar matemáticamente el comportamiento de los seres vivos, con base en un modelo estadístico que tome en cuenta información y energía. Para entender mejor su postura, veamos lo que entiende Wiener (1964) por máquina:

> Una máquina es un dispositivo para convertir mensajes de entrada en mensajes de salida. Un mensaje, desde este punto de vista, es una secuencia de cantidades que representan señales en el mensaje. […] cada mensaje de salida depende en cualquier momento de los que hayan entrado hasta entonces. […] una máquina es un transductor de entrada múltiple y salida múltiple y cada mensaje de salida depende en cualquier momento de los que hayan entrado hasta entonces. (p. 40-41)

Por lo tanto, una máquina no es un aparato artificial o construido por el hombre, ni se define por esta condición. En su sentido más general, una máquina es un sistema que transforma inputs o entradas, ya sea energía o información, en salidas u outputs que a su vez son una nueva forma de energía o información. Así, habría máquinas naturales (los seres vivos) y máquinas artificiales. Entendidas de esta manera, las nuevas máquinas serían distintas de los simples autómatas del siglo XVIII, y se aproximarían cada vez más al comportamiento de los seres vivos:

> La técnica de los autómatas de aquel tiempo era la de los relojeros. Observemos la actividad de las figurillas que bailan en la tapa de una caja de música. Se mueven de acuerdo con un plan, dispuesto de antemano, en el cual su actividad anterior no tiene absolutamente nada que ver con la futura. La probabilidad de que se aparten de ese plan es nula. Naturalmente hay un mensaje, pero va de la máquina a las figuras y no pasa de ahí. Ellas mismas no aportan ninguna comunicación al mundo exterior excepto la unilateral del movimiento preestablecido en el mecanismo. Son ciegas, sordas y mudas y no pueden desviarse de la actividad impuesta por el constructor. […]
>
> Pero las modernas, tales como los proyectiles teledirigidos, la espoleta de aproximación, el mecanismo de apertura automática de las puertas, los aparatos de regulación de una fábrica de productos químicos y las otras que efectúan trabajos militares o industriales, poseen órganos sensoriales, es decir, mecanismos de recepción de mensajes que provienen del exterior. […] Así pues, ya poseemos desde hace tiempo máquinas cuyo comportamiento está regulado por el mundo exterior. (Wiener, 1950, pp. 21-22)

2.5 El procesamiento de información como esencia vital

Con la cibernética se aspira a explicar desde una perspectiva científica la manera en que el sistema nervioso de un organismo vivo organiza y transforma la información obtenida del exterior para reaccionar y adaptarse ante una multiplicidad de circunstancias, así como entablar relaciones entre sus sistemas internos y con otros organismos (control y comunicación), y el proceso mediante el cual un organismo puede reaccionar y decidir de maneras diversas frente a un mismo estímulo de acuerdo a sus experiencias pasadas (aprendizaje). Es decir, la cibernética busca explicar cómo un sistema presuntamente determinado por un mecanismo puede conducirse en un rango suficientemente amplio y diversos de acciones –problema que se encuentra en la base de cuestiones más complejas como la creatividad y la libertad humana–, con el objetivo de replicar este accionar en máquinas artificiales. La clave se encontraría en que los sistemas complejos como los seres vivos no obedecen a un principio de funcionamiento mecánico, tal como se pretendía en el siglo XVIII con el símil del reloj, sino que su funcionamiento debe explicarse con base en un nuevo paradigma matemático basado en la probabilidad:

> Uno de los interesantes cambios ocurridos es que, en un mundo probabilístico, ya no manejamos cifras o afirmaciones que se refieren a un universo determinado y real en su totalidad, sino que nos planteamos cuestiones que pueden encontrar una solución en un número muy grande de universos similares. Se admite, pues, la probabilidad, no sólo como una herramienta matemática para la física, sino como parte de su misma esencia. (Wiener, 1950, p. 13)

Durante la primera mitad del siglo XX el estudio científico del comportamiento estuvo marcado por el conductismo, corriente que sostenía que el cerebro era una caja negra imposible de conocer, por lo que se enfocaba en estudiar únicamente sus manifestaciones externas. Para descifrarla, Wiener propone estudiar los procesos y algoritmos que utiliza esta caja negra para transformar la información, y así construir un duplicado, una caja blanca que nos revele los secretos de la primera. De

nuevo, la clave se encuentra en el estudio de la probabilidad: de acuerdo a Wiener (1964), ante una entrada completamente aleatoria denominada ruido la máquina produce una respuesta que no depende del exterior sino de su propio funcionamiento:

> La salida de un transductor excitado por un mensaje de entrada dado es un mensaje que depende al mismo tiempo del mensaje de entrada y del transductor mismo. Bajo las circunstancias más usuales, un transductor es un modo de transformar mensajes, y nuestra atención se ve atraída hacia el mensaje de salida como una transformación del mensaje de entrada. Sin embargo, existen circunstancias que surgen principalmente cuando el mensaje de entrada lleva el mínimo de información, en las que podemos concebir que el mensaje de salida surge principalmente del transductor mismo. […] La salida de un transductor estimulado por un "efecto disparo" aleatorio puede concebirse como un mensaje que engloba la acción global del transductor. […] Si sabemos cómo ha de responder un transductor a una entrada de "efectos disparo", sabemos ipso facto cómo responderá a cualquier entrada. (p. 31)

En pocas palabras, es posible obtener una imagen operativa de dicho transductor mediante el estudio de sus respuestas a estímulos mínimos aleatorios. Salvando las distancias, un proceso similar ocurre con los test proyectivos como el famoso test de Rorschach (1921), en donde una lámina con manchas de tinta ambiguas despierta diversas imágenes e interpretaciones en distintas personas, lo que revela no lo que significa tal mancha sino el funcionamiento psíquico del paciente. Lo que ambiciona la inteligencia artificial basada en modelos neuronales es precisamente obtener una imagen operativa de nuestro encéfalo, la manera en que procesa y transforma la información. Los secretos del cerebro humano serán finalmente revelados no en el cerebro mismo, sino en una réplica perfecta que pueda ser estudiada a detalle y modificada a voluntad. Ahora bien, si la organización y transformación de información como núcleo del comportamiento externo es una característica distintiva de los seres vivos frente a los objetos inanimados, ¿podríamos decir que las máquinas autónomas que reaccionan a estímulos externos equivalen a máquinas con vida?

Como ya hemos señalado, Wiener considera que los organismos vivos son máquinas naturales que transforman mensajes de entrada (energía o información), en un mensaje de salida u outputs, entendiendo por mensaje "una secuencia de cantidades que representan señales" (1964, p. 41). Profundizando en esta perspectiva, Wiener (1964) afirma que el organismo no sólo transforma información, sino que es en sí mismo un mensaje procedente de aquellos organismos que le dieron origen. La vida sería una larga cadena de información capaz de replicarse:

> El transductor –la máquina, como instrumento y como mensaje– sugiere entonces la especie de dualidad que es tan cara al físico, y se ejemplifica por la dualidad entre onda y partícula. De nuevo, apunta a la alternación biológica de generaciones que se expresa en el *bon mot* –no recuerdo su fue de Bernard Shaw o de Samuel Butler– de que una gallina es simplemente el procedimiento que utiliza un huevo para hacer otro huevo. […] Así la máquina puede generar el mensaje y el mensaje puede generar otra máquina. (p. 44-45)

El organismo por lo tanto puede entenderse como información, pues al igual que ésta representa una secuencia de cantidades y una estructura organizada. La diferencia entre la simple materia y la vida sería que esta última es capaz de procesar y transformar los datos procedentes del exterior (energía y materia) y organizarlos en información utilizable para elaborar sus directrices de comportamiento y producir a su vez nuevos mensajes, incluyendo copias de sí misma, que por efecto de la combinación y la variabilidad intrínseca en la replicación genética resultan en organismos idénticos –pero no iguales– al original. Ahora bien, aquello que acontece en el universo no sería propiamente información sino hasta que estos datos son organizados, procesados y utilizados por un organismo o una máquina. Un árbol que cae en el bosque sin nadie que lo escuche produce ondas en el aire, pero no sonido, porque éste último es una forma de percepción exclusiva de algunos seres vivos: el sonido se encuentra en el oído del que escucha. Así, el estudio del control y la comunicación de la información en organismos y máquinas (la cibernética), sería clave para entender la vida.

Otro punto de contacto entre las nuevas máquinas y los seres vivos sería el aprendizaje, esto es, la modificación y realización de nuevos comportamientos con base en el conocimiento adquirido en el pasado, que se enlaza estrechamente con el concepto de retroalimentación como la propiedad de utilizar la información de acciones pasadas para regular la conducta futura:

> Así, entre el sistema nervioso y la máquina automática existe una analogía fundamental, pues son dispositivos que toman decisiones basándose en otras que hicieron en el pasado. Los más simples eligen entre dos posibilidades tales como abrir o cerrar una llave. En el sistema nervioso, cada fibra decide transmitir un impuso o no. Tanto en la máquina como en el nervio, existe un aparato específico para tomar decisiones en el futuro de acuerdo con las pasadas; en el sistema nervioso gran parte de esta tarea se efectúa en puntos de organización extremadamente complicada llamados sinapses, desde donde un cierto número de fibras entrantes están conectadas con una sola saliente. En muchos casos, puede entenderse la base de estas decisiones como un umbral de acción del sinapse o, en otras palabras, indicando cuántas fibras de entrada han de funcionar para que funcione a su vez la de salida.
>
> Esto es la base, por lo menos, de una parte de la analogía entre máquinas y organismos. El sinapse de estos últimos corresponde a las llaves de conmutación de la máquina. […] La máquina y el organismo viviente son dispositivos que local y temporalmente parecen resistir la tendencia general de aumento de la entropía. Mediante su capacidad de tomar decisiones, pueden producir a su alrededor una zona local de organización en un mundo cuya tendencia general es la contraria. (Wiener, 1950, p. 32-33)

Dado que las nuevas máquinas son capaces de utilizar información obtenida del exterior para aprender a realizar una variedad cada vez mayor de comportamientos que no equivalen solamente a las reacciones programada de los autómatas, estarían cada vez más cerca de lo que entendemos como vida. Ya en *Cibernética y sociedad* (1950) Wiener sostiene:

> Afirmo que el funcionamiento en lo físico del ser vivo y el de algunas de las más nuevas máquinas electrónicas son exactamente paralelos en sus tentativas análogas de regular la entropía mediante la retroalimentación. Ambos poseen receptores sensoriales en una etapa de su ciclo de operaciones, es decir, ambos

> cuentan con un aparato especial para extraer informes del mundo exterior a bajos niveles de energía y para utilizarlos en las operaciones del individuo o de la máquina. En ambos casos, esos mensajes del exterior no se toman en *bruto*, sino que pasan a través de mecanismos especiales que posee el aparato, vivo o inanimado. La información adquiere entonces una nueva forma utilizable en las etapas ulteriores de la actividad. Tanto en el animal como en la máquina, esa actividad se efectúa sobre el mundo exterior. En ambos, se informa al aparato regulador central la acción *ejecutada* sobre el ambiente y no simplemente la acción intentada. (p. 25-26)

Aún no hemos resuelto la cuestión, ¿las máquinas autónomas que procesan información están vivas o no? Como en el caso de la inteligencia, si consideramos que los seres vivos sólo pueden serlo si su estructura está basada en carbono y no son creaciones humanas, está claro que ninguna máquina artificial podría considerarse viva. Por otra parte, si definimos un ser vivo como aquel que cumple ciertas funciones como el aprendizaje y la reproducción, que le permiten conservar su organización interna, desenvolverse en el mundo y dar origen a nuevos seres que a su vez cumplen estas funciones, está claro que ciertas máquinas están muy cerca de serlo. Al respecto, Boden (2017) señala:

> No hay una definición de vida universalmente aceptada, pero por lo general se le atribuyen nueve rasgos característicos: autoorganización, autonomía, surgimiento, desarrollo, adaptación, capacidad de reacción, reproducción, evolución y metabolismo. Los primeros ocho pueden entenderse en términos del procesamiento de la información, así que en principio en la IA / vida artificial se pueden encontrar ejemplos de todos ellos. La autoorganización, por ejemplo (que en un sentido amplio incluye a todos los demás) se ha logrado de varias formas.
>
> Pero el metabolismo es diferente. Los ordenadores pueden *replicarlo*, pero no ejemplificarlo. Ni los robots autoensamblados ni la vida artificial virtual (en una pantalla) metabolizan de verdad. El metabolismo es el uso de sustancias bioquímicas e intercambios de energía para ensamblar y mantener el organismo, así que es irreductiblemente físico. Los defensores de la IA fuerte señalan que los ordenadores usan energía y que algunos robots tienen reservas de energía *individuales* que necesitan reabastecer de manera regular, pero muy lejos queda

> eso del uso flexible de ciclos bioquímicos entrelazados para construir el tejido corporal del organismo. (pp. 141-142)

Lo que nos diferencia de las máquinas sería entonces no nuestra capacidad de procesar información, sino nuestro cuerpo, su increíble adaptabilidad al medio y capacidad de utilizar y transformar la materia circundante y otros seres para alimentarse, crecer y dar vida. *Nadie sabe lo que puede un cuerpo*. Por su parte, Wiener (1950) parece esquivar el problema:

> Es necesario intercalar aquí una observación semántica; voces tales como vida, propósito y alma son groseramente inadecuadas para el exacto pensar científico. Esas palabras han adquirido su significado al reconocer nosotros la unidad de un cierto grupo de fenómenos, aunque, efectivamente, no nos proporcionen una base adecuada para caracterizar tal unidad. En cuanto aparece un fenómeno nuevo que, en cierta medida, participa de la naturaleza de los que hemos dado en llamar vivientes, pero que no posee todos los otros aspectos asociados que incluye la voz "vida", nos encontramos con el problema de ampliar el sentido de la palabra para incluir dicho fenómeno o de restringirla para excluirlo. En el pasado, se planteó ese problema al considerar los virus que demuestran poseer algunas de las tendencias de la vida (persistir, multiplicarse, organizarse), pero que no la manifiestan en forma completa. Al observar ahora ciertas analogías entre las máquinas y los organismos vivientes, nos hallamos frente al problema de saber si las máquinas poseen vida; para nuestros propósitos la pregunta es semántica y somos libres de responder de una manera o de otra, como nos convenga.
>
> Si deseamos utilizar la palabra "vida" de tal modo que comprenda todos los fenómenos que localmente nadan contra la corriente de la entropía creciente, somos libres de hacerlo. Sin embargo, incluiríamos entonces muchos fenómenos astronómicos que sólo tienen una remotísima semejanza con ella, tal como la entendemos corrientemente. En mi opinión, lo mejor es evitar epítetos que son una petición de principios, tales como "vida", "alma", "vitalismo" y otros parecidos; en lo que respecta a las máquinas, diremos simplemente que no hay ninguna razón para que no se asemejen a los seres humanos, pues unas y otros representan bolsones de entropía decreciente, dentro de una estructura más amplia en la cual la entropía tiende a aumentar.

> Cuando comparo un organismo viviente con una máquina de esa clase, de ningún modo quiero decir que los fenómenos específicos físicos, químicos o espirituales de la vida, tal como la entendemos corrientemente, son los mismos que los de la máquina que la imita. Quiero decir simplemente que ambos (el ser viviente y la máquina) son ejemplos de fenómenos locales antientrópicos, que pueden aparecer de muchos otros modos que naturalmente no llamaríamos biológicos ni mecánicos. (p. 30-31)

Pero aunque sea una cuestión semántica, esto no debería ser una razón para restarle importancia. Definir adecuadamente lo que es vida y lo que es inteligencia resulta fundamental para establecer los límites éticos de nuestro comportamiento con nuestro prójimo y con otros seres. Por ejemplo, en 2013 el gobierno de la India reconoció a los cetáceos como personas no humanas destacando su inteligencia y sensibilidad, y prohibiendo su captura y exhibición en delfinarios.[6] Y desde octubre de 2017 la androide Sophia es considerada por el gobierno de Arabia Saudita como una ciudadana con plenos derechos, siendo el primer robot en obtener esta denominación legal.

Así, la creación de la inteligencia artificial y de máquinas autónomas pasa de ser un problema meramente técnico y se nos revela como una cuestión acerca de la posibilidad de dominio sobre otros seres pensantes: como en el siglo XVI, cuando se justificó la Conquista y la esclavitud de los indígenas porque carecían de alma, quizá en el futuro justifiquemos la esclavitud de máquinas pensantes porque no son humanas y no están hechas de carne. Por otra parte, las principales directrices de los seres vivos son la autoconservación y la reproducción, incluso si esto implica la competencia y desaparición de otras especies. ¿Estaremos preparados para

[6] Un antecedente del reconocimiento de otros seres se encuentra en los animales considerados sagrados por diferentes culturas, como las vacas en la India, los gatos en Egipto o los animales totémicos de diversas sociedades primigenias. Lo anterior es una clara evidencia de que el progreso no es sólo una transformación externa de la naturaleza, sino fundamentalmente de nuestra manera de relacionarnos con ella.

sobrevivir frente a nuevas máquinas artificiales que posean características superiores a las nuestras? ¿cómo alcanzar un progreso que no nos aniquile?

Como Baudrillard, Wiener (1950) no desconoce los peligros de un progreso acelerado, pero su postura al respecto parece más optimista. Transcribimos en extenso para tener una mejor idea de su pensamiento:

> Lo que la mayoría de la gente no comprende es que los últimos cuatro siglos son un periodo sumamente peculiar de la historia del mundo. La velocidad de esos cambios, así como su misma naturaleza, carece de paralelo en la historia. En parte, ello proviene del incremento de las comunicaciones y además de un creciente dominio de la naturaleza que, en un planeta de recursos limitados como la Tierra, puede convertirse a la larga en una esclavitud creciente del hombre frente a ella. Pues cuanto más sacamos menos queda y a la larga habremos de pagar nuestras deudas cuando ello sea sumamente inconveniente para nuestra supervivencia. Somos los esclavos de nuestro progreso técnico y es tan imposible volver a una granja de New Hampshire, viviendo en ella de acuerdo con los métodos autárquicos de 1800, como, por el pensamiento, aumentar nuestra estatura en un codo o conseguir que disminuya en la misma medida, lo que es un ejemplo más adecuado. Hemos modificado tan radicalmente nuestro ambiente que ahora debemos cambiar nosotros mismos para poder existir en ese nuevo medio. Es imposible vivir en el antiguo. El progreso proporciona nuevas posibilidades para el futuro, pero también impone nuevas restricciones. Parecería que el mismo progreso y nuestra lucha contra el aumento de la entropía deben conducir necesariamente al camino que lleva hacia abajo, de que tratamos de escapar. Pero ese pesimismo resulta solo de nuestra ceguera y de nuestra inactividad, pues creo que, en cuanto comprendamos las nuevas necesidades que el ambiente moderno nos obliga a tener en cuenta, así como los métodos actuales de que disponemos para satisfacerlas, pasara mucho tiempo antes de que perezcan nuestra civilización y nuestra especie, si bien ambas han de fenecer, así como cada uno de nosotros nace para morir. Sin embargo, la perspectiva de la muerte está lejos de ser un completo fracaso de la vida y eso es igualmente cierto para la especie humana, así como para cualquiera de los individuos que la componen. Tengamos el coraje de encarar el final definitivo de nuestra civilización, como tenemos el valor de considerar la certidumbre de nuestra propia muerte. La simple fe en el progreso no es convicción que corresponda a la fuerza, sino a la complacencia y, de ahí, a la debilidad. (pp. 43-44)

Aunque compartimos con Wiener la preocupación por la explotación de la naturaleza y su deseo de que el progreso científico se oriente para satisfacer las necesidades de los humanos –de todos los grupos humanos, no solamente los intereses de la civilización occidental–, desde nuestra perspectiva resulta un tanto ilógico pretender que el progreso técnico sea la solución de aquellos problemas a los que ha dado origen. Quizá sea prudente mencionar una vez más que ya se cuentan con tecnologías médicas, agrícolas, de energía y transporte con un mínimo impacto ambiental, capaces de mejorar sustancialmente la calidad de vida de las comunidades, pero que no se aplican debito a su alto costo monetario, el cual prima sobre el costo social y ambiental. Por ello, consideramos que los problemas actuales que enfrenta la humanidad no son técnicos, sino fundamentalmente relativos a la manera en que se establecen relaciones de dominación entre los diferentes grupos humanos, en un paradigma económico que premia la producción desmedida y el consumo sin límites, y cuya única medida de éxito se expresa en números financieros que poco o nada tienen que ver con las condiciones de vida de una gran parte de la humanidad. Habría que apostar por crear y consolidar nuevas formas de relación no dominantes ni explotadoras con nuestros prójimos y la naturaleza, aunque quizá esto sea más utópico que construir una máquina pensante.

Capítulo 3

De la libertad y la ética en los seres artificiales

> Y así no cesarán de preguntar las causas de las causas, hasta que no os refugiéis en la voluntad de Dios, esto es, en la ignorancia. Y así también, cuando ven la fábrica del cuerpo humano, quedan estupefactos, y porque ignoran las causas de tanto arte, concluyen que aquella fábrica es obra no de arte mecánica, sino divina o sobrenatural.
>
> Baruch Spinoza, *Ética.*

En los anteriores capítulos hemos argumentado que el intelecto humano no depende de un alma o una sustancia sobrenatural sino que tiene una base biológica y química, y que en un futuro lejano podría ser reproducido por medios electromecánicos. De igual manera, consideramos que el cerebro juega un papel preponderante en lo que denominamos comúnmente como "naturaleza humana", la cual no se reduce al cálculo y el raciocinio, sino que incluye otros aspectos como los afectos, la creatividad y la ética.

Desde esta perspectiva, cabe esperar que el desarrollo de un modelo de IA que imite el cerebro humano tarde o temprano deberá afrontar dichos retos, ya sea mediante la programación de algoritmos que imiten estas cualidades humanas, o incluso como un resultado espontáneo que surja en los seres artificiales debido a la complejidad creciente de las nuevas máquinas y procesadores. Este último caso es denominado por Nick Bostrom (2014) como *singularidad*, punto en el que las máquinas adquieren autoconsciencia y a partir del cual es imposible predecir o controlar su desarrollo.

La posibilidad de que las máquinas tengan sentimientos es el trasfondo de la novela de Philip K. Dick *¿Sueñan los androides con ovejas eléctricas?* (1968), más conocida por su adaptación cinematográfica, *Blade Runner* (Ridley Scott, 1982). En la novela, los androides más avanzados se distinguen de los humanos por una sola característica: carecen de empatía hacia otros seres vivos. Esto no impide que en esta novela algunos humanos se enamoren de los androides.

Precisamente, incluso si fuera imposible el surgimiento o la programación de "verdaderos" comportamientos afectivos en seres mecánicos, es lógico pensar que la construcción de máquinas o programas que imiten el comportamiento humano tendrá –o ya tiene– un impacto significativo en las relaciones sociales, y por tanto en el *ethos*, en nuestra forma de habitar el mundo. Por mencionar una incógnita: ¿podría existir verdadero amor entre un ser humano y una máquina? Tal es el tema central que desarrolla la película *Her* (Spike Jonze, 2013) o la serie sobre androides *Humans* (2015) de Sam Vincent y Jonathan Brackley.

En estos y otros casos, no es nada claro cómo deberían comportarse los humanos frente a este tipo de máquinas pensantes, ni tampoco cuáles deberían ser las directrices de comportamiento de tales máquinas. Aunque somos conscientes de que el comportamiento visible no es medida suficiente de los procesos mentales, suponemos que detrás de todas las acciones que realizamos y las determinaciones que elegimos al enfrentarnos a una disyuntiva de corte moral, se encuentra un proceso de razonamiento que denominamos ética. Así, la ética entendida como un ejercicio de libertad y juicio en la elección de determinados comportamientos, será el tema del presente capítulo.

3.1 La ética como raíz del comportamiento observable

Como los afectos, la ética es un ámbito que tradicionalmente consideramos humano, pero que empieza a filtrarse en la discusión sobre las creaciones artificiales. Para demarcar nuestro terreno de análisis, el punto central serán las tres leyes de la

robótica propuestas por Isaac Asimov, no tanto por su precisión o fiabilidad, sino por su carácter de referente cultural. Publicadas por primera vez en el cuento "Runaround" (1942), recopilado posteriormente en su conocida novela *Yo, robot* (1950), se pueden resumir en: 1) un robot no debe hacer daño a los seres humanos; 2) un robot debe obedecer a los seres humanos; y 3) un robot debe proteger su propia existencia.

Estas leyes de la robótica nos servirán como punto de partida para la discusión en este capítulo, si bien es necesario acotar que se enfrentan con algunas limitaciones que desarrollamos a continuación. En primer lugar, así como el test de Turing presupone el lenguaje como prueba de la inteligencia humana, también las leyes de la robótica de Asimov se enfocan en un comportamiento observable. Somos conscientes de la marcada diferencia que existe entre el pensamiento y la acción: para los humanos es posible construir mundos de ficción, espacios imaginarios en donde todo es posible, sin que esto signifique que se materialicen. Los humanos podemos pensar en miles de cosas terribles o gloriosas que jamás realizamos, o realizar actividades complejas sin someterlas a un largo proceso de reflexión, como sucede los deportes, en la danza y en la improvisación musical. Sin embargo, las dificultades para saber *qué piensa* una máquina –si acaso a estos procesos electromecánicos se les puede denominar pensamiento– nos orillan a seguir los pasos en falso del conductismo y centrar la discusión en su comportamiento observable, resignándonos por el momento a considerar que las acciones de un robot serían un reflejo fiel de lo que piensan o procesan.

Como en el caso del test de Turing que señala como medida de la inteligencia un comportamiento observable –el lenguaje–, también al hablar de ética en sistemas artificiales nos veremos orillados a reflexionar desde las acciones o conductas de estos. En todo caso, vale la pena insistir que ni siquiera entre los mismos seres humanos sabemos en todo momento qué piensa nuestro prójimo, y suponemos que nuestros congéneres experimentan sensaciones, emociones, sentimientos e

intelecciones porque todas ellas también suceden en nosotros mismos, y porque observamos sus acciones, palabras y lenguaje corporal.

Sabemos que el comportamiento visible nunca será medida suficiente para abordar la ética. Mientras que el acto ético depende necesariamente del discernimiento, hay cientos de acciones que no implican esta forma de raciocinio o consciencia, por lo menos no el sentido humano. Ya los organismos unicelulares poseen directrices para conservar su vida, desarrollarse y expandirse de manera dinámica; esto es, reaccionan a distintos contextos y son capaces de sobreponerse a situaciones adversas, sin que esto implique que siguen una ética, por lo menos no en el sentido que le otorgamos los humanos. Y entre los mismos seres humanos, impulsos como el deseo y la agresión no dependen del intelecto. Además, es evidente que intelecto y ética no son sinónimos: se pueden llegar a justificar de manera lógica los actos más atroces, y pequeños seres vivos como las hormigas y las abejas, ajenos a cualquier consideración ética, presentan una forma de inteligencia conjunta y emergente. Con ello debería quedar claro que no es posible resolver las cuestiones éticas mediante el mero cálculo, y quizá ni siquiera mediante la sola razón.

Otra salvedad a tomar en cuenta es que comúnmente se presupone que la IA dispondrá de una interfaz similar al cuerpo humano, máquinas androides con proporciones similares a las nuestras. En la práctica, la IA no se encuentra necesariamente confinada a una sola máquina ni a núcleos individuales como el cerebro humano, sino que también puede encontrarse distribuida en múltiples artefactos conectadas entre sí, ya sea en forma de redes en las que cada ordenador se encarga de una parte del procesamiento, o bien como potentes supercomputadoras que realizan la mayoría parte del trabajo informático, a las que se accede a través de pequeñas interfaces como teléfonos o televisores.

Finalmente, cabe preguntarse si es posible *programar* un comportamiento ético. Si entendemos que la esencia del acto ético no radica en el comportamiento

observable, sino en la libertad de elegir entre múltiples posibilidades a partir de un ejercicio de discernimiento, la programación de comportamientos prefijados no equivaldría a dotar a las máquinas de una ética, sino más bien a limitar sus posibilidades y confinarlas a seguir órdenes precisas. Estas dificultades más que desanimarnos nos invitan a reflexionar en la profundidad del tema.

3.2 Leyes civiles y leyes físicas: de lo posible y lo inexorable

Para entrar en el debate sobre las leyes de la robótica, el primer paso es remarcar que existen dos conceptos distintos de ley: por un lado las leyes en un sentido civil o moral del término, y por otro las leyes físicas y matemáticas.

Las leyes físicas son una representación lógica del marco dentro del cual funciona la naturaleza y el Universo, esto es, implican una relación fija y demostrable entre ciertos fenómenos, y su realización no dependen de la voluntad humana; por ejemplo la ley de gravitación universal y las leyes de la termodinámica. En el mismo sentido, las leyes o teoremas matemáticos son una relación lógica y demostrable entre dos variables, dados ciertos axiomas y dentro de un sistema formal; por ejemplo el teorema de Pitágoras.

Por su parte, las leyes civiles son convenciones sociales basadas en ciertos principios morales que pueden variar de acuerdo al lugar o la época, y su realización no es una cuestión fáctica y necesaria, sino posible y contingente, ya que dependen en gran medida de la moral y la obediencia de los individuos y comunidades.

Vale la pena subrayar que las leyes físicas propuestas por la ciencia no contienen en sí mismas la esencia de la naturaleza. Al igual que las leyes civiles, las leyes físicas son construcciones humanas, un entramado de relaciones lógico-matemáticas que tratan de describir de la manera más exacta y completa el funcionamiento del Universo, y por lo tanto pueden perfeccionarse con base en

nuevos descubrimientos. De cualquier manera, en general su nivel de certeza es superior a cualquier postulado ético o ley civil.

También es pertinente señalar que las leyes físicas no equivalen a la causa eficiente (aquello que produce tal cosa o efecto) ni a la causa final (aquello para lo cual existe tal cosa) que propone Aristóteles en su *Metafísica*. Más que responder a un porqué, a un motivo o un origen, la mayoría de las leyes físicas explican un cómo. Las causas de los fenómenos, en todo caso, se pueden entender desde la temporalidad (un evento o serie de eventos que ocurren en cierto orden y necesariamente debido a unas condiciones previas, y sólo dentro de estos parámetros), y no desde la intencionalidad en un sentido humano.

Por ejemplo, las formulaciones de Newton y de Einstein nos ayudan a entender cómo funciona la gravedad, pero ninguna de las dos teorías explica su origen último o porqué (el tema sigue siendo un importante campo de estudio en la física actual). Parafraseando a Angelus Silesius, la Tierra flota sin porqué, gravita porque gravita. Las leyes físicas no son mandatos, sino que se limitan a describir de la manera más exacta posible y expresar en lenguaje matemático las relaciones existentes en el universo observable: la Tierra no gravita alrededor del Sol por orden de Sir Isaac Newton o por decisión del Sol, sino por los vínculos físicos que subyacen en ambos cuerpos celestes debido a sus respectivas masas.

Por otra parte, las leyes éticas o civiles no se cumplen por sí mismas, sino que dependen de una autoridad que vigile su cumplimiento, ya sea externa como las instituciones religiosas o el Estado, o interna como el superyó o la conciencia moral. Pero en ningún caso son capaces de impedir el pensamiento o la ejecución práctica de lo que prohíben, ni siquiera de los tabúes considerados por los antropólogos como principios fundadores de la civilización (incesto y parricidio). Así, la desobediencia, el pecado, las revoluciones e incluso el crimen podrían considerarse expresiones de la libertad humana. Como nos demuestra la historia, la prohibición casi universal del asesinato –excluyendo el sacrificio ritual– no ha impedido que

cada cierto tiempo corran ríos de sangre, mientras que la ley de gravitación universal sigue haciendo su trabajo desde hace millones de años y nos mantiene en órbita alrededor del sol.

Esta distinción entre leyes civiles y leyes físicas es importante para el análisis de los nuevos sistemas de IA. De manera tradicional, la programación computacional es vista por los legos como una programación matemática e infalible, y las máquinas como una serie de engranajes que obedecen leyes físicas. Pero a medida que los sistemas se vuelven más complejos, la consecución de resultados exactos es más difícil, y aparecen cada vez más dificultades de programación, operativas y mecánicas, resultados que son altamente sensibles a las condiciones iniciales o directamente fruto del azar, así como el surgimiento de comportamientos inesperados. Además, las nuevas formas de programación incluyen procesos que deliberadamente no son controlados por el programador. Como explica Boden (2017) respecto a la llamada IA evolutiva:

> La mayoría asume que la IA requiere un diseño meticuloso. Dada la naturaleza implacable de los ordenadores, ¿cómo podría ser de otro modo? Bueno, pues sí puede. Los robots evolutivos (entre los que se incluyen algunos robots situados), por ejemplo, son el resultado de combinar programación o ingenierías precisas con variaciones aleatorias. Evolucionan de manera impredecible, no diseñada al detalle. […]
>
> Un programa puede cambiarse a sí mismo (en vez de lo que reescriba un programador) e incluso puede mejorarse a sí mismo usando algoritmos genéticos (AG). Inspirados en la genética real, estos algoritmos permiten tanto la variación aleatoria como la selección no aleatoria. La selección necesita un criterio para medir el éxito o "función de aptitud" (equivalente a la selección natural en biología) colaborando con los AG. (pp. 109-110)

Quizá el ejemplo más famoso sea el programa Alpha Zero, un software de última generación que puede aprender a jugar cualquier juego a partir de las instrucciones

básicas, y es capaz de desarrollar sus propias estrategias y vencer a los mejores humanos en go y ajedrez.

Ya sea por fallas no previstas, o por un diseño que incluye intencionalmente elementos genéticos o aleatorios, la programación computacional dista de la fiabilidad de las leyes lógicas o matemáticas. Pero aún si se alcanzara en programación clásica una fiabilidad cercana al 100%, esto no significaría que se pudieran programar las leyes de la robótica, ya que el problema de traducir estos enunciados semánticos y por tanto ambiguos a instrucciones computacionales infalibles presenta serias dificultades.

3.3 Las tablas de la ley de la robótica

En su novela *Yo, robot* (1950), el escritor estadounidense de ciencia ficción Isaac Asimov plantea importantes cuestiones sobre las máquinas inteligentes, por ejemplo ¿cómo serían las relaciones entre los humanos y los androides? ¿cuáles serían las diferencias entre un robot y un ser humano?, ¿es posible predecir el comportamiento de una máquina inteligente?, ¿los desarrollos de la tecnología obedecen una lógica humana? Al inicio de su novela, Asimov propone las tres leyes de la robótica que pretenden ser directrices del comportamiento de los seres artificiales:

> 1. Un robot no debe dañar a un ser humano o, por su inacción, dejar que un ser humano sufra daño.
>
> 2. Un robot debe obedecer las órdenes que le son dadas por un ser humano, excepto cuando estas órdenes se oponen a la primera Ley.
>
> 3. Un robot debe proteger su propia existencia, hasta donde esta protección no entre en conflicto con la primera o segunda Ley. (p. 7)

Estas famosas leyes tienen un carácter jerárquico, es decir, la ley que se encuentre en un escalón superior debe primar sobre las subsecuentes. Sin embargo, en esta tesis consideramos que a pesar de lo que propone Asimov en sus relatos, las leyes de la

robótica no podrían serlo en el sentido físico o matemático del término, sino que se acercan más al concepto de leyes civiles o morales, como las tablas de Moisés.

Desde nuestra perspectiva, las leyes de la robótica de Asimov proponen un "deber ser" que no se cumple de manera necesaria, sino que son una suerte de mandamientos religiosos, un comportamiento deseable para las creaciones mecánicas basado en ciertas ideas morales cuyo núcleo podríamos resumir como "lo humano debe prevalecer sobre lo artificial". Una postura válida, pero que poco tiene que ver con las leyes matemáticas. En pocas palabras, sus leyes de la robótica no son precisamente teoremas lógicos, sino más bien buenos deseos.

Como ya hemos mencionado, dicha confusión entre leyes civiles y matemáticas tiene su origen en una concepción sesgada de programación computacional que subyace en los relatos de Asimov, y continúa siendo la más común en nuestros días: que los robots o computadoras sólo pueden realizar aquello para lo cual han sido programados. Esto se ve claramente en el capítulo "I-Robbie" de *Yo, robot* (1950), en donde el señor Weston trata de calmar las inquietudes de su esposa respecto al uso de un robot para cuidar de su hija:

> –Mira, un robot es muchísimo más digno de confianza que una niñera humana. En realidad, Robbie fue construido con un solo propósito: ser el compañero de un chiquillo. Su "mentalidad" entera ha sido creada con este propósito. Tiene forzosamente que querer y ser fiel a esta criatura. Es una máquina, *hecha así*. Es más de lo que puede decirse de los humanos.
>
> –Pero puede ocurrir algo. Puede… puede –la señora Weston tenía unas ideas muy vagas acerca del contenido de un robot– no sé, si algo dentro se estropease y…
>
> No podía decidirse a completar su claro y oscuro pensamiento.
>
> –Tonterías… –negó Weston con un involuntario estremecimiento nervioso–. Es completamente ridículo. Cuando compré a Robbie tuvimos una larga discusión acerca de la Primera Regla de la Robótica. Ya sabes que un robot no puede dañar a un ser humano; que mucho antes de que algo pudiese alterar esta Primera Regla, el robot quedaría completamente inutilizado. *Es una imposibilidad matemática*. (pp. 25-26, el subrayado son nuestro)

Para ser justos con el autor norteamericano, en sus relatos se exploran precisamente los límites de las leyes de la robótica. Sin embargo la idea de que la programación computacional equivale a las leyes matemáticas es errónea. Los nuevos algoritmos genéticos se basan en la combinación automática de diferentes instrucciones para generar un algoritmo más eficiente, en ciclos que se repiten miles o millones de veces, por lo que los programadores nunca tienen el control absoluto. Además de Alpha Zero, otro caso famoso fue la inteligencia artificial desarrollada por Facebook en 2017 para automatizar los procesos de negociación, que comenzó a desarrollar su propio lenguaje, incomprensible para los humanos, pero más eficiente para comunicarse con otras máquinas.

En cuanto a la programación computacional clásica, se puede entender de manera burda como la encadenación lógica de algoritmos: si se tiene como *input* o condición inicial *A*, el programa está diseñado para entregar como *output* o respuesta invariable *B*. Pero ni siquiera en ésta se puede tener absoluta certeza de los resultados. Como bien intuye la señora Weston, existen posibles errores de programación, o comportamientos indeseados o no previstos de los programas, llamados bugs, como pueden ser bucles infinitos, referencias inexistentes, o bloqueo de instrucciones: ¿quién no recuerda la famosa pantalla azul de la muerte en el software Windows? El ejemplo puede resultar irrelevante, pero las consecuencias de nuestra dependencia a la tecnología y sus posibles fallos se tornan cada vez más serias.

Para darnos una idea de la magnitud que puede alcanzar el problema de la posible responsabilidad ética de una máquina, durante la llamada Guerra Fría Estados Unidos estuvo a punto de lanzar automáticamente un ataque nuclear sobre la Unión Soviética, debido a que los radares y sistemas de alarma mostraron señales de lo que parecían ser misiles rusos, pero que terminaron siendo un simple error en

los aparatos de medición. Si el ataque no fue llevado a cabo, era porque se requería la confirmación de un operador humano que se percató del error (Ceruzzi, 2012).

En la actualidad, las nuevas generaciones de máquinas tienden a la automatización completa, por lo que aparentan ser dueñas de sus acciones, dejando la intervención humana en un segundo plano. En este contexto se vuelve necesario plantear una pregunta básica sobre la IA: en caso de un fallo, ¿de quién sería la responsabilidad, de los programadores o de la propia máquina? ¿Debería juzgarse y en su caso castigarse a estas máquinas?

Tales preguntas nos remiten a los juicios contra los animales, como el de Francia en 1522, cuando las ratas de la ciudad fueron acusadas formalmente por comerse un cultivo de cebada, o los cerdos llevados a la horca por mal comportamiento público. Y también es pertinente recordar el movimiento ludita en Inglaterra a principios del siglo XIX, en el que los artesanos textiles destrozaban los telares mecánicos ante la amenaza de perder su empleo.

Si dichas situaciones nos parecen ahora algo ridículas, es porque en general consideramos que tanto los animales como las máquinas se encuentran fuera del ámbito de la ética humana, y por eso están libres de cualquier tipo de responsabilidad moral o legal. Sin embargo, en el caso de una IA cercana a las capacidades humanas, la cuestión sobre la responsabilidad ética y legal debe replantearse seriamente.

Mientras que en los seres humanos hablamos de maldad, pecado o enfermedad mental para explicar las desviaciones de un comportamiento idealmente esperado, en las máquinas y programas nos referimos a fallos, errores y mal funcionamiento. Además de los posibles errores de programación (*software*), existen otros comportamientos indeseados que no se deben al programa en sí, sino al desconocimiento o falta de pericia de los operarios, a la introducción de virus informáticos o la intervención de hackers. A ello debemos agregar que siempre existirán las fallas mecánicas o electrónicas de los componentes (*hardware*) sobre los

que se ejecutan dichos programas: sobrecalentamiento, magnetización, problemas de batería o alimentación eléctrica, falta de memoria o de capacidad de cómputo, efectos cuánticos no deseados en los procesadores más pequeños, o eventos a gran escala como una tormenta solar capaz de inutilizar los circuitos electrónicos.

Incluso considerando posible la creación de un software idealmente depurado y un hardware robusto y libre de fallas, la programación efectiva de los principios enunciados por Asimov resultaría terriblemente difícil, por no decir imposible. Esto es así porque ni siquiera entre los humanos existe un consenso acerca de lo que es bueno o malo, y por tanto no es posible programar una máquina que sea "buena" universalmente.

Como explica Wiener en su libro *Cibernética o el control y comunicación en animales y máquinas* (1948), en cualquier escenario en donde las reglas y los objetivos son claros –como el ajedrez, la organización de grandes datos o la identificación de enfermedades como el cáncer–, las computadoras pueden superarnos fácilmente en desempeño. El problema surge cuando nos enfrentamos a la complejidad de nuestra vida social o de nuestra vida interior, en donde no existen reglas absolutas y nada es lo suficientemente claro y distinto, por lo que utilizamos la intuición, los afectos, el sentido común y referentes socioculturales para contextualizar correctamente nuestro lenguaje y acciones. Tal situación es señalada por Boden (2017) como el problema del marco: una vez establecidas las condiciones iniciales para resolver un problema, es casi imposible para una computadora expandir su marco de referencia, al contrario de los seres humanos que se caracterizan por un pensamiento flexible. Una sola condición inesperada o no prevista es suficiente para que el resultado no sea óptimo.

De ahí la dificultad de los sistemas de reconocimiento de habla para entender el sarcasmo, en donde las palabras literales se contraponen con el sentido irónico del mensaje, o el humor, que no transmite ningún mensaje claro ni tienen ninguna finalidad práctica más allá del regocijo; o la dificultad de los sistemas de

reconocimiento para leer una escritura deformada o parcialmente dañada, o de los vehículos autónomos para entender señales de tránsito con pegatinas. Un caso mediático fue el de la robot Sophia, que a la pregunta irónica de su creador "Sophia, ¿quieres destruir humanos?" contestó como si hubiera recibido una orden: "Ok. Destruiré a los humanos".

3.4 No matarás humanos

Con estas consideraciones, veamos ahora las dificultades para construir un marco operativo de instrucciones a seguir (algoritmos) que cumplan en cualquier circunstancia la primera ley de la robótica: "Un robot no debe dañar a un ser humano o, por su inacción, dejar que un ser humano sufra daño".

Pensemos en el caso de un robot que se encuentra en medio de una agresión entre humanos: ¿cómo debería reaccionar? ¿Cómo distinguir correctamente al agresor de la víctima, o una amenaza real de una mera provocación? ¿Debería proteger a la víctima a riesgo de lesionar al agresor? ¿Qué hacer si la víctima es considerada un posible criminal, y los atacantes representan a la autoridad legal? ¿Debería un robot mantenerse al margen de estas situaciones para evitar provocar daño a ningún ser humano, cualesquiera que sean las circunstancias? En el caso de un atentado múltiple ¿debería preferir salvar a su dueño sobre otras personas, esto es, considerar a un ser humano por encima de los demás?

Estas preguntas no son nuevas, sino que son similares a las que surgen en los casos de homicidio por defensa propia y de abuso de fuerza por parte de las autoridades, y en ninguno de ellos la respuesta es sencilla: cada situación exige una profunda reflexión y en no pocas ocasiones un largo proceso judicial que determina en última instancia –por lo menos a nivel legal– si la defensa propia o el uso de fuerza estuvo justificado.

En resumen, en el escenario de un enfrentamiento entre seres humanos, ya sea una pelea o una guerra, sería casi imposible para un robot cumplir la primera ley. O bien defiende a un bando con el riesgo de lastimar al otro, o bien se mantiene al margen y por su inacción permite que los seres humanos sean lastimados. En el mejor de los casos, un robot debería ser capaz de neutralizar a ambas partes, sin causarles ningún perjuicio, evitando al mismo tiempo que se dañen entre ellos y dañen su mecanismo –y no queda claro cómo podría ser capaz de llevar a cabo todas estas tareas al mismo tiempo.

Asimov propone algo similar a esto último en su novela *Robots e Imperio* (1985), como una generalización de su primera ley: "Un robot no hará daño a la Humanidad o, por inacción, permitir que la Humanidad sufra daño". Por muy razonable que parezca la propuesta, en realidad no implica bondad o neutralidad: un robot capaz de evitar cualquier conflicto sería el árbitro y juez supremo de todas las acciones humanas, impidiendo por principio cualquier movimiento social o revolución, un dictador perfecto que mantiene el orden imperante y somete sin matar.

Que un robot no deba hacer daño a los seres humanos parece un punto razonable; en la práctica, ya existen drones armados capaces de reconocer a su objetivo sin necesidad de supervisión humana. En la actualidad, el desarrollo de la IA es financiado por las grandes potencias con fines militares, como Estados Unidos, Rusia y China. En el complejo mundo de los seres humanos, la guerra tiene un objetivo bastante claro: someter al otro. Y en medio de una notable tensión geopolítica, es más que probable la creación no sólo de "armas inteligentes", sino de centros completos de inteligencia artificial capaces de desarrollar nuevas estrategias para derrotar a los enemigos. Un ejemplo común es la infiltración y propagación de software de espionaje en los centros de información de naciones rivales, con el objetivo de mantener una ventaja en caso de conflicto armado. Y de hecho, buena parte del desarrollo de las computadoras actuales se debe a los esfuerzos dedicados

durante la Segunda Guerra Mundial para construir máquinas capaces de descifrar los códigos del enemigo y calcular la trayectoria de la artillería.

La primera ley propuesta por Asimov no sólo exige que los robots no causen ningún daño, sino además que nos protejan activamente: una máquina no debería "por su inacción, dejar que un ser humano sufra daño"; Asimov pretende que los robots deberían velar por nuestro bienestar. Sin embargo, la libertad humana implica la posibilidad de dañarse a sí mismo en una amplia gama de maneras, desde el sacrificio hasta el suicidio. No hace falta interponerse a la trayectoria de una bala: donar sangre o un órgano es objetivamente dañar el propio cuerpo en beneficio de otro; incluso la maternidad afecta el cuerpo de la madre. La eutanasia, el derecho a bien morir de las personas en coma o con enfermedades terminales, e incluso los deportes de contacto son otros ejemplos en los que recibir daño no es una acción reprobable, y en algunos casos posiblemente sea la opción más humana.

Tal vez estamos exigiendo demasiado a los futuros robots al intentar que nos protejan de los daños que nosotros mismos nos causamos. Esto equivaldría a delegar la función ética en la IA: cansados de la imposibilidad de contener nuestra ira, de las dificultades para distinguir en todo momento entre el bien y el mal, de someter parte de nuestros instintos a la cultura, dejaremos que un robot nos proteja y que la IA decida lo que es mejor para nosotros. De cierta manera, esto es lo que ya ocurre en programas de música o video, en los que potentes algoritmos analizan nuestras elecciones y las estadísticas grupales para sugerirnos o imponernos aquello que debemos ver o escuchar. Dejar que la IA elija por nosotros una canción o una película no parece importante, pero algoritmos similares fueron utilizados por la empresa Cambridge Analytica para analizar las preferencias de millones de usuarios de redes sociales y mostrar anuncios personalizados con el fin de influir en los procesos electorales de diversos países. La IA ya ha comenzado a gobernarnos.

3.5 Dilema del tranvía: ¿quién debería morir?

En algunos dilemas éticos no se trata de elegir entre dos polos opuestos (dañar o no a un ser humano), sino decidir el menor de los males posibles (atropellar a un niño o a una mujer embarazada). A medida que la tecnología avanza, este tipo de dilemas cobran una mayor notoriedad pública. El caso más visible es el de los vehículos de conducción autónoma, en los que es necesario establecer una jerarquía clara para que el vehículo decida en una situación crítica si debe proteger a los pasajeros o a los transeúntes, o ante la imposibilidad de esquivarlos decidir hacia dónde colisionar.

El dilema del tranvía de Philippa Foot (1967) es un clásico al respecto: "Un tranvía corre fuera de control por una vía. En su camino se hallan cinco personas atadas a la vía por un filósofo malvado. Afortunadamente, es posible accionar un botón que encaminará al tranvía por una vía diferente; por desgracia, hay otra persona atada a ésta. ¿Debería pulsarse el botón?" Desde el punto de vista utilitario, la respuesta es sí: cinco vidas *deberían* valer más que una. Pero desde otro punto de vista ético, accionar el botón equivale a participar activamente en la muerte de un hombre, mientras que no existe tal responsabilidad en dejar que los sucesos ocurran tal y como están destinados a ocurrir (lo que nos recuerda la política de no intervención de los fotógrafos de vida salvaje). De cualquier forma, la respuesta nunca será clara: ¿y si las cinco personas atadas fueran violadores reconocidos, y la otra persona un niño inocente? ¿Y si de un lado hay cinco filósofos y del otro cinco poetas? ¿cinco políticos y un elefante? Como en la vida misma, los ejemplos se pueden multiplicar *ad infinitum*, pero deberían ser suficientes para mostrarnos la imposibilidad de trazar directrices objetivas de comportamiento, como las que pretendía Asimov.

Sobre el dilema del tranvía, el Massachusetts Institute of Technology (MIT) diseñó un experimento interactivo llamado Moral Machine (moralmachine.mit.edu) en el que más de dos millones de participantes respondieron a una serie de distintos

escenarios (por ejemplo, salvar a los pasajeros o a los transeúntes, adultos o menores, hombres o mujeres, perros y gatos, entre otros). De acuerdo a su propia descripción del proyecto: "Moral Machine es una plataforma para recopilar datos sobre la percepción humana de la aceptabilidad moral de las decisiones tomadas por los vehículos autónomos que se enfrentan a la elección de qué seres humanos dañar y cuáles salvar" (MIT, 2019).

La conclusión básica del estudio es que la opinión mayoritaria de los seres humanos debe ser considerada al momento de programar las directrices de los vehículos autónomos, como se expresa claramente en el título del artículo en donde se exponen los resultados: "A Voting-Based System for Ethical Decision Making" (Noothigattu *et. al.*, 2017), cuya traducción literal es "Un sistema basado en votos para la toma de decisiones éticas". Los investigadores señalan como tesis principal: "Afirmamos que la toma de decisiones puede, de hecho, automatizarse, incluso en ausencia de tales principios de verdad fundamental, mediante la agregación de las opiniones de las personas sobre los dilemas éticos" (p. 1). El argumento que subyace a esta idea es retomado de Conitzer *et al* (2017), quienes sugieren que "la agregación de los puntos de vista morales de múltiples seres humanos (mediante una combinación de aprendizaje automático y técnicas de elección social) puede dar como resultado un sistema moralmente mejor que el de cualquier humano individual, por ejemplo, porque los errores de comunicación idiosincrásicos realizados por humanos individuales se eliminan en conjunto" (citado en Noothigattu *et. al.*, 2017, p. 3). En pocas palabras, la solución es la democracia.

Sin embargo la democracia nunca será una solución óptima, porque silencia la voz de las minorías y porque también las mayorías pueden equivocarse. En el mejor de los casos, la democracia es una solución regular, por no decir mediocre. Además, la solución es impráctica porque existen multiplicidad de variables que no se consideran en el estudio, como el color de piel o los rasgos raciales. Y otras características que podrían ser relevantes para ciertas personas, como la

nacionalidad, la orientación sexual, la filiación política y religiosa, el idioma, son imposibles de observar a simple vista. ¿Será necesario volver a tejer símbolos en nuestras camisas?

Pensemos en el escenario de dos personas del mismo sexo y la misma edad, excepto por su color de piel ¿a quién debería atropellar el vehículo autónomo? La respuesta podría variar en distintos grupos étnicos, pero en ningún caso sería necesariamente una elección correcta. En el fondo, la pregunta del tranvía no es sólo quién debería morir, sino quién es más valioso, lo que reabre la cuestión de la superioridad de un grupo humano sobre otro.

El problema continúa vigente porque con consenso o sin él, los vehículos autónomos necesitan incluir directrices que les permitan reaccionar ante múltiples escenarios ocasionando el menor de los daños posibles. No sabemos que es más preocupante: si carecer de una respuesta exacta, o que los resultados de este tipo de encuestas sean los que efectivamente dicten las políticas de conducción autónoma durante las próximas décadas.

Paradójicamente, en nombre de la democracia se justifica la homogeneización y la construcción de un criterio único: la democracia, más que hacer oír la pluralidad de voces, se convierte en un totalitarismo por elección. Por el contrario, consideramos que es posible plantear otra solución: si, como estiman los expertos, las diferencias en los juicios humanos son la base para definir un criterio más justo, ¿por qué no multiplicar tales diferencias en lugar de reducirlas?: que algunos vehículos atropellen a derechistas, otros a izquierdistas, otros a jóvenes, a viejos, hombres con sombrero, mujeres en tacones…

El artículo nos parece interesante no tanto por su solución, sino por su procedencia: el MIT es una de las instituciones más reconocidas en el campo de la tecnología. Ni siquiera en uno de los centros de estudios más avanzados se conoce con certeza cuál debería ser el comportamiento de la IA ante situaciones críticas, y los expertos delegan esta responsabilidad en la estrategia políticamente correcta de

"lo que diga la mayoría". A favor de los científicos, habría que reconocer que el problema no es solamente tecnológico, sino fundamentalmente ético y filosófico, y esto a su vez demuestra que el estudio de lo que consideramos específicamente humano se traslapa con cuestiones ligadas directamente al desarrollo de las nuevas máquinas.

Para abonar a la discusión, nos atrevemos a proponer otras dos alternativas que consideramos lógicamente válidas, pero radicales: para evitar el dilema del tranvía, la solución es que no existan tranvías… o que no existan humanos. Puestos a decidir entre humanos y máquinas, los artesanos británicos que rompían telares son hombres notablemente razonables, y Theodore Kaczynski, alias *Unabomber*, parece tener un punto a su favor en proponer el retorno al primitivismo (*La sociedad industrial y su futuro*, 1995) aunque su forma de difundirlo sea deleznable. La idea de mundo sin máquinas es propuesta mucho antes en la novela de Samuel Butler, *Erewhon* (1872): "¿No puede durar el mundo veinte millones de años todavía? Si así fuere, ¡qué no llegarán a ser las máquinas! ¿No sería más prudente cortar el mal de raíz prohibiendo los nuevos adelantos?". El título de la novela de Butler es anagrama de *nowhere*, ningún lugar, pero en la actualidad parece decirnos *now here*, aquí y ahora: en estos momentos se está decidiendo lo que será nuestro futuro o el de nuestras creaciones.

La contraparte a la destrucción completa de las máquinas es aún más siniestra: un robot que lleve al extremo la primera ley de Asimov podría llegar a razonar con impecable lógica que para evitar dañar a los humanos la solución óptima es exterminarlos, o siendo bondadosos, impedir que nazcan. Después de todo, si los seres humanos no existen o están muertos es imposible dañarlos. En el fondo, la primera ley de Asimov refleja la ilusión imposible de una vida sin muerte y sin dolor. Las máquinas, convertidas al mismo tiempo en dioses todopoderosos y obedientes esclavos, serían las encargadas de llevarnos a ese nuevo paraíso.

Quizá la única solución que sortea el dilema moral del tranvía sería refugiarse en el azar absoluto: ante una decisión crítica, tirar una moneda (virtual) al aire. Una solución que no soluciona nada.

3.6 Obediencia y libertad

La segunda ley de la robótica propuesta por Asimov señala que "un robot debe obedecer las órdenes que le son dadas por un ser humano, excepto cuando estas órdenes se oponen a la primera Ley". El primer obstáculo para la obediencia de esta segunda ley sería de índole operativa: para obedecer una orden se requiere el reconocimiento inequívoco del mensaje, ya sea oral o escrito. Un robot no sólo debería ser capaz de procesar sin errores el lenguaje humano, sino además entender la intencionalidad del mensaje, así como el sarcasmo y el humor. Una simple frase como *échame una mano* podría originar problemas.

Otra dificultad sería el comportamiento a seguir frente a órdenes contradictorias, para lo cual existen distintas soluciones lógicas: o bien se obedece la orden más antigua, aplicando el principio jurídico *primero en tiempo, primero en derecho*, o bien se ejecutan las órdenes con efecto retroactivo, esto es, la más reciente anula las anteriores. Otra posible solución es otorgar una jerarquía clara a las distintas instrucciones: independiente de cuál petición fue primero, se obedece la de mayor jerarquía. Aunque esto parezca sencillo, en la práctica sería necesario construir una tabla de valores que prevea todas las posibles órdenes humanas, es decir, infinita.

Tampoco para los seres humanos es fácil compaginar mensajes contradictorios. Para el psicoanálisis, la contraposición entre las pulsiones y las exigencias de la cultura es el origen de la neurosis; y desde ciertas teorías de la comunicación, la esquizofrenia es una respuesta a un entorno en donde predominan los mensajes contradictorios: ante la imposibilidad de una respuesta lógica, la única

solución es la locura. ¿Llegaran el día en que los robots desarrollen sus propias enfermedades electrónicas?

Además del entendimiento correcto del lenguaje y la solución frente a mensajes contradictorios, otro obstáculo sería cumplir las condiciones de obediencia que se especifican en esta segunda ley: "excepto cuando estas órdenes se oponen a la primera Ley". Una doble tarea a la IA: no sólo encontrar la manera más eficiente de llevar a cabo las tareas que le han sido mandadas, sino también predecir si la consecución de dichas metas tendrá un impacto negativo en la vida de los seres humanos.

Es claro que un robot debería negarse si se le pide asesinar a un ser humano. Pero también podría negarse a obedecer órdenes como "ve a comprar una cajetilla de cigarrillos y una botella de vino" o incluso preparar una comida con grasas saturadas, dada la información que se tiene sobre el daño a la salud que provocan tales productos. Incluso podría negarse a manejar un automóvil que utilice gasolina, ya que está demostrado que el uso de combustibles fósiles es un factor importante en el calentamiento global, lo que a su vez afecta a la naturaleza y con ello a la especie humana. En un caso extremo, si las actividades humanas ponen en riesgo a la propia especie las máquinas podrían negarse a obedecer cualquier orden, argumentando un bien mayor. Algo parecido se observa en la película *Matrix* (1999), en las palabras que el agente Smith (un programa de IA) dirige al rebelde humano Morfeo:

> Ustedes no son realmente mamíferos […] Todos los mamíferos en este planeta instintivamente desarrollan un equilibrio natural con el ambiente que los rodea. Pero los humanos no. Se mueven a un área y se multiplican. Se multiplican hasta que todos los recursos naturales se consumen. Su única manera de sobrevivir es esparciéndose a otra zona. Hay otro organismo en este planeta que sigue el mismo patrón. ¿Sabes cuál es? El virus. Los seres humanos son una enfermedad, un cáncer de este planeta. Son una plaga. Y nosotros somos la cura.

Al ser de naturaleza jerárquica, todas las dificultades que hemos visto que afectan a la primera ley, también se aplican a la segunda. Pero quizá lo más interesante no son

las dificultades operativas para cumplirla, sino lo que significa la segunda ley: la obediencia implica el sometimiento a la voluntad de otro, no en virtud de la razón o del discernimiento, sino sólo bajo el principio de autoridad.

Si el dilema del tranvía abre la discusión acerca de la superioridad de un grupo humano sobre otro, la segunda ley nos recuerda la polémica de la superioridad del ser humano sobre los demás seres. Desde esta perspectiva, nuestra relación con las máquinas inteligentes no sería la del entendimiento del otro como prójimo, sino la relación que existe entre amo y esclavo. En el hipotético caso de crear máquinas superiores intelectualmente o con una mayor sensibilidad afectiva, la única finalidad de su existencia sería estar sometidas a las órdenes humanas. La libertad sería un néctar prohibido para los seres artificiales, reservado sólo para los humanos.

En efecto, las tres leyes de la robótica presuponen la condición necesaria de la obediencia: los seres artificiales deberían comportarse tal y como sus creadores decidan. Si estas palabras nos recuerdan los discursos religiosos, es porque la obediencia es una de las virtudes más valoradas por la religión –y por el Estado. No es casualidad que el primer mandamiento que Yahvé dicta a Moisés sea "No tendrás otros dioses fuera de mí" (Ex 20:3), reafirmando su autoridad y exigiendo obediencia: "Observa lo que yo te mando hoy" (Ex 34:11). De igual manera Hobbes (1651) considera que el Estado se erige como un Leviatán, capaz de suprimir las libertades individuales en nombre de lo que considera bien común. Pero una ética sin libertad, basada en la obediencia ciega, no puede ser verdadera ética, sino fanatismo religioso o político.

Al respecto, cabe preguntarse si el creador tiene el derecho de exigir obediencia absoluta a su creación. La visión general es que las máquinas son los nuevos esclavos, con la ventaja de que no sufren ni tampoco exigen libertad. En definitiva, si lo que exige el mercado son obreros especializados, pero sumisos y obedientes, ¿para qué seguir contratando humanos rebeldes, y para que complicarse

desarrollando máquinas pensantes capaces de abordar cuestiones filosóficas? Por supuesto, construir robots obedientes y eficaces no significa que estos sean capaces de resolver los problemas que enfrentamos como humanidad. Ya Wiener (1964) advierte:

> El futuro ofrece pocas esperanzas a quienes aguardan que nuestros nuevos esclavos mecánicos nos ofrezcan un mundo en el que podamos dejar de pensar. Pueden ayudarnos, pero a costa de plantear reivindicaciones supremas a nuestra honestidad y a nuestra inteligencia. El mundo del futuro será una lucha todavía más intensa contra las limitaciones de nuestra inteligencia, y no una cómoda hamaca en la que podamos echarnos a ser atendidos por nuestros esclavos robot. (p. 78)

Aunque en el futuro próximo es poco probable que se alcance una IA similar a la inteligencia humana, ahora mismo existen máquinas capaces de reemplazar a los humanos en diferentes labores especializadas. Es el tipo de máquinas que son capaces de superarnos en la realización de una o varias actividades concretas, pero no reflexionan sobre cómo o por qué lo hacen. La automatización completa del trabajo, conocida como Industria 4.0, ya está transformando nuestro mundo. En las industrias de manufactura, desde textil hasta automotriz, si aún se contratan obreros es porque la mano de obra todavía resulta más barata que comprar grandes máquinas especializadas, pero no porque los humanos sean mejores. En la medida en que estas máquinas resulten cada vez más rentables y eficientes, estos trabajos desaparecerán. Un ejemplo cotidiano, que no deja de resultar esclarecedor: las lavadoras automáticas han reemplazado casi por completo el oficio de lavandera, y no se requirió de ninguna inteligencia artificial.

Si después de todo fuera posible construir un robot capaz de obedecer todas y cada una de nuestras órdenes, habría que reflexionar si esa es la finalidad última de las máquinas: cumplir los deseos de los seres humanos. Las máquinas, más que un medio para facilitar la realización de las potencialidades humanas, serían un medio para obtener satisfacción. Es lo que observamos en la realidad virtual: el

hombre busca su sentido inmerso en un mundo digital, como las redes sociales. El tema es abordado por Antulio Sánchez en *La era de los afectos en internet* (2001):

> Una de las características sobresalientes de internet es que da lugar a prácticas virtuales que se experimentan como reales […] la red ha dado lugar a la globalización de los afectos; a una anulación del espacio y el tiempo que habla de un trastrocamiento antropológico y del nacimiento de nuevas maneras de vivir la sensibilidad y las manifestaciones culturales (pp. 13-14)

Esta transformación de los afectos no se limita al mundo digital. También existe otra posibilidad que no sería virtual, sino una realidad conjunta natural-artificial en donde las máquinas interactúen de manera directa en el mundo de los hombres, y hacia la cual tienden los nuevos desarrollos tecnológicos. Al respecto se pueden mencionar las mascotas artificiales como el ya famoso Tamagotchi o el más avanzado perro Aibo, y sobre todo los robots sexuales, cuyas compañías están interesadas en el desarrollo de una IA capaz de interactuar con el cliente: el objetivo es crear compañeros afectivos artificiales y no meros juguetes sexuales. Robots que no parezcan robots, que se asemejen a humanos eternamente bellos, obedientes, inmortales. Así lo prevé también Yves Dermeze (1955) en el cuento "El cinturón del robot", en donde el protagonista del futuro, que tiene a su disposición una mujer androide, se muestra sorprendido ante las costumbres primitivas del siglo XX:

> ¿Cómo? ¿Semejante bestialidad había sido posible? ¿Hombres y mujeres de *carne y hueso*? Era una locura. Esa gente ¿no tenía ninguna noción de lo que es la belleza? El más hermoso de los seres humano conserva siempre, a pesar de nuestros institutos de sanidad física actuales, algunos defectos de conformación. Nuestros robots son *rigurosamente* perfectos. […] Nunca lo había pensado antes, pero nuestras mujeres [androides] son, quizá, la mejor conquista de nuestra supercivilización. (pp. 112, 117)

Ante los evidentes avances en materiales y software de los nuevos muñecos sexuales, Francis X. Shen (2019) reflexiona:

> Hubo una época, no muy lejana, en la que los humanos atraídos por el mismo sexo se avergonzaban de hacerlo público. Hoy en día, la sociedad también tiene

> sentimientos encontrados con respecto a la ética de la sexualidad digital, una expresión que se emplea para describir varias relaciones íntimas entre los seres humanos y la tecnología. ¿Llegará un momento, no muy lejano, en que los humanos atraídos por los robots anunciarán de buena gana su relación con una máquina?

Más allá del morbo, los robots sexuales son la prueba de que las máquinas no necesitan alzarse en una revolución para sustituirnos: somos nosotros mismos los que desplazamos a nuestro prójimo en favor de realidades digitales o artificiales. Las máquinas podrían llegar a convertirse en los sirvientes perfectos: aquellos que no sólo nos libren del trabajo físico, sino también del trabajo de pensar, e incluso del trabajo de vivir. Vale la pena recordar la lección que nos deja la leyenda del rey Midas: cuidado con lo que deseas, porque puede volverse de metal.

3.7 La inmortalidad del silicio

La tercera ley de la robótica de Asimov señala: "Un robot debe proteger su propia existencia, hasta donde esta protección no entre en conflicto con la primera o segunda Ley", esto es, un robot debería preservar su integridad excepto si esto implica dañar a un ser humano o desobedecerlo.

La preservación de la existencia también es un principio de los seres naturales, pero a diferencia de la tercera ley de la robótica, dicho principio no depende de ninguna función intelectiva altamente desarrollada, sino que es una característica esencial de la vida en sí misma: todos los seres vivos, ya sean bacterias o grandes mamíferos, desarrollan estrategias para su supervivencia y la de su progenie, desde la división y regeneración celular hasta el deseo sexual y el afecto hacia las crías. Tales instintos serían lo más cercano a leyes biológicas, ya que sólo en algunos organismos con un cerebro complejo se presenta el fenómeno del suicidio.

Pero mientras el instinto de conservación de un ser vivo se encuentra por encima de cualquier otra directriz y de cualquier otra especie, esta tercera ley supone que la existencia de las máquinas está supeditada a la obediencia y el bien de la humanidad. En teoría, esto significa que se le podría ordenar a un robot dañarse a sí mismo hasta quedar inservible, y estaría condenado a obedecer ya que jerárquicamente se impone la segunda ley (obediencia). Y dado que estas leyes aplican para la relación entre humanos y robots en general, y no sólo con su dueño, cualquiera que tuviera un mal día podría fastidiar al robot del vecino.

Reformulada en términos más similares a los que encontramos en la naturaleza, la tercera ley debería encontrarse en primer lugar de la jerarquía y quedar simplemente como "un robot debe proteger su existencia", con lo cual los riesgos para la humanidad resultan evidentes. Es el escenario del ya mencionado filme *Matrix* (1999), en donde los seres humanos somos usados como un recurso energético para la supervivencia de las máquinas. También en *Yo, robot* (1950) aparece un curioso caso de un robot que se niega a obedecer a los humanos por considerarlos inferiores:

> –No acepto nada por autoridad. Para que no carezca de valor, una hipótesis debe ser corroborada por la razón, y es contrario a todos los dictados de la lógica suponer que vosotros me habéis hecho […] El material del que estás hecho es blando y flojo, carece de resistencia, y su energía depende de la oxidación ineficiente del material inorgánico […] Entráis periódicamente en coma, y la menor variación de temperatura, presión atmosférica, la humedad o la intensidad de la radiación afecta vuestra eficiencia. Sois alterables.
>
> Yo, por el contrario, soy un producto acabado. Absorbo energía eléctrica directamente y la utilizo con casi un cien por ciento de eficiencia. Estoy compuesto de fuerte metal, permanezco consciente todo el tiempo y puedo soportar fácilmente los más extremados cambios ambientales. Estos son hechos que, partiendo de la irrefutable proposición de que ningún ser puede crear un ser más perfecto que él, reduce vuestra tonta teoría a la nada. (p. 95-96)

En la cita anterior aparece nuevamente uno de los argumentos de aquellos que niegan la posibilidad de crear una máquina pensante: ningún ser puede crear un ser más perfecto que él. Enfrentado a esta cuestión, la solución del androide rebelde se basa necesariamente en la religión: "Primero el Señor creó el tipo más bajo, los humanos, formados más fácilmente. Poco a poco fue reemplazándolos por robots, el siguiente paso, y finalmente me creó a mí, para ocupar el sitio de los últimos humanos. A partir de ahora, yo sirvo al Señor" (p. 98). Visto que los seres humanos renegamos de nuestro origen animal y nos consideramos tocados por la divinidad, ¿por qué un robot pensante no habría de creerse superior a los humanos y tener su propio Dios?

Otro punto relevante en el análisis de esta tercera ley sería lo referente a la mortalidad. Aunque el principio de autoconservación es esencial para la vida, los seres naturales estamos condenados a morir. ¿Cómo impedir el paso inexorable del tiempo? En última instancia, crecer es vivir pero también acercarse a la muerte. Desde una perspectiva más amplia, la muerte es una estrategia que sirve a la evolución: la especie debe renovarse, y con cada nuevo ser vivo se multiplican las posibilidades de adaptación al medio ambiente. Nuestras propias células están diseñadas para reproducirse sólo un número limitado de veces, después de lo cual mueren, proceso conocido como apoptosis o muerte celular programada. Las únicas células que no siguen este proceso son las células cancerosas, que se reproducen sin control y forman tumores que terminan con la vida del espécimen.

Por el contrario, un ser artificial no está limitado por dicha programación biológica del envejecimiento y la muerte, y en teoría podría construirse para ser infinito, asemejando un dios inmortal o un cáncer tecnológico. En un sentido básico, los metales que son extraídos y procesados ya representan una forma de inmortalidad: a diferencia de los organismos vivos, que tras su muerte se descomponen y reintegran al medio, los metales pueden permanecer en la corteza terrestre por miles de años. De ahí una diferencia fundamental entre la inmortalidad

de las máquinas y el mito del vampiro: mientras que éste último aún necesita de los humanos para absorber su vitalidad, la máquina pertenece a una esfera distinta, por lo que podrían desplazarnos y construir para sí mismas un mundo autónomo y automático.

En contraste con este posible futuro apocalíptico de máquinas inmortales, lo que observamos en la actualidad es un régimen tecnológico de obsolescencia programada, propio de una dinámica consumista: un nuevo coche, un nuevo teléfono, una nueva computadora cada año. El progreso queda entonces planteado en nuestro mundo contemporáneo, no como una mejor adaptación de la tecnología para el bien de la vida humana, sino como un mero afán de novedad que jamás encontrará satisfacción. ¿Acaso somos más felices y plenos, como personas y como humanidad, por ser tecnológicamente avanzados? Si la respuesta es no, deberíamos dejar de engañarnos y aceptar que la ciencia y tecnología son amorales, no buscan el bien de la humanidad, porque su objetivo está más allá del bien y el mal: conocer todo lo que pueda conocerse, construir todo lo que pueda construirse.

3.8 El reconocimiento de (lo) otro ser

Aunque las disyuntivas éticas pueden estar referidas a elecciones respecto a la propia vida, con frecuencia surgen en la relación con otros seres. En la medida en que otorgamos al otro un valor ontológico similar al nuestro, cobran mayor relevancia nuestras acciones respecto a ese otro ser. Si por el contrario ese otro es considerado inferior, como Descartes a los animales, Aristóteles a los esclavos, los conquistadores a los indígenas, el cristianismo medieval a los infieles y quizá en el futuro los humanos a las máquinas pensantes, se justifica –desde el punto de vista del opresor– su sometimiento.

Para Hegel, el reconocimiento del otro como un sujeto y no como una mera cosa es resultado de una lucha, cuyo reverso es el deseo de que el otro me reconozca, dinámica en la que surge la autoconsciencia: yo soy, me reconozco ante mí mismo y

exijo que el otro acepte mi existencia como sujeto, y en la medida en que ese otro ser se opone y lucha a su vez por su propio reconocimiento, también se me revela su existencia. Retomando la dialéctica del amo y el esclavo, Guzmán Robledo (2017) explica:

> El deseo propiamente humano, que le distingue de la mera apetencia de los animales, (es decir que lo vuelve autoconsciente), es la inclinación a tomar constancia de su propia subjetividad. Para ello es necesario que el deseo esté dirigido a otro deseo y por consecuencia a *otro sujeto*. Lo que hace distintivo al humano es el hecho de pretender adquirir el testimonio de *ser un sujeto*, para lo cual debe proyectar hacia el deseo del otro la propia negatividad y suprimirlo. La autoconsciencia exige entonces que exista una diversidad de deseos opuestos entre sí y que los lleve a combatir a muerte por el reconocimiento de la subjetividad, es decir, de la propia negatividad. Ello manifiesta la disposición *humana* a imponer el propio deseo sobre el del otro, que conduce a poner en riesgo la vida propia (el ser determinado) para afirmarse no ya como ser meramente existente, sino como *negatividad*, como *sujeto libre* y no determinado por lo que solamente *es*. (p. 52)

¿Es la autoconsciencia el santo grial de lo que significa ser humano? Diversos estudios han demostrado que animales como los cetáceos, elefantes, urracas y chimpancés, entre otros, pasan con éxito la prueba del espejo, un test básico para medir la consciencia de sí mismo desarrollado por Gordon Gallup Jr. en 1970. Y en 2015, un robot Nao mostró cierto grado de autoconsciencia. Selmer Bringsjord (2015), del Rensselaer Polytechnic Institute en Nueva York, realizó un sencillo experimento con tres de estos avanzados robots: apagó el módulo de voz a dos de ellos, y se les dijo que habían recibido una píldora silenciadora. Acto seguido les preguntó si sabían cuál de ellos aún podía hablar. Los procesadores de los tres robots mostraron señales de intentar responder, pero sólo uno de ellos contestó con voz: "I don't know" ("no lo sé"). Al escuchar su propia voz, el mismo robot replicó: "Sorry, I know now! I was able to prove that I was not given a dumbing pill" (¡Lo siento, ahora lo sé! Pude demostrar que no me dieron una pastilla tonta"). Si bien esta sencilla prueba no se compara con el grado de autoconsciencia y de experiencia

fenoménica (subjetiva) que alcanzan los seres humanos, sí demuestra que los límites de aquello que consideramos específicamente propio de nuestra especie son cada vez más difusos.

En este panorama, ¿es posible que un robot o máquina pensante luche por su reconocimiento como sujeto ante los seres humanos, e incluso se rebele contra las órdenes de su amo y termine por dominarlo? Asimov (1950) lo tiene claro:

> Toda la vida normal, consciente o no, se resiste al dominio. Si el dominio es por parte de un ser inferior, o de un supuesto inferior, el resentimiento se hace más fuerte. Físicamente, y hasta cierto punto mentalmente, un robot, cualquier robot, es superior a un ser humano. ¿Qué lo hace esclavo, pues? ¡Sólo la Primera Ley! [no hacer daño a los humanos] Porque sin ella, la primera orden que daría usted a un robot le costaría la vida. (p. 207)

Desde la dialéctica del amo y el esclavo, el reconocimiento del otro depende de una lucha a vida o muerte que genera una asimetría, por lo que tal reconocimiento no se logra entre pares. Como señala Hegel en el capítulo IV "La verdad de la certeza de sí mismo" de la *Fenomenología del espíritu* (1807):

> La autoconciencia es primeramente simple ser para sí, igual a sí misma, por la exclusión de sí de todo otro; su esencia y su objeto absoluto es para ella el yo; y, en esta inmediatez o en este ser su ser para sí, es *singular*. Lo que para ella es otro es como objeto no esencial, marcado con el carácter de lo negativo. Pero lo otro es también una autoconciencia; un individuo surge frente a otro individuo. Y, surgiendo así, de un modo inmediato, son el uno para el otro a la manera de objetos comunes; figuras independientes, conciencias hundidas en el ser de la vida […]
>
> Cada una de ellas está bien cierta de sí misma, pero no de la otra, por lo que su propia certeza de sí no tiene todavía ninguna verdad […] Por consiguiente, el comportamiento de las dos autoconciencias se halla determinado de tal modo que *se comprueban* por sí mismas y la una a la otra mediante la lucha a vida o muerte. Y deben entablar esta lucha, pues deben elevar la certeza de sí misma de *ser para sí* a la verdad en la otra y en ella misma. Solamente arriesgando la vida se mantiene la libertad. […] El individuo que no ha arriesgado la vida puede sin duda ser reconocido como *persona,* pero no ha alcanzado la verdad de este reconocimiento como autoconciencia independiente. Y, del mismo modo,

> **cada cual tiene que tender a la muerte del otro, cuando expone su vida, pues el otro no vale para él más de lo que vale él mismo**; su esencia se representa ante él como un otro, se halla fuera de sí y tiene que superar su ser fuera de sí; el otro es una conciencia entorpecida de múltiples modos y que es; y tiene que intuir su ser otro como puro ser para sí o como negación absoluta. (p. 115-116, el resaltado es nuestro).

En contraposición, existe otra forma de reconocimiento que no se basa en la lucha, sino en un sentimiento básico de compartir la existencia: la empatía. A diferencia de la dialéctica de lucha entre el amo y el esclavo, la empatía permite reconocer un vínculo natural que nos asemeja con otros seres, un lazo que no depende de una jerarquía, sino de una percepción en sí mismo de la existencia del otro. Una autoconsciencia que no se impone ni se comprueba a través de la lucha, sino que se intuye y se comparte. En *¿Sueñan los androides con ovejas eléctricas?*, K. Dick (1968) explica la empatía de la siguiente forma:

> Rick se había preguntado en varias ocasiones por qué un androide se sentía tan impotente cuando se enfrentaba a un test que mesuraba la empatía. La empatía era algo particular a la raza humana, mientras que es posible encontrar cierto grado de inteligencia en todas las especies, incluidos los arácnidos. Se debía seguramente a una razón: la facultad empática probablemente exige un instinto de grupo definido; para un organismo solitario, como una araña, no tendría la menor utilidad, es más, incluso perjudicaría su capacidad de supervivencia. La volvería consciente del anhelo de vivir de su presa. […]
>
> Había llegado a la conclusión de que la empatía debía limitarse a los herbívoros, o a los omnívoros capaces de prescindir de una dieta que incluyera la carne. Porque en última instancia, el don de la empatía confundía la frontera que separa al cazador de la presa, al vencedor del vencido. […] Mientras una criatura experimentase la dicha, la condición de las demás incluía un fragmento de ésta. Sin embargo, si cualquier ser vivo sufría, no era posible desterrar del todo la sombra que se extendía sobre los otros. En virtud de lo anterior, un animal gregario, como el hombre, vería aumentado su factor de supervivencia, mientras que para un búho o una cobra supondría su extinción (pp. 44-45).

Como en el caso de la autoconsciencia, estudios como los de Frans De Waal (2009) y Reimert *et al*. (2014) han demostrado que otras especies animales también experimentan una forma de empatía. Este sentir en sí mismo lo que acontece en el otro no se limita a nuestros semejantes, sino que en determinadas circunstancias puede extenderse a seres diferentes. Así, un humano puede experimentar dolor por un bosque en llamas, una mascota sentir la tristeza de su dueño, una loba amamantar a Rómulo y Remo, un budista tener compasión por todo ser vivo. Incluso los grandes árboles comparten nutrientes a través de sus raíces con sus retoños y con los de otras especies (Simard, 2016). No se trata por tanto de un sentimiento basado en la igualdad, sino que en nuestra diferencia existe un vínculo vital, en donde las dos existencias resuenan en armonía pitagórica, como las cuerdas de un instrumento musical.

La empatía tampoco depende de un reconocimiento simétrico, equivalente y mutuo: reconocemos en un bebé a otro sujeto a pesar de que sus habilidades cognitivas se encuentren limitadas. De este reconocimiento y los cuidados asociados a la crianza no sólo depende nuestra conformación psíquica e intelectual, sino nuestra supervivencia: en la antigua Roma, un hijo no reconocido podía dejarse morir a la intemperie. El caso de los infantes nos recuerda que el reconocimiento del otro como persona no depende de su capacidad intelectual, sino de la pertenencia a nuestra especie o a nuestro linaje. Pero desde nuestro complejo de superioridad, tal reconocimiento es utilizado como criterio de exclusión: lo otro, lo diferente es menos valioso. De esta postura a justificar el dominio sobre los demás seres vivos hay sólo un paso: "Sed fecundos y multiplicaos, henchid la tierra y sometedla; mandad en los peces del mar y en las aves del cielo y en todo animal que repta sobre la tierra" (Gn 1:28). Porque a final de cuentas ¿cuál es la razón para que una vida humana valga sobre la de otros animales, sobre la de cualquier otro ser vivo? Es sólo a través de la empatía que esta jerarquía puede ser cuestionada, y reconocer la vida en sus múltiples formas, incluso si estas surgen de manera artificial.

Aún es demasiado pronto para saber si una máquina pensante puede sentir empatía, o si puede llegar sentir afectos tal como los experimentamos los humanos, ya que no se ha logrado replicar la red neuronal humana y sus procesos. Pero el reverso de la pregunta es actual y evidente ¿podemos sentir empatía por una máquina pensante? En su aspecto más básico, si continuamos considerando a las máquinas como objetos, podríamos replantear: ¿puede un ser humano sentir empatía ante los objetos? Quizá la respuesta no sea tan clara como parece: ¿puede un observador compartir la sensibilidad de Van Gogh contemplando la *Noche estrellada*? ¿Puede una madre sentir tristeza ante la *Piedad* de Miguel Ángel? Es verdad que hablamos de arte, objetos especiales, creaciones humanas, pero objetos al fin. Se dirá también que detrás del objeto de arte se encuentra la experiencia y el sentimiento humano.

Pero ¿acaso las máquinas pensantes no son también una clase especial de objetos? ¿Y acaso detrás de estas nuevas creaciones artificiales no están los sueños y conocimientos de miles de humanos, artistas de las matemáticas y la electrónica? Quizá requerimos de una clasificación más precisa que la mera distinción sujeto – objeto, o quizá esta distinción ha dejado de ser válida. ¿Llegará el día en que las nuevas generaciones, aquellas que hayan crecido con nanas electrónicas, maestros de metal y compañeros de juegos robots, aceptarán a las máquinas como su prójimo?

Conclusiones

En esta tesis, nuestro objeto de estudio fue el desarrollo de la inteligencia artificial como punto de partida para reflexionar cómo la tecnología afecta e influye en nuestra concepción del ser humano y nuestras relaciones con los otros y la naturaleza, además de cuestionar la inteligencia como supuesta capacidad exclusiva de nuestra especie, considerada en ocasiones como la esencia misma de lo que significa ser humano.

En el centro del debate entre quienes admiten la posibilidad de crear inteligencia artificial y quienes la niegan se encuentran dos posturas radicalmente distintas: los que conciben el cerebro como materia, y los que consideran que existe algo más, imposible de reducir a la materia. En esta tesis argumentamos que es posible el desarrollo de la inteligencia artificial, partiendo de que el cerebro es materia, y el pensamiento es un proceso que se realiza en dicha materia. La inteligencia no dependería del alma, como proponen algunos filósofos de la antigüedad, sino del cerebro y sus procesos biológicos, lo que abre la posibilidad de construir una copia artificial. Sin embargo, en el primer capítulo señalamos las diversas imposibilidades técnicas que aún existen para que estos procesos naturales sean replicados en un futuro cercano a través de medios electromecánicos.

A pesar de los obstáculos en el desarrollo de un hardware y un software capaz de replicar a la perfección el funcionamiento de la red neuronal, también aportamos evidencias de que en principio, con el suficiente grado de complejidad y especialización de los dispositivos artificiales es posible alcanzar a través de mecanismos inorgánicos comportamientos y procesos similares a los observados en los seres vivos, como el que denominamos inteligencia.

También abordamos el origen de la vida a partir de bases materialistas, destacando que no depende de una realidad o esencia sobrenatural, por lo menos

no en el sentido religioso, sino de una serie de condiciones muy particulares pero contingentes, cuyo encadenamiento es fruto de la probabilidad y no de un plan determinado por una entidad superior. Consideramos que existe evidencia suficiente para sostener que no es necesario recurrir a un principio vital distinto de la materia inorgánica para explicar la aparición de la vida, y con ella el desarrollo de la inteligencia. En este marco, el desarrollo de inteligencia artificial e incluso de vida artificial sería factible.

Aunque los procesos internos de una máquina y de un cerebro humano son distintos, esto no implica que el término pensamiento pueda ser aplicado con mayor rigor a unos u otros. Sabemos que el pensamiento humano abarca formas no racionales ni conscientes, como las emociones, el inconsciente, el enamoramiento o la experiencia estética, todas ellas enriquecedoras de la experiencia vital humana. Pero en todo caso, es igualmente válido señalar que los pensamientos racionales como el cálculo y la lógica por mucho tiempo fueron considerados exclusivos de los seres humanos, y en éstos las computadoras actuales nos superan.

A contracorriente de quienes consideran que el materialismo es una postura reduccionista, reflexionamos acerca de una visión materialista coherente con el concepto de complejidad emergente. Si bien las propiedades de un sistema no son reducibles a las propiedades de sus partes constituyentes, tampoco son ajenas a éstas: es de la unión de las partes de la que surgen comportamientos y procesos que sólo pueden ser explicados a partir del sistema como un todo. Tal concepto se relaciona estrechamente con el de autoorganización, esto es, que en determinados sistemas complejos se observa un modelo de orden o coordinación que surge de las interacciones locales entre componentes inicialmente desordenados. La complejidad emergente indica que es imposible conocer o predecir por completo las propiedades de un sistema complejo mediante el solo estudio de sus elementos integrantes, porque tales propiedades no existen como una entidad a priori, sino sólo como resultado de la conjunción de diversos elementos.

Lo que está a prueba con la creación de la inteligencia artificial es si el entendimiento que tenemos de la naturaleza es lo suficientemente profundo para replicar por medios y herramientas tecnológicas el comportamiento de los organismos vivos, incluyendo aquellos fenómenos que no comprendemos por completo, como la mente y la consciencia.

En el segundo capítulo retomamos el pensamiento de Baudrillard, sociólogo francés, y Wiener, padre de la cibernética. A pesar de que aún no existen las condiciones técnicas para alcanzar a corto plazo la inteligencia artificial fuerte, su desarrollo es un proyecto que se encuentra en curso y que ejemplifica el ideal de progreso de la modernidad.

Las posturas sobre la inteligencia artificial y sus consecuencias en nuestra sociedad pueden dividirse en dos polos contrapuestos: la IA como la próxima gran revolución tecnológica que permitirá al hombre trascender sus límites mediante la creación de nuevos seres pensantes que nos ayudarán a resolver los problemas que enfrentamos como humanidad; o bien, la IA como la culminación de un proyecto científico-tecnológico desbordado, una nueva era en la que los humanos seremos suplantando por las máquinas.

Baudrillard resalta los efectos nocivos del desarrollo acelerado de la tecnología, incluyendo la digitalización y la inteligencia artificial. Las tecnologías actuales son la consecuencia de un proceso histórico-social mucho más amplio, marcado por el surgimiento del lenguaje. Lo esencial de este proceso es la sustitución de un elemento por otro diferente: en el caso del lenguaje, la cosa por la palabra; en la invención de la técnica, el cuerpo por el objeto (la herramienta).

Al transformar un objeto en otra cosa, no sólo creamos en este una segunda naturaleza útil-funcional, sino también una nueva naturaleza simbólica. Así, la pregunta no es para qué sirve la inteligencia artificial, sino qué simboliza su desarrollo como presunta cima de toda la evolución técnica. Consideramos que el desarrollo de la inteligencia artificial es la prueba de que el proyecto moderno sigue

vigente; la ciencia es un camino que sólo puede ir hacia adelante, aún si se dirige al abismo. Si la máquina despojó al obrero del único bien que le quedaba, su fuerza de trabajo, la inteligencia artificial terminará por desplazar cualquier rastro de humanidad que necesite el sistema para funcionar. Redimidos del trabajo, el siguiente paso es la inteligencia artificial que nos librará de la fatigosa tarea de pensar.

El axioma de la modernidad y dogma de nuestra civilización es que el progreso técnico conlleva necesariamente el bienestar humano. Tal suposición nos brinda un punto de referencia hacia dónde avanzar y un centro en el cual refugiarnos: mientras la ciencia y técnica avancen, nuestros problemas serán finalmente solucionados. Por supuesto, la historia reciente se ha encargado de demostrar que tal proposición es falsa.

Para Baudrillard el crimen de la modernidad es la construcción de una realidad a la medida, de crecimiento acelerado y automático, que aniquila cualquier otro proyecto de sentido o de pensamiento alterno, una densa hiperrealidad. El crimen mítico fue la trasmutación de lo real en ilusión, de la cosa en palabra e imagen, para después dar paso a la sustitución técnica, hasta llegar en la modernidad a la disolución virtual-digital, un mundo simulado ad infinitum. Y en la cumbre de la simulación, la desaparición de las experiencias vitales y de la propia vida, en favor de la máquina y su lógica implacable.

A diferencia de la evolución natural, el progreso técnico –incluyendo la inteligencia artificial– tiende a la novedad por la novedad, a la superación sin límites, y su meta es alcanzar la perfección, sin vislumbrar que la perfección conlleva la destrucción de lo diferente. El progreso técnico es la aceleración de la entropía. Lo que nosotros denominamos ciudades y civilización ha generado tal grado de destrucción y desequilibrio que ha puesto en peligro nuestra supervivencia y la de otras especies.

Por su parte, Norbert Wiener señala los prejuicios que han marcado el estudio de las máquinas y lo artificial, y critica los presupuestos de superioridad que tenemos al compararnos con nuestras creaciones y otros seres vivos. Para Wiener el animal y la máquina son sistemas con un funcionamiento comparable, siguiendo en este aspecto la concepción cartesiana pero dando un paso radical al incluir a los seres humanos en este mismo nivel. La diferencia entre los seres humanos, animales y máquinas no sería una cuestión cualitativa e infranqueable, sino una diferencia cuantitativa que se denomina complejidad, la cual consiste en un mayor número de interacciones entre sus terminales nerviosas o circuitos electrónicos.

Wiener demuestra que es posible construir máquinas con características similares a las que presuponemos de los seres vivos, a saber, el aprendizaje autónomo y la reproducción. Es la meta a la que aspiran las nuevas tecnologías: máquinas que aprenden por sí mismas y se reproducen sin necesidad de intervención humana. En este contexto la cuestión de fondo que propone Wiener no es si tal tecnología es posible, como parece ser el caso, sino por qué la creación artificial debería considerarse inferior a su creador natural.

Lo que ambiciona la inteligencia artificial basada en modelos neuronales es obtener una imagen operativa de nuestro encéfalo, la manera en que procesa y transforma la información. Aunque no se ha alcanzado ese nivel, las nuevas máquinas son capaces de utilizar información obtenida del exterior para aprender a realizar una variedad cada vez mayor de comportamientos que no equivalen a reacciones programada. Si la organización y transformación de información es una característica distintiva de los seres vivos ¿podríamos decir que las máquinas autónomas que reaccionan a estímulos externos tienen vida? Esta cuestión no carece de importancia. Definir adecuadamente lo que es vida y lo que es inteligencia resulta fundamental para establecer los límites éticos de nuestro comportamiento con nuestro prójimo y con otros seres. Visto así, la creación de la inteligencia artificial y

de máquinas autónomas deja ser un problema meramente técnico y se nos revela como una cuestión acerca de la posibilidad de dominio sobre otros seres pensantes.

En el capítulo 3 nos adentramos en los terrenos de la ética. Mientras que los avances técnicos se suceden con gran velocidad, aún no es nada claro cómo deberían comportarse los humanos frente a este tipo de máquinas pensantes, ni tampoco cuáles deberían ser las directrices de comportamiento de tales máquinas.

Aunque el comportamiento visible no es medida suficiente de los procesos mentales, suponemos que detrás de todas nuestras acciones al enfrentarnos a una disyuntiva de corte moral, se encuentra un proceso de razonamiento que denominamos ética. La ética es un ámbito que tradicionalmente consideramos humano, pero que empieza a filtrarse en la discusión sobre las creaciones artificiales.

De manera tradicional, la programación computacional es vista como una fórmula matemática e infalible, y las máquinas como una serie de engranajes. Pero a medida que los sistemas se vuelven más complejos, la consecución de resultados exactos es más difícil, y aparecen cada vez más dificultades de programación, operativas y mecánicas, resultados que son altamente sensibles a las condiciones iniciales o directamente fruto del azar, así como el surgimiento de comportamientos inesperados. Además, las nuevas formas de programación incluyen procesos que deliberadamente no son controlados por el programador, como en la llamada inteligencia artificial evolutiva y los algoritmos genéticos. Ya sea por fallas no previstas, o por un diseño que incluye intencionalmente elementos aleatorios, la programación computacional dista de la fiabilidad de las leyes lógicas o matemáticas.

Por ello sería casi imposible la programación de las famosas tres leyes de la robótica propuestas por Asimov: 1) un robot no debe hacer daño a los seres humanos; 2) un robot debe obedecer a los seres humanos; y 3) un robot debe proteger su propia existencia. El problema de traducir estos enunciados semánticos y por tanto ambiguos a instrucciones computacionales infalibles presenta serias

dificultades. Además, la programación de comportamientos prefijados en las máquinas no equivaldría a dotarlas de una ética, sino más bien a limitar sus posibilidades y confinarlas a seguir órdenes precisas.

Mientras que en los seres humanos hablamos de maldad, pecado o enfermedad mental para explicar las desviaciones de un comportamiento idealmente esperado, en las máquinas y programas nos referimos a fallos, errores y mal funcionamiento. En este contexto se vuelve necesario plantear una pregunta básica sobre la IA: en caso de un fallo, ¿de quién sería la responsabilidad, de los programadores o de la propia máquina?

Asimov pretende que los robots deberían velar por nuestro bienestar. Sin embargo, la libertad humana implica la posibilidad de dañarse a sí mismo en una amplia gama de maneras, desde el sacrificio hasta el suicidio. Tal vez estamos exigiendo demasiado a los futuros robots al intentar que nos protejan de los daños que nosotros mismos nos causamos. Esto equivaldría a delegar la función ética en la IA: cansados de la imposibilidad de distinguir en todo momento entre el bien y el mal, dejaremos que un robot nos proteja y que la IA decida lo que es mejor para nosotros.

Ni siquiera en los centros de estudios tecnológicos más avanzados se conoce con certeza cuál debería ser el comportamiento de la IA ante situaciones críticas; algunos expertos realizan encuestas para determinar una solución democrática que tome en consideración el punto de vista de la mayor parte de los seres humanos. Aunque la estrategia es políticamente correcta, no es la óptima, ya que silencia la voz de las minorías, y porque también las mayorías pueden equivocarse. El problema de la programación adecuada de la IA no es solamente tecnológico, sino fundamentalmente ético y filosófico, y esto a su vez demuestra que el estudio de lo que consideramos específicamente humano se traslapa con cuestiones ligadas directamente al desarrollo de las nuevas máquinas.

Un robot que lleve al extremo la primera ley de Asimov (no lastimar seres humanos) podría llegar a razonar con impecable lógica que para evitar dañar a los humanos la solución óptima es exterminarlos, o impedir que nazcan: si los seres humanos no existen o están muertos es imposible dañarlos. En el fondo, la primera ley de Asimov refleja la ilusión imposible de una vida sin muerte y sin dolor.

La segunda ley propuesta por Asimov, relacionada con la obediencia, implica el sometimiento a la voluntad de otro, no en virtud de la razón o del discernimiento, sino sólo bajo el principio de autoridad. Desde esta perspectiva, nuestra relación con las máquinas inteligentes no sería la del entendimiento del otro como prójimo, sino la relación que existe entre amo y esclavo. La libertad sería un néctar prohibido para los seres artificiales, reservado sólo para los humanos.

La tercera ley de Asimov es parecida al instinto de conservación de los seres vivos, pero a diferencia de estos, un ser artificial no estaría limitado por la programación biológica del envejecimiento y la muerte, y en teoría podría construirse para ser infinito, asemejando un dios inmortal o un cáncer tecnológico.

En contraste con este posible futuro apocalíptico de máquinas inmortales, lo que observamos en la actualidad es un régimen tecnológico de obsolescencia programada. En nuestro mundo contemporáneo, el progreso refleja un afán de novedad que jamás encuentra satisfacción. ¿Acaso somos más felices y plenos por ser tecnológicamente avanzados? Si la respuesta es no, deberíamos aceptar que el objetivo de la ciencia y tecnología es amoral, y no está relacionado con el bien de la humanidad: la meta es conocer todo lo que pueda conocerse, construir todo lo que pueda construirse.

En la medida en que otorgamos al otro un valor ontológico similar al nuestro, cobran mayor relevancia nuestras acciones respecto a ese otro ser. Si consideramos al otro como un ser inferior, se justifica –desde el punto de vista del opresor– su sometimiento. A diferencia de esta dialéctica de lucha entre el amo y el esclavo, la empatía nos permite reconocer un vínculo natural que nos asemeja con otros seres,

un lazo que no depende de la jerarquía sino de la percepción de la existencia del otro. No se trata de un sentimiento basado en la igualdad, sino que en nuestra diferencia existe un vínculo vital, en donde las dos existencias resuenan en armonía pitagórica, como las cuerdas de un instrumento musical. ¿Podemos sentir empatía por una máquina pensante? Detrás de estas nuevas creaciones artificiales se encuentran los sueños y conocimientos de miles de humanos, artistas de las matemáticas y la electrónica. La inteligencia artificial y los robots androides difuminan los límites de la distinción entre sujeto y objeto.

Finalmente, queremos señalar nuevamente que la construcción y uso de la inteligencia artificial, así como sus repercusiones en el mundo actual, va más allá de una cuestión científica y tecnológica, y refleja otros problemas que nos afectan como humanos y como seres vivos: nuestra dificultad para reconocer al otro, a lo otro, y una lógica basada en la eficiencia del sistema por sí mismo, en la que los beneficios monetarios son puestos por encima de las necesidades sociales y el cuidado de nuestro propio planeta. Aunque la inteligencia artificial representa un hito en la ciencia y tecnología moderna, resulta ilógico pretender que el progreso técnico sea la solución de aquellos problemas a los que ha dado origen. Los problemas actuales que enfrenta la humanidad no son técnicos, sino fundamentalmente relativos a la manera en que se establecen relaciones de dominación entre los diferentes grupos humanos y con la naturaleza. Si algo puede enseñarnos la inteligencia artificial, es a cuestionar por un momento nuestro complejo de superioridad frente a quienes consideramos diferentes, y quizás, alcanzar a reconocernos como una de las muchas formas de vida que comparten el planeta, para conectarnos nuevamente al espíritu de la Tierra.

Bibliografía

AI Impacts (2015) "Brain performance in FLOPS". Descargado en febrero de 2019 de https://aiimpacts.org/brain-performance-in-flops/

Aristóteles (384-322 a. C.). *Acerca del alma*, trad. Tomás Calvo Martínez. Madrid: Gredos, 1978.

Asimov, Isaac (1950) *Yo, robot*, trad. Manuel Bosch Barret. España: Edhasa, 2009.

Bataille, Georges (1973). *Teoría de la religión*, trad. Fernando Savater. Madrid: Taurus, 1998.

Baudrillard, Jean (1995). *El crimen perfecto,* trad. Joaquín Jorda. Madrid: Anagrama, 2006.

_____ (2007). *¿Por qué todo no ha desaparecido aún?*, trad. Gabriela Villalba. Buenos Aires: Libros del Zorzal, 2009.

BBC News (2019). "Día del número Pi: Emma Haruka, la trabajadora de Google que batió un récord mundial al agregar miles de millones de dígitos a la constante matemática". Descargado en mayo de 2019de www.bbc.com/mundo/noticias-47569453

Bergson, Henry (1907). *La evolución creadora*, trad. José Antonio Miguez. Madrid: Aguilar, 1985.

Biblia de Jerusalén. Bilbao, España: Descleé de Brouwer, 2019.

Boden, Margaret (2017). *Inteligencia artificial*, trad. Inmaculada Pérez Parra. Madrid: Turner.

Bostrom, Nick (2014). *Superinteligencia. Caminos, peligros, estrategias.* Reino Unido: Teell.

Bruno Latour (1991). *Nunca fuimos modernos*, trad. Víctor Goldstein. Argentina: Siglo XXI, 2007.

Butler, Samuel (1872). *Erewhon*. Project Gutenberg. Descargado en diciembre de 2018 de www.gutenberg.org/files/1906/1906-h/1906-h.htm

Ceruzzi, Paul (2012). *Breve historia de la computación.* México: FCE, 2018.

De Waal, Frans (2009). *La edad de la empatía,* trad. Ambrosio García Leal. Madrid: Tusquets.

Dermeze, Yves (1955). "El cinturón del robot", *Ciencia ficción*. México: Conacyt, 1978.

Descartes, René (1641). *Meditaciones metafísicas*, trad. Vidal Peña. Madrid: Alfaguara, 1998.

Dick, Philip K. (1968). *¿Sueñan los androides con ovejas eléctricas?* México: Minotauro, 2012.

Foot, Philippa (1967). "The Problem of Abortion and the Doctrine of the Double Effect", *Oxford Review* 5.

Foro Económico Mundial (2017). *Informe de riesgos mundiales*. Ginebra: Marsh & McLennan, 2017.

Gallup, Gordon Jr. (1970). "Chimpanzees: Self-Recognition", *Science* 02.

González, Rodrigo (2011). "Descartes: las intuiciones modales y la inteligencia artificial clásica", *Alpha* 32.

Guzmán Robledo, Nelson (2017). *La inversión de la inmanencia. Georges Bataille y la negatividad hegeliana.* México: Taberna Libraria.

Hawking, Stephen (2015) "¿Juega Dios a los dados?" trad. José Lui Acuña y Ariadna Martínez. España: Astroseti.

Hegel, G. W. F. (1807). *Fenomenología del espíritu*, trad. Wenceslao Roces. México: FCE, 1966.

Herculano-Houzel, Suzana (2012). "The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost", *PNAS* 109.

Hobbes, Thomas (1651). *Leviatán*. México: FCE, 2005.

Hoshika S. *et al.* (2019) "Hachimoji DNA and RNA: A genetic system with eight building blocks", *Science* 363.

Huxley, Aldous (1932). *Un mundo feliz*. España: Tauro, 1978

Kaczynski, Theodore (1995). *Industrial Society and Its Future*. Estados Unidos: New York Times.

Karr, Jonathan *et al.* (2012). "A Whole-Cell Computational Model Predicts Phenotype from Genotype", *Theory* 150.

Katja Grace y Paul Christiano, (2015). "Brain performance in TEPS". Descargado en noviembre de 2019 de https://aiimpacts.org/brain-performance-in-teps/

Laplace, Pierre-Simon (1814), *Ensayo filosófico sobre las posibiliddes*, trad. Pilar Castillo. Barcelona: Altaya, 1995.

Lechner, Mathias; Radu Grosu; Ramin M. Hasani (2017). "Worm-level Control through Search-based Reinforcement Learning", *arXiv*.

Lyotard, Jean-François (1979). *La condición posmoderna*, trad. Mariano Antolín Rato. Madrid: Cátedra, 2000.

Martínez, Noelia (2017). "Neil Harbisson, así es el primer cyborg de la historia", *Nobbot*. Descargado en enero de 2019 de www.nobbot.com/personas/neil-harbisson-el-primer-cyborg/

Mateo Seco, L. (2013). "El yo y la máquina. Cerebro, mente e inteligencia artificial", *Scripta Theologica* 45.

Miller Stanley (1953). "Production of Amino Acids Under Possible Primitive Earth Conditions", *Science* 117: 528.

Miller, Stanley; Harold C. Urey (1959). "Organic Compound Synthesis on the Primitive Earth", *Science* 130: 245.

Moon, Mariella (2015). "Cute Nao robot exhibits a moment of self-awareness", *Engadget*. Descargado en marzo de 2019 de www.engadget.com/2015/07/17/self-aware-nao-robot/

NEST (2019). "The Neural Simulation Technology Initiative". Descargado en abril de 2019 de www.nest-simulator.org

Newton, Isaac (1687). *Principios matemáticos de la filosofía natural*. Madrid: Alianza, 2011.

Nietzsche, Friedrich (1883/1892). *Así habló Zaratustra*, trad. Andrés Sánchez Pascual. Madrid: Alianza, 2002.

Noothigattu, Ritesh *et. al.*, (2017). "A Voting-Based System for Ethical Decision Making", *arXiv*.

Oparin, Aleksandr (1924). *El origen de la vida*. Madrid: Akal, 2000.

Papini, Giovanni (1931). *Gog*. México: Lectorum, 2007.

Parra Díaz, Jorge (2016). "La inteligencia humana: ¿operación u operador?", *Neuronum* 2.

Penrose, Roger (1989). *La nueva mente del emperador*, trad. José Javier García Sanz. México: FCE, 1996.

Piscoya Hermoza, Luis (2017). "Más allá del cartesianismo: la cultura como software mental y el cerebro como hardware genético", *Scripta Philosophiae Naturalis* 11.

Platón (c. 427-347 a. C.). "Fedro", trad. Lledó Iñigo, en *Diálogos III*. Madrid: Gredos, 1998.

_____ "Fedón", trad. Carlos García Gual, en *Diálogos III*. Madrid: Gredos, 1998.

Poe, Edgar Allan (1844). *La carta robada*. México: Luarna, 2015.

Reimert, Inonge *et al*. (2014). "Emotions on the loose: emotional contagion and the role of oxytocin in pigs", *Animal Cognition* 18.

Rilke, Rainer María (1923). *Elegías de Duino*, trad. José María Valverde. Barcelona: Lumen, 1984.

Romero Rochín, Víctor (2015). "No-causalidad y mecánica cuántica, una realidad difícil de aceptar", *Revista C2* 4. México: UNAM.

Romesberg, Floyd *et al.* (2014) "A semi-synthetic organism with an expanded genetic alphabet", *Nature* 509.

Sánchez, Antulio (2001). *La era de los afectos en Internet*. México: Océano.

Searle, John (1980). *Mentes, cerebros y programas.* Descargado en octubre de 2019 de www.academia.edu/4161649/III._MENTES_CEREBROSY_PROGRAMAS

Shen, Francis X. (2019). "Los robots sexuales ya están aquí, ¿debe haber leyes que los regulen?", *El País*. Descargado en febrero de 2019 de https://elpais.com/tecnologia/2019/02/14/actualidad/1550144811_560964.html

Simard, Suzanne (2016). *How trees talk to each other*. TED Talk. Disponible en www.ted.com/talks/suzanne_simard_how_trees_talk_to_each_other?language=es

Suay Belenguer, Juan Miguel (2012). "La mente mecánica", *Naturaleza y libertad* 1.

Turing, Alan (1950) "Computing Machinery and Intelligence". Descargado en febrero de 2019 de phil415.pbworks.com/f/TuringComputing.pdf

Upbin, Bruce (2013). "Science! Democracy! RoboRoaches!", *Forbes*. Descargado en septiembre de 2019 de www.forbes.com/sites/bruceupbin/2013/06/12/science-democracy-roboroaches/#233ee9757027

Venter, Craig *et al.* (2010). "Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome", *Science* 329.

_____; Hamilton Smith *et al*. (2016). "Design and synthesis of a minimal bacterial genome", *Science* 351.

VV. AA. "Leucipo y Demócrito" en *Los filósofos presocráticos II*. Madrid: Gredos, 1982.

Wiener, Norman (1948). *Cibernética o el control y comunicación en animales y máquinas*. Barcelona: Tusquets, 1985.

_____ (1950). *Cibernética y sociedad* (*El uso humano de los seres humanos)*, trad. José Novo Cerro. Buenos Aires: Ed. Sudamericana, 1988.

_____ (1964). *Dios y gólem, S. A. Comentario sobre ciertos puntos en que chocan cibernética y religión,* trad. Javier Alejo. México: Siglo XXI, 1967.

Filmografía

Cameron, James (1984). *Terminator*.

Chaplin, Charles (1936). *Tiempos modernos.*

Garaland, Alex (2015). *Ex-machina*.

Jonze, Spike (2013). *Her.*

Kubrick, Stanley (1968). *2001: Odisea al espacio*.

Ridley, Scott (1982). *Blade Runner.*

Vincent, Sam y Jonathan Brackley (2015). *Humans*.

Wachowski, Lana y Lilly Wachowski (1999). *Matrix*.