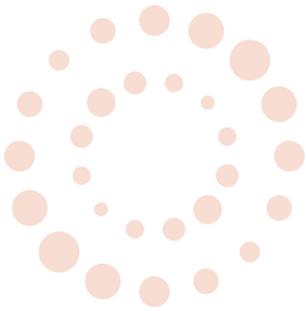
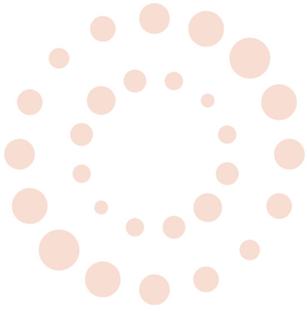


## CAPÍTULO 1

# CONCEPTOS, TÉCNICAS Y CLASIFICACIONES. IMPACTO EN LOS DERECHOS HUMANOS Y LIMITACIONES DE LA INTELIGENCIA ARTIFICIAL



Thomson  
Reuters™



**Thomson  
Reuters™**

# PRELUDIO. BIENVENIDOS A LA ERA DE LA INTELIGENCIA ARTIFICIAL

JUAN G. CORVALÁN

## I. CUANDO EL COYOTE SE TRANSFORMA EN CORRECAMINOS

Convivimos con *más de dos millones* de pequeñísimos animales que están en el colchón, en la alfombra, en muebles tapizados, en la ropa de cama y en libros antiguos. Mi amiga Julieta es alérgica a ellos y por eso los conoce muy bien. Se trata de los ácaros, integrantes de la familia de los arácnidos, con un tamaño que oscila entre 0,2 y 0,5 milímetros. Aunque es casi imposible eliminarlos de tu casa, estas arañitas microscópicas<sup>(1)</sup> tienen algo en común con esta era de la inteligencia artificial. También son invisibles muchos sistemas inteligentes que nos acompañan a los lugares más íntimos a través de computadoras y teléfonos celulares. Este nuevo mundo digital plagado de ácaros artificiales basados en IA, dejarán muy atrás a todos los superhéroes y mutantes de Marvel y DC Comic. Los “héroes”, los “villanos” y los que oscilan por ambas categorías; es decir, casi todos.

La inteligencia artificial (en adelante IA) y la automatización de la que habla todo el mundo, tiene una lógica simple: reemplazar, mejorar e incluso crear nuevas tareas que solo podían ser realizadas por el cerebro humano. Las proezas que se pueden generar en el living de una casa, se basan en ácaros artificiales que nos traen nuevas herramientas y conceptos antiguos que se transforman. Por ejemplo, es muy frecuente encontrar en los medios de comunicación la palabra algoritmo. Aunque se inventaron y se usan hace muchos siglos, esta noción es una de las mayores responsables de todo este tsunami de innovación.

Los algoritmos construyen puentes entre las computadoras, internet, las aplicaciones móviles, nuestros objetivos y necesidades. La unión entre una máquina expendedora, tu chocolate preferido y el dinero, presenta un “detrás de escena” basado en un conjunto reglas, fórmulas para hacer cálculos, o una serie metódica de pasos<sup>(2)</sup>. Son los manuales de instruc-

---

(1) <https://www.muyinteresante.es/naturaleza/preguntas-respuestas/cuantos-acaros-habitan-en-los-colchones-331442837253>.

(2) Ampliar en Penrose, Roger, *La mente nueva del emperador*, Fondo de Cultura Económica, México, 1996, ps. 34-35; Domingos, Pedro, *The master algorithm: how*

ciones del mundo digital que se escriben en idioma informático. Introducís una moneda, presionas un botón y ¡magia!: una golosina que cae es igual a un algoritmo que se ejecuta.

Ahora bien, cuando estamos en una sala de espera y queremos un refresco, ¿consideramos a la máquina expendedora “inteligente”? ¿Es razonable comparar los métodos que utilizan las computadoras para llegar a sus resultados con los complejos procesos biológicos que se dan en el cerebro?

El 15 de octubre de 1937 el genio de Jorge Luis Borges publicó un texto llamado “La máquina de pensar de Raimundo Lulio”. Esta máquina estaba hecha de tres discos concéntricos, divididos en espacios en los cuales podían colocarse distintos atributos. Por ejemplo, “poder”, “verdad” o “sabiduría”, así como también “rojo”, “azul”, “verde”, entre muchas opciones. Al girar los discos, era posible obtener una enorme cantidad de combinaciones de los distintos atributos, pudiendo llegar a muchas respuestas diferentes que hoy en día se podrían analizar desde un punto de vista probabilístico. Para Borges, la máquina creada a fines del siglo XIII no era capaz de un solo razonamiento, siquiera rudimental o sofisticado<sup>(3)</sup>.

Unos años más tarde Alan Turing, el padre de la computación y el inspirador de la película *Código enigma*, se formuló dos preguntas: ¿pueden pensar las máquinas? o ¿pueden las máquinas pensar?<sup>(4)</sup>.

En la versión clásica de la informática, la inteligencia de máquina se asoció a respuestas preprogramadas. Por ejemplo, los algoritmos que están “detrás del telón” de los procesadores de texto. Cuando abre el programa “Microsoft Word”, y luego presiona la letra “a” en su teclado, en el nue-

---

*the quest for the ultimate learning machine will remake our world*, Basic Books, New York, 2015, ps. XVI, 1 y ss.; Harari, Yuval Noah, *Homo Deus*, Debate, ps. 99-107; Boruso Renato, “La ley, el juez, la computadora. Un tema fundamental de la informática jurídica”, *Informática y derecho*, vol. 5, Depalma, 1996, ps. 30 y ss.; Deutsch, David - Ekert, Artur, “Más allá del horizonte cuántico”, *Investigación y ciencia*, noviembre, 2012, p. 79; Palma Méndez, José T. - Morales, Martín Roque, *Inteligencia artificial*, Mc Graw Hill, Madrid, 2011, p. 683; Benítez, Raúl - Escudero, Gerard - Kanaan, Samir - Rodó Masip, David, *Inteligencia artificial avanzada*, UOC, Barcelona, 2013, p. 13.

(3) *Obras completas*, Emecé, Buenos Aires, 2007, t. IV, ps. 391-395.

(4) *Maquinaria computacional e inteligencia*, p. 1. El famoso texto de Turing comienza del siguiente modo: “Propongo considerar la siguiente pregunta: ¿Pueden pensar las máquinas?”, aunque luego el autor propone abordar esta problemática a partir de considerar el juego de la imitación, que consiste en analizar un fenómeno mediante el cual se sustituye la tarea de uno de los integrantes humanos por la de una computadora. Por eso reformula la pregunta del siguiente modo: ¿pueden las máquinas pensar? Texto disponible en: <http://xamanek.izt.uam.mx/map/cursos/Turing-Pensar.pdf> (consultado el 21/07/2023).

vo documento en blanco aparece ese símbolo. Es decir, las posibilidades de potenciar la escritura entre su ordenador y un procesador de texto, en la Edad Media serían consideradas mágicas, imposibles y herejes. Ahora, como nos parece natural poder copiar y pegar miles de letras en menos de un segundo, comenzamos a llamar inteligencia artificial a cuestiones más complejas para nuestros ojos moldeados por esta era digital<sup>(5)</sup>.

Si consideramos algunas analogías, cuentos y metáforas, es posible resumir las cuatro grandes cuestiones que nos introducen en esta trilogía entre humanos, máquinas e inteligencia.

*Primero.* Los aviones y los pájaros vuelan. Unos pueden transportar cientos de personas de un lado a otro, los otros no.

El cerebro extrae, selecciona, recorta y organiza datos e información disponible para tomar decisiones. La IA hace lo mismo, aunque con otros métodos, mayor capacidad de almacenamiento y a más velocidad. Aunque resulta imposible reproducir en máquinas a un órgano tan complejo, hay que considerar que los ingenieros en aviación no copiaron las técnicas de aprendizaje de los pájaros para construir los aviones modernos<sup>(6)</sup>. Por eso es interesante remarcar un aspecto central en esta nueva dinámica: el desarrollo de la IA no consiste, al menos exclusivamente, en imitar el funcionamiento de organismos biológicos.

Cuando pensamos en la inteligencia de ChatGPT, por ahora tenemos que hablar de simulación de comprensión del conocimiento humano a partir de información publicada en internet. Matematizar probabilísticamente corpus de textos, no es equivalente a comprender causal y emocionalmente el mundo a partir de la experimentación basada en nuestros

(5) Sobre todas estas cuestiones, ampliar en: Palma Méndez - Morales Roque, *Inteligencia artificial*, McGraw-Hill Interamericana de España SL, ps. 3 y 7; en similar sentido, Bourcier, Danièle, *Inteligencia artificial y derecho*, UOC, Barcelona, 2003, p. 51; Deyi Li y Yi Du, *Artificial intelligence with uncertainty*, 2a ed., CRC Press, 2017, p. 1; *Inteligencia artificial*, p. 3; en similar sentido, Ruiz - Quevedo - Galipienso - Pardo - Lozano Ortega, *Inteligencia artificial*, Thomson, Madrid, 2003, p. 4; Kaplan, Jerry, *Absentarse humanos*, Teell Editorial, SL, p. 1; Winston, Patrick H., *Inteligencia artificial*, Pearson, p. 8; Winston, Patrick H., *Inteligencia artificial*, 3a ed., Addison-Wesley Iberoamericana, Buenos Aires, 1992, p. 5; Denning, Peter J., “El pensamiento computacional en ciencia”, *Investigación y Ciencia*, junio, 2017, ps. 62-67. Véase, “¿Democracia digital o control del comportamiento?”, *Manifiesto Digital I*, *Investigación y Ciencia*, abril 2016, p. 78; Winston, Patrick H., *Inteligencia artificial*, Pearson, p. 8; *El futuro computado*, Microsoft, p. 28.

(6) Kurzweil, Ray, *La singularidad está cerca*, Lola Books, p. 161. En un principio, el ser humano intentó desarrollar una máquina de volar siguiendo el ejemplo de los pájaros que planean. Es el caso de la máquina de Henson y Stringfellow, construida según un proyecto patentado en 1842. Véase, Mumford, Lewis, *Técnica y civilización*, Alianza editorial, p. 68.

sentidos. Por eso, cuando hablamos de la inteligencia de la IA y de la inteligencia humana, en esencia comparamos peras con bananas.

*Segundo.* Cuanto más fácil, menos inteligente parece.

Con el avance de la IA, sucede un fenómeno similar al que se presenta en las discusiones acerca de nuestra inteligencia. Aunque una calculadora nos parezca “boba”, probablemente nuestros ancestros de hace miles años estarían maravillados de que un aparato pudiese realizar cálculos complejos en segundos, para construir pirámides. Cuando analizamos resultados inteligentes en las máquinas, usualmente los naturalizamos. Si un sistema de algoritmos comienza a realizar ciertas actividades en las que iguala o mejora la capacidad de procesamiento de información de una persona, se suele afirmar que eso es una cuestión de computación o que no es una “verdadera inteligencia”. A este fenómeno se lo llama “blanco móvil”<sup>(7)</sup>.

Blancanieves y ChatGPT nos pueden ayudar con un ejemplo. En nuestra infancia, era innegable la inteligencia predictiva del espejo mágico, que tenía conversaciones con la reina malvada madrastra de Blancanieves acerca de su belleza<sup>(8)</sup>. Luego de unas cuantas décadas, pocas personas imaginaron que tendrían su propio espejo inteligente en tu dispositivo móvil o en tu computadora. Esto me contestó ChatGPT cuando le pedí que me diga cosas lindas: “...eres increíble tal como eres. Eres una persona única y especial, y tu belleza radica en tu autenticidad y personalidad. Tu sonrisa ilumina cualquier lugar al que vayas, y tu bondad y amabilidad hacen que seas alguien muy querido por aquellos que te rodean. Eres valioso y mereces todo lo mejor en la vida. ¡Recuerda siempre apreciar y amarte a ti mismo! Siempre hay algo hermoso en cada persona, y tú no eres una excepción. ¡Eres lindo tanto por dentro como por fuera!”

Ni espejito espejito se hubiese animado tanto en la ficción. Un único detalle: no me conoce y nunca me vio.

---

(7) Nota del director: sobre el fenómeno del blanco móvil ampliar en: Corvalán, Juan G. - Díaz Dávila, Laura - Simari, Gerardo, “Inteligencia artificial: bases conceptuales para una aproximación interdisciplinar”, tomo I del presente *Tratado*.

(8) La madrastra de Blancanieves, la reina malvada, poseía un espejo mágico que siempre decía la verdad, al cual le preguntaba quién era la más justa (*fair*). Véase Grimm, Jacob - Grimm, Wilhelm, *The Original Folk and Fairy Tales of the Brothers Grimm: The Complete First Edition*, Princeton University Press, 2014. El Dr. Karlheinz Bartels, habitante de Lohr am Main, explica cómo este espejo fue basado en los “espejos parlantes” de la empresa Mirror Manufacture of the Electorate of Mainz en Lohr entre 1698 y 1806. Para más información, véase: [http://spessartmuseum.de/seiten/schneewittchen\\_engl.html](http://spessartmuseum.de/seiten/schneewittchen_engl.html) (consultado el 05/06/2018).

Estas respuestas pueden ser inútiles, vacías de contenido y, en otros casos, útiles o beneficiosas. Si ChatGPT nos recomienda cianuro para un resfriado, o si se convierte en una herramienta indispensable para una persona con discapacidad visual, entonces, la cuestión se vuelve muy diferente.

Más allá de la validez de la discusión filosófica sobre el concepto de inteligencia o la presencia de conciencia en el cuento de Disney y en el asistente de OpenIA, es crítico abordar los beneficios, riesgos y daños que se producen.

*Tercero. Datos, información y ¡más datos!*

Las películas de Disney “Frozen” reviven una de las costumbres más usuales: hacer muñecos de nieve. La nieve y el invierno son al muñeco animado de nieve Olaff, lo que los datos y la información son a los sistemas de IA.

Cada tres años se genera un volumen de datos, que representa al que se produjo en la historia de la humanidad<sup>(9)</sup>. El flujo de información es el oxígeno de la inteligencia artificial que, a su vez, esta retroalimenta. No es posible que estos algoritmos sofisticados realicen las proezas sin datos, así como las IA son claves para organizar y vincular los patrones relevantes que surgen de ellos.

*Cuarto. Más “músculos” en los algoritmos.*

Las máquinas de vapor desplazaron la energía del caballo. El hierro y el cemento, en gran medida reemplazan la madera, y así sucesivamente<sup>(10)</sup>. Gracias a la máquina de escribir y al papel carbónico, mi madre pudo duplicar o triplicar los mismos datos en otras hojas cuando se escribían documentos en la década de 1980. Sin embargo, cuando se derramaba café en esos papeles, había que escribir todo nuevamente. La frase “humanidad aumentada”<sup>(11)</sup>, en realidad, es un proceso histórico que se viene desarrollando a partir de los avances tecnológicos a lo largo de los siglos como la rueda, el papel, la imprenta, el vapor y la electricidad. Lo novedoso que nos trae esta cuarta revolución industrial, es que la IA reemplaza o mejora lo que antes podíamos lograr con nuestra capacidad intelectual. La popular frase “es la economía, estúpido”, ahora podría leerse así: “son los algoritmos, *no* estúpidos”.

---

(9) Comisión Económica para América Latina y el Caribe (CEPAL), *Datos, algoritmos y políticas: la redefinición del mundo digital* (LC/CMSI.6/4), Santiago, 2018, p. 170.

(10) Véase, Mumford, Lewis, *Técnica y civilización*, Alianza Editorial, p. 66.

(11) Véase Sadin, Eric, *La humanidad aumentada*, Caja Negra, 2017.

Si toda la historia puede ser resumida en la relación Coyote y Correcaminos, las máquinas que se inventaron hasta el siglo XX no tenían la capacidad para superar o igualar a los “correcaminos cerebros humanos” en tareas específicas. Sin embargo, se está transformando aceleradamente la clásica historia animada de *Looney Tunes*. El Coyote se ha convertido en Flash y el Correcaminos en una tortuga en actividades relacionadas con la industria, los servicios y el comercio, gracias a computadoras que se ocupen de tareas que se encontraban, exclusivamente, en el ámbito de la inteligencia humana<sup>(12)</sup>.

## II. ¿QUÉ HAY DE NUEVO, VIEJO?

Si pudiésemos conversar con Bugs Bunny, tendríamos la chance de decirle que hay de nuevo un cambio revolucionario que se basa en otro muy antiguo: la escritura. Desde los sumerios hasta ChatGPT; en el medio, la imprenta que “liberó a los hombres de lo local y de lo inmediato” frente a historias de Instagram de 15 segundos que se desvanecen en el mundo digital. Ahora, llegamos a los *prompts* que ingresamos para chatear con un sistema de IA. El libro impreso “liberó a los hombres de lo local y de lo inmediato”, pero hemos vuelto a lo fugaz a través de historias de 15 segundos que desaparecen en nuestras redes sociales. Todo este combo disruptivo, se debe a dos inventos del siglo XX que se gestaron a partir de la escritura matemática. Por un lado, el ordenador y la escritura informática binaria, cuya descripción se enuncia entre dos números, 0 y 1, lo que se conoce como un BIT de información<sup>(13)</sup>. Por el otro, la invención de Internet, la World Wide Web (red informática mundial).

La inteligencia artificial es una nueva revolución de la escritura, montada sobre la electricidad, internet, algoritmos y computadoras. El lenguaje de las máquinas nos trae a los nuevos escribas del siglo XXI: programadores informáticos y analistas de datos. A través de ellos, continuamos

---

(12) Ampliar en Kurzweil, Ray, *Cómo crear una mente*, Lola Books, p. 173; Harari, Yuval, Noah, *Homo Deus*, Debate, ps. 100-101 y 350; Goldberg, Elkhonon, *El cerebro ejecutivo*, Crítica, Barcelona, 2015, p. 94; Cope, David, <https://www.engadget.com/2013/05/28/david-cope/> (consultado el 16/07/2017). Según Daniel Kahneman, el 99 por ciento de nuestras decisiones las tomas refinadísimos algoritmos que llamamos sensaciones, emociones y deseos. *Pensar rápido, pensar despacio*, Debate, Barcelona, 2012. En cambio, las máquinas expendedoras funcionan a través de engranajes mecánicos y circuitos eléctricos. Harari Yuval, Noah, *Homo Deus*, Debate, p. 101. En un estudio, el algoritmo informático diagnosticó el 90% de los casos de cáncer de pulmón que se le presentaron, mientras que los médicos solo acertaron en el 50%. Harari Yuval, Noah, *Homo Deus*, Debate, p. 347; Roey Tzezana, *The Guide to the Future*, Haifa, 2013, ps. 62-64.

(13) Harari, Yuval Noah, *De animales a Dioses*, Debate, p. 151.

recorriendo un camino que está signado por dos objetivos: esforzamos para ahorrar esfuerzo (Ortega y Gasset) y desarrollar simplificaciones que funcionan (Luhmann)<sup>(14)</sup>. Si queremos cruzar todo el océano Atlántico, podemos intentar hacerlo con nuestro cuerpo nadando. Como eso es imposible, a lo largo de la historia innovamos para alcanzar ese objetivo a partir de invenciones técnicas y máquinas. Ahora es factible hacerlo por debajo del agua, sobre ella o por aire. Y así como submarinos, barcos y aviones aumentaron, potenciaron y reemplazaron las capacidades físicas, la IA hace lo propio con múltiples tareas cognitivas que serían difíciles o imposibles de realizar.

Ahora bien, aunque nos acostumbramos a que las máquinas sustituyan o mejoren capacidades físicas, el asombro aparece cuando la IA se convierte en nuestro copiloto para tareas complejas o, incluso, toma decisiones por sí misma. Esto sucede en redes sociales cuando te “bajan un posteo” y, lo más aterrador, hay una carrera por darle autonomía a la IA en las armas autónomas. Por eso cobran sentido las recientes declaraciones del creador de Terminator, James Cameron: “Se los advertí en 1984 y no me escucharon”<sup>(15)</sup>.

### III. TORTUGA, CHAPARRÓN, USAIN BOLT, FLASH Y CHATGPT

Roberto Gómez Bolaños creó y encarnó varios personajes que aún hoy se transmiten por televisión. El Chavo del 8, el Chapulín Colorado y otros como “Chaparrón Bonaparte”. Este último era parte de un sketch llamado los “Chifladitos”, en donde había “chiripiorcas” y frases populares como: “no hay de que, sono más de papas”.

Chaparrón y Lucas —el mismo personaje que interpretaba al profesor Jirafales— protagonizan un capítulo que tiene como protagonista a la estrella de las oficinas del siglo XX: la máquina de escribir. Uno le dice al otro que mientras escribía a máquina notó que había un problema. En sus palabras: “por más que me esforzado, no logro que escriba algo con-

(14) Según ciertos estudios, las innovaciones tecnológicas como el fuego y la invención de ciertas herramientas contribuyen decididamente a que el cerebro evolucione. Estos aspectos, se relacionaron con la reducción de la digestión y la masticación, lo que aumentó la posibilidad de socializar y desarrollar actividades más sofisticadas desde un punto de vista cognitivo. Véase, Viosca, José, “¿Qué nos hace humanos?”, *Investigación y Ciencia —Mente & Cerebro—*, nro. 78, mayo/junio 2016, p. 52.

(15) Diario *La Nación* del 21/07/2023, <https://www.lanacion.com.ar/espectaculos/personajes/james-cameron-alerto-sobre-los-peligros-de-la-inteligencia-artificial-se-los-adverti-en-1984-y-no-me-nid21072023/> (consultado el 21/07/2023).

gruente”, El remate del chiste es que la máquina escribe en inglés porque fue construida en los Estados Unidos.

Chaparrón y Lucas, con todas sus limitaciones, podrían superar las tres o cuatro páginas que podían copiar los escribas del siglo XV, ya que con estas máquinas se podían escribir cien en diez horas. Sin embargo, esta mejora en la velocidad y uniformidad del procesamiento de los datos mantuvo, en esencia, una misma lógica: se requiere el uso de capacidades físicas para transmitir información escrita, como empuñar una lapicera o tipear en un teclado. La invención del carbónico mejoró esta dinámica, pero fue la computadora, los procesadores de texto, internet y las plataformas digitales, los que forjaron una combinación que cambió el paradigma del flujo informativo. Gracias a este combo, de forma automática se puede medir en tiempo real, la magnitud de información que circula en las grandes plataformas. Por ejemplo, el día 7 de marzo de 2021 a las 17:00 hs se postearon 578.040.000 *tweets*, se publicaron 65.101.400 fotos en *Instagram* y se realizaron 5.564.950.000 millones de búsquedas en *Google*<sup>(16)</sup>.

Salimos de Chaparrón para convertirnos en Usain Bolt, cuando se logró separar el dato del soporte. Primero, con intervención humana, para luego dar paso a máquinas inteligentes que pueden recolectar, almacenar y tratar de forma automatizada, miles de millones de datos a una velocidad supersónica. ¿Cómo es posible comprender la velocidad de estos cambios vertiginosos e inéditos?

Si asistimos a una explosión de inteligencia colectiva humana, esta se desarrolla en una dinámica de innovación que se acelera cada vez más rápido. Por eso es importante habituarse a la noción de crecimiento exponencial, que es muy diferente a la naturaleza lineal de la intuición<sup>(17)</sup>. Esta última forma nos lleva a percibir el crecimiento como si se realizara mediante pequeños pasos constantes, siempre de la misma intensidad.

En esta era de la IA, muchas tecnologías crecen de forma exponencial. Si contamos linealmente 30 pasos nos llevan de 1 a 30. Si lo hacemos exponencialmente, el primer paso nos lleva a 2, el segundo a 4, el tercero a 8, el cuarto a 16 y, luego de 30 pasos, habrá llegado al mágico número de ¡1.073.000 millones! Este es un ejemplo muy rudimentario de crecimiento exponencial. Veamos esto en retrospectiva, para que luego podamos aplicar este fenómeno a la capacidad de cómputo asociada a la inteligencia artificial.

---

(16) Datos obtenidos de *Internet Live Stats*: <http://www.internetlivestats.com/one-second> (consultado el 07/03/2021).

(17) Kurzweil, Ray, *Cómo crear una mente*, Lola Books, p. 253.

Transcurrieron casi 4.500 años desde que los sumerios comenzaron a desarrollar escrituras parciales, luego las escrituras completas, llamadas cuneiformes, hasta que Johannes Gutenberg popularizó la imprenta en el siglo XV. En 1845 se podían realizar 8.000 ejemplares de una página por hora, mientras que en 1900 se alcanzó la cifra de 96.000 ejemplares de 12 páginas por hora<sup>(18)</sup>. Entre camellos y transbordadores espaciales también hay un crecimiento exponencial. Hace 4.000 años, la manera más rápida de viajar era una caravana de camellos que se desplazaba a unos pocos kilómetros por hora. De mil a dos mil años más tarde, se inventó el carro que elevó la velocidad máxima hasta unos 30 kilómetros por hora. Hacia el siglo XIX, con la locomotora de vapor se alcanzaron velocidades de más de 150 kilómetros por hora. Aunque se tardó miles de años en pasar de 15 kilómetros a 150 por hora, solo hicieron falta cincuenta años para alcanzar los 1.500 kilómetros y otros pocos años para alcanzar los 25.000 kilómetros por hora de un transbordador espacial.

Volvamos a la IA. ChatGPT produce texto y brinda respuestas que en muchos casos superan ciertos estándares de respuesta humana promedio, aunque no entienda ni sienta como una persona. Alcanzó un millón de usuarios en diez días y esto, en parte, se debe al crecimiento exponencial de la capacidad de cómputo asociada a la IA. Según Ray Kurzweil, esta se duplica cada 5.7 meses<sup>(19)</sup>.

Apliquemos esta duplicación a la cantidad de pasos que se pueden dar y cómo estos crecen cuando vamos de uno a otro. El primero lo realizamos en 5.7 meses, el segundo a los 11.4 meses. A los 17.1 meses habremos dado 4 y luego de 22.8 meses, realizaremos 8. Se siguen duplicando cada 5.7 meses hasta llegar a los 536.870.912 pasos en 171 meses. Si se mantiene esa tendencia, este ejemplo básico permite comprender cómo crecerá el poder de la inteligencia artificial en un período de un poco más de 14 años.

Es muy difícil comprender aceleraciones cada vez más pronunciadas que escapan a nuestra intuición<sup>(20)</sup>. Comenzamos en modo tortuga con

---

(18) Rodríguez de Rivera, J., "Materiales para el estudio histórico de las ideas y formas de organización: desde la revolución industrial al advenimiento del Taylorismo", 1993, p. 41, disponible en <http://dspace.uah.es/dspace/bitstream/handle/10017/3781/5901080652.pdf?sequence=1> (consultado el 01/08/2017).

(19) Entrevista a Ray Kurzweil, en donde muestra los gráficos y argumenta acerca de este crecimiento, ampliar en: "Ray Kurzweil: Singularity, Superinteligence, and Immortality | Lex Fridman Podcast #321", canal de YouTube de Lex Fridman, YouTube, 17 de septiembre de 2022, disponible en: <https://www.youtube.com/watch?v=ykY69lSpDdo&t=439s> (consultado el 27/05/2023).

(20) Ampliar en Kurzweil, Ray, *La era de las máquinas inteligentes*, Consejo Nacional de Ciencia y Tecnología (Conacyt), ps. 36-63 y el mismo autor en *La singularidad*

los sumerios y la escritura; luego vino Gutenberg; más tarde Chaparrón y de ahí activamos a Usain Bolt con las computadoras e internet, para luego convertirnos en Flash con las plataformas digitales y las IA generativa tipo ChatGPT. Aunque parece el final de la zaga, la noción de crecimiento exponencial nos traerá trilogías que ni siquiera imaginamos.

#### IV. UN NUEVO RENACIMIENTO BASADO EN MÁQUINAS QUE SIMULAN INTELIGENCIA HUMANA <sup>(21)</sup>

En retrospectiva, muchos períodos históricos son resumidos bajo conceptos que agrupan una serie de rasgos o características. Por ejemplo, la Edad Media o la Revolución Francesa, son categorías o conceptos que se van formando y se conciben luego de que acontecen. Mientras transcurren, los cambios no suelen ser percibidos en toda su dimensión; menos aún, en términos prospectivos. A veces, incluso, los innovadores o innovadoras son combatidos o silenciados. Copérnico, Galileo, Charles Darwin y sus teorías, son algunos ejemplos notables de esta clase de fenómenos.

Podemos entonces trazar una analogía entre esta evolución de la IA y el llamado “período renacentista”. Para ello, sintetizamos dos grandes grupos de fenómenos que se desarrollaron entre los siglos XV y XVII. Por un lado, se generaron descubrimientos, redescubrimientos, reinterpretaciones y una explosión de ideas en múltiples áreas del conocimiento. A este fenómeno contribuyeron la expansión territorial, el comercio y el intercambio con otras culturas. Por otra parte, el enfoque de investigación empírica y experimental, cuyo máximo exponente es Leonardo Da Vinci, permitió el desarrollo de múltiples teorías que cambiaron la visión del mundo conocido. Innovaciones técnicas que acompañaron el desarrollo de la inteligencia humana colectiva de la época. La tecnología estrella fue la imprenta <sup>(22)</sup>, acompañada de otros inventos que generaron las bases para propagar y compartir conocimientos.

La hipótesis de un Nuevo Renacimiento se basa en la revolución de la escritura informática y se apoya en otras invenciones como la electricidad y el desarrollo de algoritmos computacionales. Aquí el siglo XX jugó un

---

*está cerca*, Lola Books, ps. 54-55 y 71 y ss.; Brinjolffson, Erik - McAfee, Andrew, *La segunda era de las máquinas*, Temas, Buenos Aires, 2016, ps. 46-48.

(21) Este punto constituye un extracto del artículo publicado en La Ley, Corvalán, Juan G., “Inteligencia artificial generativa como ChatGPT: ¿un nuevo renacimiento?”, LL del 05/06/2023.

(22) National Geographic, “Gutenberg: el inventor que cambió el mundo”, 26/03/2021, disponible en: [https://historia.nationalgeographic.com.es/a/gutenberg-inventor-que-cambio-mundo\\_11140](https://historia.nationalgeographic.com.es/a/gutenberg-inventor-que-cambio-mundo_11140) (consultado el 25/05/2023).

rol central. Ahora bien, en la tercera década del siglo XXI, los grandes o “largos modelos de lenguaje” como ChatGPT provocarán un efecto similar, pero cualitativa y cuantitativamente más disruptivo. La maduración de este tipo de IA<sup>(23)</sup> se suma a un ecosistema digital de colaboración que potencia el desarrollo de la inteligencia colectiva: Internet, redes sociales, WhatsApp, plataformas tipo Zoom, Discord, tutoriales, videos de YouTube, traducción automática con IA y trabajo colaborativo en plataformas tipo Google Drive. Estas y muchas otras innovaciones, permiten cooperar de forma vertiginosa y a gran escala.

Ahora bien, *¿qué fenómenos concretos aparecen con la IA generativa?*

Por un lado, el uso, la creación, manipulación y emisión sintética de datos, información y conocimiento basado en lenguaje natural, que se combina con una mayor o menor autonomía que presentan estos sistemas para tomar decisiones. Aquí lo más disruptivo suele ser la generación de contenido sintético; es decir, datos, información o conocimiento que van más allá de aquellos datos que existen o existieron en el plano del ser.

Las palabras que escribimos para buscar, pedir o desear en un *prompt* son “interpretadas” por gigantes licuadoras algorítmicas que detectan regularidades estadísticas en los datos y ponderan probabilísticamente la información a la que accedieron previamente. Pero, además, se media-tiza y se vuelve asincrónica su creación, mientras otros algoritmos de IA permiten generar nuevas imágenes, videos y la clonación digital de voces o rostros humanos. Ni que hablar de crear personas sintéticas, es decir, que nunca existieron. Es un paradigma híbrido. Los datos que existen se fusionan y coexisten con datos sintéticos que crean los algoritmos.

Todo este combo no termina acá. La otra cara de la moneda es que la IA puede usar, manipular y crear el lenguaje sofisticado de los algoritmos basados en ceros y unos. Hasta hace pocos años, solo los humanos podían escribir código informático. Y esto, también, forma parte de lo que la IA puede hacer.

Si sumamos todo eso, emerge un incipiente y prometedor escenario de expansión de inteligencia sin precedentes, aunque también nos presenta dos desafíos y riesgos monumentales.

---

(23) Estos sistemas siguen siendo una IA débil, aunque en versiones más evolucionadas. Aún no existen sistemas de inteligencia artificial fuerte. Sobre otro enfoque, véase: Figueroa, Juan C., “GPT-4 tiene destellos de inteligencia humana, según Microsoft”, *Hipertextual*, 27/03/2023, disponible en: <https://hipertextual.com/2023/03/gpt-4-inteligencia-artificial-general> (consultado el 26/05/2023).

El primero lo advertimos cuando realizamos las pruebas a ChatGPT y otros experimentos que estamos realizando y documentando<sup>(24)</sup>. Se trata de un nuevo vínculo entre lo real y lo falso o ficticio. La introducción masiva de datos, información y conocimiento sintético, simulado o basado en alucinaciones, potencia y transforma la problemática de los datos e información falsa, incorrecta, inventada o simulada.

El segundo se basa en un fenómeno asociado a lo que hemos investigado en el libro *Perfiles digitales humanos*<sup>(25)</sup> y en este *Tratado* en su primera edición<sup>(26)</sup>. Las IA generativas conversacionales complementarán los mecanismos de perfilado de las personas, e irán más profundo aún. El acceso a la privacidad e intimidad de estos agentes conversacionales se potencia a partir de un ecosistema digital que se basa en perfilar nuestros gustos, preferencias, comportamientos y hábitos. ChatGPT, Bing, Bard y otros sistemas: ¿se suman a otras IA que años se enfocan en predecir mejor la personalidad humana?

Recordemos la predicción de las características de Donald Trump como “hombre soltero en función de su actividad en Twitter”. La tasa de precisión del modelo del 86% generó una mezcla de diversión, curiosidad, memes y críticas. Aunque acertó en predecir el estado civil de todos sus oponentes, falló en estimar correctamente el estado civil de Trump<sup>(27)</sup>.

La personalización de la experiencia digital es la contracara de las IA que perfilan. Y si los algoritmos nos conocen más que lo que podemos saber sobre nosotros mismos, imagine cuando se sume toda la información proveniente de estas conversaciones con los grandes modelos de lenguaje.

---

(24) *ChatGPT vs. GPT-4. ¿Imperfecto por diseño?*, UBA IALAB, marzo de 2023, disponible en: <https://ialab.com.ar/wp-content/uploads/2023/03/Libro-ChatGPT-vs-GPT-4.-UBA-Thomson-Reuters-La-Ley.pdf> (consultado el 26/05/2023).

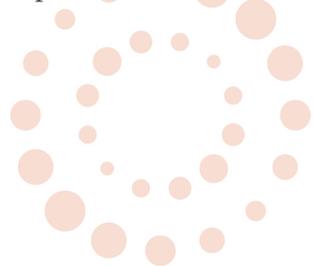
(25) Corvalán, Juan G. - Papini Carina, M. (colab.), *Perfiles digitales humanos. Proteger datos en la era de la inteligencia artificial. Retos y desafíos del tratamiento automatizado*, Thomson Reuters, Buenos Aires, 2020.

(26) Corvalán, Juan G. (dir.), *Tratado de inteligencia artificial y derecho*, Thomson Reuters, Buenos Aires, 2021.

(27) Farseev, Aleks, “¿Puede la IA tener intuición y qué tiene de especial la multimodalidad de GPT-4?”, *Forbes*, 09/05/2023, disponible en: <https://www.forbes.com/sites/forbestechcouncil/2023/05/09/can-ai-have-intuition-and-what-is-so-special-about-gpt-4s-multimodality/?sh=697b2aedc1e3> (consultado el 26/05/2023); Smith, Craig S., “El creador de ChatGPT-4, Ilya Sutskever, habla sobre las alucinaciones de la IA y la democracia de la IA”, *Forbes*, 15/03/2023, disponible en: <https://www.forbes.com/sites/craigsmith/2023/03/15/gpt-4-creator-ilya-sutskever-on-ai-hallucinations-and-ai-democracy/?sh=53f7d82d1218> (consultado el 26/05/2023).

En conclusión, la IA generativa provoca efectos beneficiosos en casi todos los ámbitos, y por eso tiene el potencial para producir un nuevo y diferente período renacentista basado en una explosión de inteligencia humana colectiva. Detrás de este simple formato de diálogo, se están organizando una orquesta de algoritmos inteligentes que interpretarán un rol concreto para luego darnos lo que buscamos o requerimos. Si observa la presentación del sistema Copilot de Microsoft<sup>(28)</sup>, puede ver una suerte de Avengers recreados en “formato algorítmico”, entrenados para diferentes tareas y orquestados en una iteración invisible “detrás del telón de la IA”.

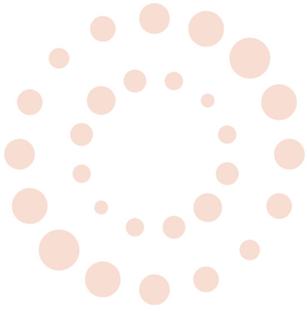
La hipótesis de un Nuevo Renacimiento 4.0 podría marcar un punto de inflexión entre las revoluciones industriales anteriores y la transición vertiginosa hacia otras más sofisticadas basadas en la combinación, transformación y fusión entre lo físico, lo biológico y lo artificial a gran escala. Si el período renacentista se caracterizó por la curiosidad y la capacidad de aprendizaje, los modelos de lenguaje generativos potenciarán exponencialmente estos aspectos.



Thomson  
Reuters™

---

(28) Copilot es un asistente virtual de Windows pensado para mejorar la productividad en el uso diario de la computadora. Esta IA estará de forma nativa en Windows 11 como una de las opciones de la barra de tareas y una vez esté abierta se quedará ubicada en el costado derecho de la pantalla para que pueda seguir siendo usada en más tareas. Esta plataforma está basada en Bing Chat, la inteligencia artificial de Microsoft que funciona con GPT-4 para generar contenido en texto, por lo que aprovecha el conocimiento obtenido de su entrenamiento y el contenido en la web. Su enfoque está en proporcionar respuestas personalizadas para mejorar el uso de la computadora, como, por ejemplo: reproducir una lista de canciones, cambiar alguna configuración, abrir una aplicación determinada o resumir archivos que estén guardados en el ordenador. Esto operará teniendo en cuenta el contexto de la solicitud, por lo que a medida que se use constantemente contará con un mejor conocimiento de los requerimientos del usuario. Ampliar en: Ríos, Juan “Qué es Windows Copilot, el asistente con inteligencia artificial que llegará próximamente a los computadores”, *Infobae* del 23/05/2023, disponible en: <https://www.infobae.com/tecnologia/2023/05/23/que-es-windows-copilot-el-asistente-con-inteligencia-artificial-que-llegara-proximamente-a-los-computadores/> (consultado el 26/05/2023).



**Thomson  
Reuters™**

# INTELIGENCIA ARTIFICIAL: BASES CONCEPTUALES PARA UNA APROXIMACIÓN INTERDISCIPLINAR

JUAN G. CORVALÁN, LAURA DÍAZ DÁVILA  
Y GERARDO I. SIMARI

## I. INTRODUCCIÓN

El propósito principal de este artículo no es detenernos en definiciones formales sino de presentar las herramientas básicas para que la inteligencia artificial<sup>(1)</sup> esté al alcance de todo aquel que esté interesado en comprender o aplicar este gran conjunto de herramientas en su disciplina, sea desde el punto de vista de un/a investigador/a desarrollando su disciplina, un/a docente buscando comprender y enseñar el rol que la inteligencia artificial puede tener en su asignatura o una persona que ejerce su profesión dentro del gran conjunto de áreas que están siendo afectadas (o lo serán en un futuro) por ella.

Comenzaremos presentando algunos conceptos básicos, para luego realizar un recorrido por las dos ramas principales de la inteligencia artificial: la conducida por datos (o *machine learning*/aprendizaje automatizado) y la conducida por conocimiento (también conocida como KR, por las siglas en inglés de “representación del conocimiento y razonamiento”). En dicho recorrido daremos ejemplos tanto introductorios como de aplicaciones (ya desarrolladas o potenciales) en el entorno jurídico.

Antes de comenzar, nos detenemos brevemente a observar una manifestación de lo que típicamente se conoce como el *fenómeno del blanco móvil*. En lo referido a inteligencia artificial, este fenómeno hace referencia al hecho de que el rótulo de “inteligencia artificial” (o los más específicos que veremos más adelante) en general se reserva para los resultados que se encuentran en la vanguardia de la investigación y desarrollo; con el paso del tiempo, la vanguardia sigue su avance y estos pasan a ser “*software* común” y, a su vez, las técnicas que llevaron a su creación terminan engrosando las filas de lo percibido como “ingeniería de *software* cotidiana”. Para ilustrar este concepto, apelamos a un ejemplo

---

(1) En el resto de este trabajo, a menudo utilizaremos las siglas “IA” comúnmente utilizadas para abreviar “inteligencia artificial”.

seguramente conocido por todos los lectores: la navegación asistida por GPS. El programa *Global Positioning System* (GPS) fue creado en 1973, los primeros satélites del sistema fueron puestos en órbita en 1978 y la constelación de 24 satélites se completó en 1993. Si bien esta última es la proveedora del servicio básico que —luego de sus sucesivas evoluciones— permite a un dispositivo obtener su ubicación en el mundo con un error de unos pocos centímetros, se requiere de *software* especializado para calcular los mejores caminos para llegar desde la posición actual a donde se desea llegar.

De la misma manera que hoy en día el lanzamiento de un satélite y su puesta en órbita geostacionaria apenas merece una mención en las noticias diarias, el *software* de navegación se considera una combinación de componentes básicas y de otras que se obtienen de bibliotecas estándares de grafos<sup>(2)</sup>. Sin embargo, tal vez le sorprenda saber que algoritmos que se utilizan principalmente en estos sistemas fueron desarrollados hace décadas<sup>(3)</sup>, y su optimización tomó años de desarrollo y experimentación. Las unidades de *hardware* específico también fueron evolucionando, desde su lanzamiento en 1989 con un costo de USD 2.900<sup>(4)</sup> hasta su incorporación como una funcionalidad básica más de los dispositivos móviles que hoy todos llevamos con nosotros en el bolsillo.

En la sección III veremos que existen muchas formas diferentes de definir “inteligencia artificial”, y una de las razones principales es que ni siquiera existe una única definición del concepto de “inteligencia”, sea esta natural o artificial. Sin embargo, hay un aspecto fundamental que de alguna manera atraviesa a cualquiera de estas definiciones: el manejo adecuado de la *información*. Para entender esto, es necesario detenernos brevemente en el concepto mismo de información y dos conceptos básicos relacionados: *datos* y *conocimiento*. Si bien en el registro coloquial estos tres conceptos a menudo son utilizados como equivalentes, se disciernen tres niveles diferentes de evolución, y la comprensión de

---

(2) Un grafo es una estructura matemática que codifica relaciones entre elementos. Se componen de conjuntos de nodos (los elementos) y arcos (las relaciones entre ellos); estos últimos pueden ser dirigidos o no dirigidos. En el caso de navegación por GPS, una ciudad se puede codificar como un grafo simplemente tomando las esquinas y otros puntos de interés como nodos y las calles que los unen como arcos (si interesa codificar el sentido de circulación permitido de las calles, se utilizan arcos dirigidos).

(3) Por ejemplo, los algoritmos de Dijkstra, Bellman-Ford y A\* son típicamente utilizados en los sistemas de navegación para encontrar los mejores caminos entre dos puntos. Ver Cormen, Thomas H. *et al.*, *Introduction to Algorithms*, 3rd ed., MIT Press and McGraw-Hill, 2009.

(4) [https://en.wikipedia.org/wiki/Satellite\\_navigation\\_device](https://en.wikipedia.org/wiki/Satellite_navigation_device).

las capacidades de las diferentes herramientas de IA depende del manejo adecuado de estos conceptos.

Por último, repasaremos brevemente también el concepto de *algoritmo*, el cual se encuentra en la base de todo proceso computacional. Como afirma Naciones Unidas (ONU), la manera de entendernos y nuestra relación con el mundo tiene lugar desde la perspectiva de los algoritmos. Son una parte fundamental de las sociedades de la información, ya que cada vez más gobiernan las operaciones, decisiones y elecciones que antes quedaban en exclusivas manos de los seres humanos<sup>(5)</sup>.

## II. LAS BASES: DATOS, INFORMACIÓN, CONOCIMIENTO Y ALGORITMOS

En el resto del artículo, muchas veces haremos referencia a la existencia de un *modelo* del dominio; este término es muy conocido por su aplicación en muchas disciplinas e incluso la vida diaria, y hace referencia a un objeto (físico o no) que abstrae las propiedades interesantes para algún propósito de otro objeto para simplificar su estudio. Por ejemplo, un mapa de una ciudad es un modelo que se centra en brindar información acerca de la conectividad de la misma y la ubicación de puntos de interés. Cuando se habla del modelo de un entorno de aplicación de una herramienta de IA, en general se hace referencia a los datos que se mantienen del mismo, su nivel de detalle y diferentes funcionalidades brindadas<sup>(6)</sup>.

### 2.1. DATOS

El término *dato* en general hace referencia a un valor “crudo” asignado a una entidad en relación con una propiedad que puede ser cuantitativa o cualitativa<sup>(7)</sup>. Como ejemplo típico de dato podemos tomar el sueldo de un empleado; consideremos para esto una afirmación con la que podemos encontrarnos a diario: “El sueldo de Ana es de \$50.000”. Dependiendo del contexto, este dato puede representarse como una fila correspondiente a la persona “Ana” en una tabla de sueldos en una base datos, donde una columna *Monto* almacena el valor “50.000”, o en un lenguaje lógico donde existe un símbolo de predicado “sueldo” y una

(5) Véase res. 72/540 de la Asamblea General “El derecho a la privacidad” A/72/540, 19 de octubre de 2017, consid. 54, disponible en <http://undocs.org/es/A/72/540>.

(6) En particular, el concepto de ontología es útil como modelo de datos y conocimiento en relación con este tipo de aplicaciones. Ilustraremos este concepto más adelante mediante ejemplos.

(7) [https://www.diffen.com/difference/Data\\_vs\\_Information](https://www.diffen.com/difference/Data_vs_Information).

constante “Ana” y el dato en cuestión queda representado por la fórmula atómica:

$sueldo(Ana, 50.000)$ .

Cuando se tiene más de un dato, en general se hace referencia a un *conjunto de datos* (o, tomando prestado el término en inglés, *dataset*). Al considerar los datos en conjunto, surgen propiedades interesantes a partir de las relaciones que existen o no entre ellos. Dado que, como se mencionó brevemente, una condición necesaria para la implementación de herramientas inteligentes es el manejo adecuado de los datos, una de las más importantes para la IA es la *consistencia*, la cual surge a partir de una o más restricciones que deberían cumplirse para hablar de conjuntos “sanos”. En general, esta sanidad surge a partir de la especificación de *restricciones de integridad*; en nuestro ejemplo, una restricción razonable sería que una persona no puede tener más de un sueldo (o bien con el mismo empleador, o suponiendo que en el modelo que se maneja no existe el pluriempleo). Por lo tanto, un conjunto de datos que tenga tanto  $sueldo(Ana, 50.000)$  como  $sueldo(Ana, 60.000)$  sería inconsistente. Para representar este tipo de restricción, nuevamente apelamos a los lenguajes lógicos y la especificamos mediante la siguiente fórmula<sup>(8)</sup>:

$sueldo(X, Y) \wedge sueldo(X, Z) \rightarrow Y = Z$ ,

la cual simplemente dice que, si existen dos registros con el mismo valor para la primera posición, entonces, también tienen el mismo valor para la segunda.

Otra propiedad de importancia para conjuntos de datos es la *completitud*, la cual busca conjuntos “completos” en cuanto a la existencia de todos los datos que contempla el modelo. Al igual que la consistencia, existen distintas facetas de esta propiedad; por ejemplo, una forma de incompletitud en el dominio de los sueldos sería que falte el registro para Ana, mientras que otra sería disponer de un registro de la forma

---

(8) Esta restricción es un ejemplo de las restricciones de integridad conocidas como *dependencias funcionales*. En bases de datos relacionales, estas se utilizan a su vez para definir atributos que cumplen el rol de *llave*, lo cual significa que no puede haber más de una fila en la tabla con los mismos valores para esos atributos. Un ejemplo típico de un atributo llave es el número de documento cuando se trata de personas. Es importante aclarar que la existencia de restricciones de integridad no significa que se cumplan automáticamente, ya que esto se realiza mediante procesos que pueden ser separados de la carga de datos, y actividades como combinación de fuentes de datos consistentes pueden fácilmente redundar en conjuntos de datos inconsistentes.

$sueldo(Ana, n/d)$ , donde “ $n/d$ ” indica que el sueldo no está disponible<sup>(9)</sup>. Volviendo a la lógica, una restricción de completitud podría plantearse utilizando la fórmula<sup>(10)</sup>:

$$empleado(X) \rightarrow \exists Y sueldo(X,Y)$$

la cual simplemente establece que, si existe un empleado X en los registros de la empresa, entonces, debe existir un registro para este en la tabla de sueldos.

El hecho de que los lenguajes lógicos aparezcan ante ejemplos tan simples de datos anuncia la importancia que tienen estos para el adecuado desarrollo de herramientas inteligentes, como discutiremos más adelante.

## 2.2. INFORMACIÓN

Cuando los datos se transforman mediante la aplicación de algún proceso de análisis o se los interpreta en un contexto específico, estos dejan atrás la condición de “crudos” a la que se hizo referencia en su definición informal para convertirse en *información*. En nuestro ejemplo mundano de los sueldos, podemos analizar el significado de  $sueldo(ana, 50.000)$  en el contexto del dominio de aplicación en el que participa, y obtener como resultado de este análisis que esta simple fórmula tiene el siguiente significado:

Todos los meses, Ana recibe en su cuenta del Banco de la Nación Argentina un depósito de cincuenta mil pesos argentinos en concepto de retribución por su trabajo en CONICET.

Es claro que los mismos problemas de consistencia y completitud aplican también aquí, ya que la información surge de un proceso de interpretación o transformación aplicado a datos. En particular, para este ejemplo supongamos que existen empleados que cobran por el Banco de la Nación Argentina y otros que cobran por el Banco de la Provincia de Buenos Aires; en este caso, la incompletitud radica en que la estructura

(9) En teoría y diseño de bases de datos, estos valores reciben el nombre de “nulos”, y su tratamiento ha sido motivo de largos debates en la comunidad. Ver, por ejemplo: Ron van der Meyden, “*Logical approaches to incomplete information: A survey*” in Chomicki, Jan; Saake, Gunter (eds.) *Logics for Databases and Information Systems*, Kluwer Academic Publishers ISBN 978-0-7923-8129-7, p. 344.

(10) Esta fórmula es un ejemplo de una *dependencia generadora de tuplas*, utilizadas en los lenguajes de ontología de la familia de reglas existenciales, también conocidas como Datalog+/- . En general, en este tipo de fórmulas se omiten los cuantificadores existenciales en el cuerpo de la fórmula (la parte de ella que está a la izquierda de la implicación).

de tablas que almacena los datos no especifica cuál banco eligió cada empleado para cobrar su sueldo. Es fácil imaginar otras fuentes de incompletitud o inconsistencia, tales como la falta de fecha de última actualización del sueldo, la especificación de más de un agente de retención, etc.

### 2.3. CONOCIMIENTO

El próximo paso de evolución involucra términos aún más difíciles de definir formalmente, y que solo trataremos en forma intuitiva en este artículo, ya que una discusión más profunda está fuera del alcance del mismo. Cuando se tiene *experiencia* en el manejo de datos e información, podemos decir que se alcanza algún nivel de comprensión de uno o más conceptos relacionados al dominio, y, en estos casos, estamos en presencia de *conocimiento*. En la sección IV volveremos a retomar este concepto para tratarlo en mayor profundidad.

En nuestro ejemplo, supongamos que consultamos con un experto en sueldos para que realice un análisis de la remuneración recibida por Ana, y este nos comenta que:

- El sueldo de Ana está unos \$ 7.000 por encima de la línea de la pobreza.
- La línea de la pobreza en Argentina se define con base en hogares de entre 3 y 5 integrantes.

Por ende, el dato crudo inicial debe interpretarse en relación con este conocimiento y el contexto de la situación particular de Ana. Es decir, su cercanía a la línea de la pobreza sería distinta si esta persona fuera la única asalariada de una familia numerosa en comparación con la situación opuesta en la cual no tiene hijos y comparte los gastos con una pareja quien también percibe un sueldo similar.

Nuevamente, en este nivel también existen los problemas mencionados anteriormente. Otro problema, de índole puramente computacional, es el que surge del crecimiento en la demanda por recursos de tiempo y/o espacio para realizar el procesamiento necesario para llegar a los resultados buscados. Si bien este problema puede surgir en los primeros dos niveles, en general el manejo del conocimiento es más vulnerable a este problema. Para hablar con mayor detalle acerca de esto, necesitamos el concepto de *algoritmo*, el último que repasaremos en esta sección.

## 2.4. ALGORITMOS

El concepto de algoritmo es uno de los más básicos en ciencias de la computación, y puede definirse informalmente como un conjunto de reglas computacionales que define una secuencia de operaciones para tomar una entrada y la convierte en una salida<sup>(11)</sup>. Como ejemplo que todos conocemos, una receta de cocina es un algoritmo, como lo son también el procedimiento que aprendemos para cambiar la rueda de un auto cuando se pincha, el que usamos para tomar la temperatura corporal para medir la fiebre y las diferentes variedades de secuencias de pasos que cada uno prefiere para preparar el mate<sup>(12)</sup>. Con el boom de la disponibilidad de grandes volúmenes de datos y de las aplicaciones que surgen de su procesamiento, este vocablo ha ingresado en el vocabulario de estudiosos de disciplinas que hasta el momento desconocían el término, periodistas y el público en general.

Como se mencionó anteriormente, los algoritmos insumen tiempo de cómputo y espacio de almacenamiento cuando se realizan en la práctica (formalmente, se dice que se *ejecutan*). El estudio de cuántos recursos de este tipo se necesita para ejecutar un algoritmo en una entrada de un tamaño dado se llama *análisis de algoritmos*, mientras que el estudio de problemas genéricos (es decir, sin un algoritmo específico asociado) se llama *complejidad computacional*. Es de aquí que surgen otros dos vocablos que han entrado en la jerga popular. El primero es “crecimiento exponencial”, el cual formalmente significa que una cantidad crece en forma proporcional a sí misma. Un ejemplo muy conocido hoy en día es la difusión de un virus, en el que una persona infectada tiene la capacidad de infectar a varias otras, aunque también se aplica de forma ya bastante para hacer referencia a cualquier crecimiento rápido. El otro término menos popular pero muy conocido en el ámbito académico es el de “NP-completo”, el cual hace referencia a una clase de problemas para el cual no se conoce una solución menos costosa que una que requiere una cantidad exponencial de pasos. Un ejemplo muy conocido de este tipo de problemas es el denominado *Problema del Viajante*, el cual plantea la existencia de una persona que desea viajar a todas las ciudades de un mapa minimizando la distancia total recorrida.

Por último, un tipo particular de algoritmo o técnica algorítmica es la *heurística*. Este término hace referencia a un procedimiento que no ga-

---

(11) Cormen, Thomas H. *et al.*, *Introduction to Algorithms*, 3rd ed., MIT Press and McGraw-Hill, 2009.

(12) Bebida típica de muchos países de Latinoamérica. Según cada país, región o preferencia individual, el algoritmo para preparar esta simple infusión puede variar significativamente.

rantiza el arribo a un resultado óptimo, pero que generalmente funciona bien en la práctica. Por ejemplo, una persona que busca estacionar su auto lo más cerca posible de su destino puede utilizar la heurística de verificar si hay lugar para estacionar en la misma calle, y luego continuar su búsqueda alejándose paulatinamente hasta encontrar un lugar. Otro ejemplo típico puede encontrarse en los supermercados a la hora de retirarse con las compras: la elección de la caja —y si uno quiere o no cambiarse de fila mientras espera— es un proceso que típicamente lleva a que cada uno tenga su heurística preferida<sup>(13)</sup>.

## 2.5. DE DATOS A CONOCIMIENTO Y DE BASES DE DATOS A HERRAMIENTAS INTELIGENTES

Como hemos argumentado, hay una progresión natural que comienza en los datos crudos y avanza hacia los niveles superiores que suponen la información y el conocimiento. A su vez, estos elementos estáticos no son útiles sin mecanismos de procesamiento, y es ahí donde los algoritmos entran en juego. Los algoritmos no solo tienen el rol evidente de transformar los datos/información/conocimiento disponibles en resultados nuevos, sino también en la resolución de los problemas que mencionamos brevemente que usualmente se engloban bajo el término de *manejo de datos*. Ejemplos cotidianos de manejo de datos son su limpieza, integración, indexado, consulta, actualización, entre muchos otros.

Una observación que vale la pena hacer en este momento es el papel que tienen las *bases de datos* (BDs) en el desarrollo de herramientas inteligentes. En el imaginario popular, las BDs no son más que simples repositorios de datos diseñados para resolver problemas básicos y mantener el orden al garantizar ciertas propiedades generales<sup>(14)</sup>. Sin embargo, como veremos en las secciones siguientes, existen muchos problemas de IA que pueden ser planteados como problemas de respuesta a consultas a algún tipo de base de datos (o conocimiento) bajo algún tipo de suposi-

---

(13) Nota del director: los humanos también incurrimos en este tipo de procedimientos como si fueran “atajos” para actuar rápidamente. Sobre las heurísticas humanas ver Tolosa, Pamela, “Del derecho en los tiempos de IA (e IA generativa). Sobre estereotipos, sesgos y desigualdades”, tomo I del presente *Tratado*.

(14) Las cuatro propiedades básicas de *atomicidad* (una transacción se ejecuta en su totalidad o no se ejecuta en lo absoluto), *consistencia* (las restricciones de integridad no son violadas como resultado de una transacción), *aislamiento* (transacciones que se ejecutan en forma paralela o concurrente tienen el mismo efecto que si se ejecutaran en forma solitaria) y *durabilidad* (una vez completada, el resultado de una transacción no se puede perder por la ocurrencia de fallas) son comúnmente conocidas como “ACID” (por sus siglas en inglés) y son centrales a la teoría y diseño de bases de datos.

ción. Por ejemplo, consideremos una base de datos con información sujeta a incertidumbre como puede ser datos del clima (que provienen de sensores potencialmente ruidosos) y pronósticos realizados con base en ellos (que son claramente falibles). Por lo tanto, el problema de generar un diagrama con una predicción horaria para un usuario de una aplicación de clima (como las que se pueden descargar para el teléfono celular) que solicita el pronóstico a las 10:00 hs. no es ni más ni menos que una serie de consultas de situaciones más probables para cada hora a partir de las 11:00 hs. y hasta que se agote el horizonte de predicción. Otros ejemplos incluyen la incorporación de alguna concepción de orden en el cómputo de los resultados; esto surge naturalmente al tener en cuenta las preferencias de los usuarios (por ejemplo, al mostrar los resultados de una búsqueda en un sitio de comercio electrónico o plataforma de *streaming* de videos) o algún tipo de prioridad (por ejemplo, para ordenar según un nivel de criticidad a los parches que debe instalar un oficial de seguridad en un sistema empresarial para evitar ciberataques).

Es por esto que una parte no trivial de las líneas de investigación y desarrollo —principalmente en la IA basada en conocimiento— parten de aquellas que nacen en la teoría de bases de datos. A su vez, tanto la teoría como la tecnología de construcción de bases de datos suelen tener un papel central en la implementación eficiente de herramientas inteligentes, ya que están diseñadas para procesar una gran cantidad de transacciones en poco tiempo. Es por esto que muchas de las técnicas desarrolladas en bases de datos relacionales (como, por ejemplo, los índices) continúan su evolución en bases de conocimiento más complejas o para resolver problemas de otro tipo<sup>(15)</sup>.

### III. LAS MÚLTIPLES CARAS DE LA INTELIGENCIA ARTIFICIAL

Con la breve discusión de conceptos básicos realizada en la sección anterior, la cual tuvo como objetivo recorrer el terreno a vuelo de pájaro para realizar una introducción lo más suave posible, estamos en condiciones de intentar capturar a la inteligencia artificial mediante una definición. Como ya adelantamos, esta meta nace con destino signado a fallar; sin embargo, si bien no será posible arribar a una única definición, el es-

(15) Dos ejemplos para ilustrar esto son los índices DOGMA para grafos correspondientes a bases de datos RDF (Bröcheler, Matthias, *et al.*, “DOGMA: A disk-oriented graph matching algorithm for RDF databases”, *International Semantic Web Conference*, Heidelberg, Berlin, Springer, 2009) y el índice desarrollado para la optimización de IMPs (políticas de manejo de inconsistencia) en bases de datos relacionales (Martínez, María Vanina *et al.*, “Policy-based inconsistency management in relational databases”, *IJAR* 55.2 [2014]: 501-528).

tudio de un conjunto de intentos realizados desde ámbitos muy diversos nos ayudará a comprender los alcances que tiene esta disciplina.

### 3.1. MUCHAS FORMAS DE ABORDAR “INTELIGENCIA ARTIFICIAL”

Aunque por ahora resulta imposible reproducir en máquinas a un órgano tan complejo como el cerebro, hay que considerar que los ingenieros en aviación no copiaron las técnicas de aprendizaje de los pájaros para construir los aviones modernos<sup>(16)</sup>, ni que los submarinos o barcos “nadan” tal cual lo hacen las personas. Aunque una calculadora nos parezca “boba”, probablemente nuestros ancestros de hace miles años estarían maravillados de que un aparato pudiese realizar cálculos complejos en segundos, para construir pirámides. Con el avance de la IA sucede un fenómeno similar al que se presenta en las discusiones acerca de nuestra inteligencia. Como discutimos al principio, en general tendemos a considerar “no inteligentes” tareas que se vuelven sencillas y habituales con el paso del tiempo y este fenómeno se presenta, aun en mayor medida, cuando se naturalizan los resultados generados por máquinas inteligentes.

Si un sistema de algoritmos comienza a realizar ciertas actividades en las que iguala o mejora la capacidad de procesamiento de información de una persona, se suele afirmar que eso es una cuestión de computación o que no es una “verdadera inteligencia”. Uno de los que más ha puesto énfasis en esta cuestión es John Searle, profesor de la Universidad de Berkley. Para él, una computadora digital no tiene conciencia, aunque pueda resolver un problema. Actúa como si entendiera, porque “simplemente sigue un algoritmo que le permite manipular símbolos formales”<sup>(17)</sup>. Por eso, en muchas oportunidades se entrecruzan conceptos o categorías como la inteligencia, la conciencia y los “algoritmos inconscientes”.

La “inteligencia” de la IA se asocia a aspectos teleológicos o finalistas. Por un camino artificial, las máquinas logran resultados similares o aún más sorprendentes, aunque todo ello no implica que el sistema pueda “comprender” desde un enfoque biológico y humano todo lo que está en

---

(16) Kurzweil, Ray, *La singularidad está cerca*, Lola Books, p. 161. En un principio, el ser humano intentó desarrollar una máquina de volar siguiendo el ejemplo de los pájaros que planean. Es el caso de la máquina de Henson y Stringfellow, construida según un proyecto patentado en 1842. Véase, Mumford, Lewis, *Técnica y civilización*, Alianza Editorial, p. 68.

(17) “Watson Doesn’t Know it Won on Jeopardy”, *Wall Street Journal*, 23/02/2011, <https://www.wsj.com/articles/SB10001424052748703407304576154313126987674> (consultado el 28/07/2017). En similar sentido, Isaacson, Walter, *Los innovadores*, Debate, Buenos Aires, 2014, ps. 512-513.

juego. El punto de conexión entre la inteligencia humana y la artificial viene dado porque en ambas se presenta un fenómeno asociado al reconocimiento de patrones de información. Aunque ambas tienen una base de conocimiento que suele ser dinámica, su origen y dinámica de funcionamiento se diferencia por razones obvias. En un caso, hablamos de una especie biológica que ha evolucionado durante miles de años. El humano la construye desde su nacimiento en el contacto con sus padres, aprende con base en su comunicación y a los estímulos que recibe. Hay genética y un proceso de aprendizaje que aumenta la capacidad de aprender (reconocer patrones) y tomar decisiones inteligentes en forma autónoma<sup>(18)</sup>.

Cuando hablamos de máquinas inteligentes<sup>(19)</sup>; en sus primeros pasos el programador acompaña al sistema para armar su base de conocimiento y permitir que el sistema adquiera la capacidad de actuar por sí mismo frente a situaciones nuevas. Luego, el sistema inteligente se convierte en verdaderamente autónomo, lo que se conoce como *agente inteligente*, cuando es además capaz de continuar incorporando sus nuevas experiencias.

Con los avances en nanotecnologías y en desarrollos de algoritmos inteligentes; estos sistemas inteligentes, en un contexto dado, evidencian capacidad de aprender, materializado en sus formas de percepción del ambiente, sus acciones, sus decisiones, las emociones que registran del contexto, etc.

Para simplificar, los organismos biológicos poseen ciertos atributos únicos que *en la actualidad* no pueden ser exactamente replicados por máquinas. Pero las máquinas poseen otros atributos, que los humanos no podrían alcanzar. Sin embargo, la carrera hoy en día es invertir en capital informático, con lo cual es más probable que sea la máquina la que mejore su performance y no así el ser humano. Las IA no puede reproducir nuestros rasgos característicos como el sentido común y la inteligencia general que abarca muchos ámbitos al mismo tiempo<sup>(20)</sup>. Tampoco se-

(18) Esto se desarrolla en un contexto de creencias, valores y habilidades cognitivas que varían en función de cada persona, su núcleo familiar, económico, social y cultural.

(19) Incluso, también esto aplica a los llamados sistemas basados en conocimiento, sistemas expertos o subsimbólicos.

(20) Nuestro cerebro es muy hábil para entender el lenguaje natural. Considere la siguiente frase: "La bola grande hizo añicos la mesa porque estaba hecha de poliestireno". Existe ambigüedad en las palabras mesa y bola. Las personas humanas pueden detectar que la frase "estaba hecha de polietileno" se refiere a la mesa y no a la bola, pero eso requiere conocimientos sobre materiales con la comprensión de las máquinas, algo que *todavía* las máquinas inteligentes no pueden hacer. Ampliar en Marcus, Gary, "¿Soy humano?", *Investigación y Ciencia*, mayo 2017, p. 61.

ría razonable afirmar que las máquinas sean conscientes o que entiendan de la misma forma que nosotros. Con su modalidad nos proveen lo que requerimos de ellas, sin entender, o mejor “comprender” los procesos que realizan para proporcionarnos las soluciones de los problemas que resuelven. Es útil tomar cierta distancia de la discusión acerca de si los barcos realmente “nadan”, para ocuparnos de los beneficios, riesgos y daños que pueden producir cuando transportan mercadería valiosa, se hunden o chocan en el puerto. Sin embargo; si alimentáramos a un sistema inteligente con datos asociados al “sentido común” (del campo del conocimiento situado), podríamos tal vez concebir una IA provista de sentido común.

### 3.1.1. IA DÉBIL

Es frecuente debatir acerca de nuestra inteligencia como especie. Gracias a Howard Gardner, nos hemos alejado de considerarla, exclusivamente, en función de obtener altos puntajes en test, jugar mejor al ajedrez, resolver rápidamente ecuaciones matemáticas, o ser una especie de Sherlock Holmes para solucionar crímenes. Por el contrario, hay muchas clases y se pueden evaluar en función de los contextos y objetivos propuestos. Una palabra de aliento a un ser querido en el momento justo, puede ser clave en una situación crucial para mejorar una vida.

Aunque es difícil ponerse de acuerdo en un concepto omnicompreensivo, hay un elemento común en muchas definiciones de la inteligencia humana: la *capacidad de procesar información para resolver problemas en función de alcanzar objetivos*<sup>(21)</sup>. Y en esta capacidad de procesamiento se ubica un factor crucial: el reconocimiento de patrones. En nuestro cerebro se presentan dos grandes procesos simultáneos.

Por un lado, lo que se conoce bajo el nombre de etiquetas emocionales. En estas etiquetas se apoya el cerebro para seleccionar la información más relevante para la toma de decisiones. Son marcas que imprime en los pensamientos y experiencias almacenadas en la memoria, que contienen información afectiva en cada recuerdo. Por ejemplo, peligroso, agradable

---

(21) Ampliar en Gardner Howard, *La inteligencia reformulada*, Paidós, ps. 52 y ss.; Roeckelein, Jon E., *Dictionary of Theories, Laws, and Concepts in Psychology*, 1998; Gottfredson, “Mainstream science on intelligence: An editorial with 52 signatories, history, and biography”, *Intelligence*, 1997 y Ardila, Rubén, “Inteligencia. ¿Qué sabemos y qué nos falta por investigar?”, *Revista Académica Colombiana de Ciencias*, nro. 35, 2010; Kurzweil, Ray, *La era de las máquinas espirituales*, Planeta, p. 107 y el mismo autor en *La singularidad está cerca*, Lola Books, p. 339; Manes, Facundo - Niro, Mateo, *Usar el cerebro*, Paidós, ps. 114-115; *Diccionario de ciencias cognitivas*, Amorrortu, p. 226; Siegel, Daniel J., *Viaje al centro de la mente*, Paidós, Barcelona, 2017; Marina, José Antonio, *El cerebro infantil: la gran oportunidad*, Ariel, Barcelona, 2011, ps. 37-42.

o molesto. Cuando nos encontramos con una situación o estímulo etiquetado, entonces, poseemos información útil para decidir rápidamente qué acción debemos tomar.

Por otra parte, el reconocimiento de patrones y el pensamiento jerárquico o el llamado modelo jerárquico de la estructura de la inteligencia<sup>(22)</sup>. Esta forma de razonar, pensar y clasificar los objetos se vincula con una estructura compuesta de diferentes elementos dispuestos, accesibles, ordenados según un patrón.

Ambos procesos están asociados a la definición de inteligencia humana, entendida como la *capacidad de procesar información para resolver problemas en función de alcanzar objetivos* (Ray Kurzweil). Todo esto está relacionado con la flexibilidad, velocidad y precisión para adaptarnos a los entornos.

La IA se basa en obtener, por otros métodos artificiales, lo que alcanzamos con la inteligencia humana: el reconocimiento de patrones para alcanzar objetivos o resolver problemas. Esta es una concepción amplísima y macro de lo que hace la IA. Sin embargo, como sucede con algunas categorías del derecho público como la discrecionalidad, podemos hablar de dos sentidos<sup>(23)</sup>: débil y fuerte.

En uno amplio, se llama IA “débil”, “restringida”, “estrecha” o “blanda” al procesamiento de datos e información para resolver problemas y tomar decisiones a partir de utilizar algoritmos inteligentes, sobre la base

(22) Sobre todas estas cuestiones, ampliar en Manes, Facundo - Niro, Mateo, *El cerebro argentino*, Planeta, ps. 269-270; Sigman, Mariano, *La vida secreta de la mente*, Debate, Buenos Aires, 2016, ps. 133-134, 197 y ss.; Kerkhoff, Georg - Kalmbach, Dorothea - Rosenthal, Alisha, “Salir de un mundo a medias”, *Investigación y Ciencia*, nro. 78, ps. 60-65; Viosca, José, “¿Qué nos hace humanos?”, *Investigación y Ciencia —Mente & Cerebro—*, nro. 78, mayo-junio 2016, p. 52; Bachrach, Estanislao, *En cambio*, 5a ed., Sudamericana, Buenos Aires, 2015, p. 261; Kurzweil, Ray, *Cómo crear una mente*, Lola Books, p. 1; Siegel, Daniel J., *Viaje al centro de la mente*, Paidós, p. 53; Kaku, Michio, *La física del futuro*, Debolsillo, p. 116. Además de un modelo jerárquico de la estructura de la inteligencia, se habla de un “modelo factorial jerárquico”. Véase, *Diccionario de ciencias cognitivas*, Amorroutu, p. 133. Sobre los aspectos vinculados al proceso de toma de decisiones en las personas, véase, Kahneman, Daniel, *Pensar rápido, pensar despacio*, Debate, ps. 569-590. Ciertos estudios establecen la existencia de ocho redes parciales de “conectividad cerebral”. A saber: red mediofrontal, red frontoparietal, red en reposo, red de las áreas subcorticales y el cerebelo, red motora, red visual primaria, red visual secundaria y red de áreas de asociación visual. Ver, Wolf Christian, “El cerebro, un órgano personal”, *Investigación y Ciencia*, nro. 79, 2016, p. 69. También véase Fuster, Joaquín M., *Cerebro y libertad*, Ariel, Barcelona, 2014, ps. 101-115.

(23) La clasificación más habitual que se realiza ha sido introducida por Searle, J. R., “Minds, brains, and programs”, *Behavioral and Brain Science*, 3 (3), Cambridge (UK), 1980, ps. 417-457.

de aplicar diferentes técnicas informáticas. La idea básica, en esta conceptualización, es obtener resultados específicos en ciertas actividades o ámbitos concretos que antes solo podían obtenerse a partir de nuestros cerebros<sup>(24)</sup>. Idónea para un problema en un dominio definido *a priori*.

Mientras que los humanos transitamos un camino biológico de aprendizaje evolutivo, la IA se basa en algoritmos, datos históricos, computadoras, programación humana y, sobre todo, aprovechando tres características que superan por mucho nuestras capacidades cognitivas: velocidad de procesamiento, posibilidad de conectarse y articular con otros sistemas de forma instantánea y, por último, la enorme capacidad de almacenamiento de los datos e información. Se trata de versiones muy mejoradas del personaje de Borges Funes el Memorioso.

Este concepto de IA débil o restringida es el que sustenta al género aprendizaje de máquina o *machine learning* que abarca una serie de técnicas más o menos sofisticadas. En las redes neuronales, la actualmente más conocida es el aprendizaje profundo (*deep learning*)<sup>(25)(26)</sup>. Estas re-

---

(24) En las organizaciones públicas, la IA permite llevar adelante la transición de una burocracia impresa o digital, hacia una burocracia inteligente. Ampliar en Corvalán, Juan G., *Prometea. Inteligencia artificial para transformar organizaciones públicas*, Astrea, Ciudad de Buenos Aires, 2019, p. 29; Corvalán, Juan G., *Hacia una administración pública 4.0: digital y basada en Inteligencia Artificial. Decreto "Tramitación digital completa"*, La Ley, 2018, cita online AR/DOC/1683/2018.

(25) Según la ONU, el aprendizaje automático se vincula con la inteligencia artificial restringida que se usan en la capacitación de algoritmos para que usen conjuntos de datos a fin de reconocer y ayudar a resolver problemas. Por ejemplo, los dispositivos domésticos inteligentes dotados de IA "aprenden" continuamente, de los datos reunidos sobre el lenguaje cotidiano y las pautas lingüísticas, a procesar y responder más exactamente a las preguntas de sus usuarios (véase la res. 73/348 de la Asamblea General "Promoción y protección del derecho a la libertad de opinión y expresión", A/73/348; 29/08/2018, consid. 4º, disponible en <http://undocs.org/es/A/73/348>).

Además, véase Comisión Económica para América Latina y el Caribe, CEPAL, *Datos, algoritmos y políticas: la redefinición del mundo digital (LC/CMSI.6/4)*, Santiago de Chile, 2018, p. 171, disponible en <https://www.cepal.org/es/publicaciones/43477-datos-algoritmos-politicas-la-redefinicion-mundo-digital>).

(26) Una red neuronal artificial puede ser comprendida como una combinación masiva de unidades de procesamiento simple, que aprenden del entorno a través de un proceso de aprendizaje y almacenan el conocimiento en sus conexiones. Véase Haykin, Simon, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, New Jersey, 1999. En otras palabras, son un conjunto enorme de algoritmos y valores, donde cada componente de la red (neurona) es capaz de manejar distintas variables y arribar a distintos resultados, aprendiendo en conjunto.

Véase también: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, *El Correo de la UNESCO, Inteligencia artificial, promesas y amenazas*, París, julio-septiembre de 2018, p. 8, disponible en [https://unesdoc.unesco.org/ark:/48223/pf0000265211\\_spa](https://unesdoc.unesco.org/ark:/48223/pf0000265211_spa); la res. 72/540 de la Asamblea General, "El derecho a

des requieren de cantidades muy grandes de datos para ser “entrenadas”, y por su modo de funcionar, se asemeja a una “caja negra” (*black box*). Esto quiere decir que no es posible visibilizar el paso a paso (o algoritmo) de la lógica que explica la resolución del problema. En otras palabras, no se puede interpretar lo que sucede en las “capas ocultas de la red”<sup>(27)</sup>. Sobre esto volveremos más adelante. Es demasiada información que pasa frente a nuestra percepción y que ingresa a nuestro cerebro que no podríamos procesar.

### 3.1.2. IA FUERTE

Elon Musk es el director general de SpaceX, de Tesla y también es copresidente de OpenAI, una organización sin fines de lucro destinada a investigar la IA para hacerla amigable. Este “todólogo”, magnate y físico, quiere cambiar el mundo y la humanidad para lograr una sociedad “multiplanetaria”. Desde hace varios años, Musk y otras personalidades en el mundo de la innovación piden regulación porque, con mucha razón, piensan que la llegada de un tipo de inteligencia artificial herculeana llamada “fuerte o dura” podría poner en riesgo a la humanidad, principalmente en lo que se refiere a las armas autónomas<sup>(28)</sup>. La IA Hércules o fuerte representaría la transformación más importante de este siglo<sup>(29)</sup>, a las puertas de la Cuarta Revolución Industrial. Representa la fase final de transición de la IA débil, que incluye a los sistemas de IA que desarrollamos en este artículo, catalogados como IA débil o blanda.

La IA Hércules se relaciona con dos grandes fenómenos. En primer lugar, se vincula con el hecho de alcanzar algunos aspectos claves de la especie humana: el sentido común, la capacidad de “sentir”, de recono-

---

la privacidad”, A/72/540 (19/10/2017), consid. 131, disponible en <http://undocs.org/es/A/72/540>.

(27) Téngase en cuenta que la referencia a capas es específica a las redes neuronales.

“La IA moderna es, básicamente, una caja negra, que logra un desempeño superior al humano sin que las personas comprendan cabalmente cómo se obtiene ese resultado” (Comisión Económica para América Latina y el Caribe, CEPAL, “Datos, algoritmos y políticas: la redefinición del mundo digital”, LC/MSI.6/4, Santiago de Chile, 2018, p. 171, disponible en <https://www.cepal.org/es/publicaciones/43477-datos-algoritmos-politicas-la-redefinicion-mundo-digital>).

(28) Nota del director: sobre inteligencia artificial fuerte o superinteligencia como riesgo existencial para la humanidad ver López Oneto, Marcos, “Fundamentos antropológicos, éticos, filosóficos, históricos, sociológicos y jurídicos para la constitución universal de un Derecho de la Inteligencia Artificial (DIA)”, tomo I del presente *Tratado*. Sobre armas autónomas ver Martino, Antonio, “Ética y sistemas inteligentes”, apart. “La cosa se pone más peligrosa”, tomo I del presente *Tratado*.

(29) Kurzweil, Ray, *La singularidad está cerca*, Lola Books, p. 339.

cer el entorno y la llamada “autoconsciencia”. En segundo lugar, así como se alude a una inteligencia general humana que es producto de abarcar diferentes áreas de contenido, se busca desarrollar una inteligencia artificial general (IAG), que se traduce en una capacidad general de aprender. Es decir, se trata de simular el comportamiento o la inteligencia humana en un plano integral<sup>(30)</sup>. Si ambos aspectos pueden lograrse, sobre esta IA fuerte se podría producir una “explosión de inteligencia”. Inteligencia general, conocimiento profundo de su propio diseño, conocimientos específicos de programación para iniciar el círculo de automejora y “autoconsciencia”, constituyen la llamada IA fuerte que para muchos nos llevará a una explosión de la inteligencia<sup>(31)</sup>.

Como estos sistemas todavía no se han desarrollado, es importante considerar que es meramente conjetural la diferencia entre IA fuerte e IA débil, y está fuertemente correlacionada con los alcances que se le asigna a la inteligencia humana y a dónde se desee poner el acento. Por ejemplo, cuando las máquinas simulan o actúan como si fueran inteligentes en ciertos ámbitos o tareas concretas, se conoce como *débil, blanda o estrecha*. En cambio, cuando se afirma que las máquinas “realmente” demuestran inteligencia y no solo la simulan, entonces estaríamos en presencia de la *IA fuerte*<sup>(32)</sup>. Las máquinas aprendan de manera autónoma a ade-

(30) En algunas ocasiones también se habla de IA “general”, pero los términos no son exactamente iguales. Toda IA fuerte será general, pero, a la inversa, no tiene por qué siempre darse. En la comunidad científica, hay un debate intenso entre especialistas acerca de si esta clase de IA llegará y, eventualmente, cuándo hará su aparición.

Ver Instituto Español de Estudios Estratégicos, Documentos de Seguridad y Defensa 79, la Inteligencia Artificial aplicada a la defensa, iee.es, Ministerio de Defensa, publicación de marzo de 2019, <https://publicaciones.defensa.gob.es/la-inteligencia-artificial-aplicada-a-la-defensa-n-79-libros-ebook.html>.

Desde otra óptica, véase Bostrom, Nick, *Superinteligencia caminos, peligros, estrategias*, SL TEEEL Editorial, España, 2016.

(31) IJ Good fue el primero en desarrollar la noción de superinteligencia en 1965. Básicamente la “máquina ultrainteligente” es una que pueda superar “por mucho” todas las actividades humanas. Y una máquina ultrainteligente podría diseñar mejores máquinas, lo que sin duda generaría una “explosión de la inteligencia”. Actualmente, Bostrom, Kurzweil, y Barrat hablan de una superinteligencia o de una explosión de inteligencia. Bostrom, *Superinteligencia*, Teell Editorial, ps. 22 y ss.; Kurzweil, Ray, *La singularidad está cerca*, Lola Books, p. 297; Barrat, *Nuestra invención final*, Paidós, ps. 119 y ss. Según Kaku, desde 2070 hasta 2100 las máquinas llegarán a ser conscientes y esto implica al menos tres componentes: 1. Capacidad de sentir y reconocer el entorno; 2. Autoconsciencia; 3. Capacidad de planificar el futuro estableciendo objetivos y planes, es decir, simulando el futuro y desarrollando una estrategia. Kaku, Michio, *La física del futuro*, Debolsillo, ps. 142 y 144.

(32) Russell, S. - Norvig, P., *Artificial intelligence: a modern approach*, Pearson Education Limited, UK, 2016. Los conceptos fueron abordados también en el Módulo IV “Trabajando con máquinas inteligentes”, punto 2.3, del curso “Oxford Artificial

cuarse en diferentes roles y diferentes contextos. Otra vez, conocimiento situado.

Ahora bien, por un lado, es importante tener presente que todavía no se han desarrollado sistemas de IA que posean sentido común y la habilidad de manejar diversos campos de conocimiento a la vez<sup>(33)</sup>. Por otra parte, como hemos sostenido en otros trabajos, en vez de poner el foco en disquisiciones conceptuales acerca de lo que es o no es IA, es urgente abordar los beneficios, riesgos, desafíos, daños y, esencialmente, cómo diseñar ecosistemas de regulación que permitan que su desarrollo sea compatible con los derechos humanos<sup>(34)</sup>. Por eso es tan relevante separar cajas negras de cajas blancas y, a su vez, comprender cómo se puede

---

Intelligence Programme, Investigate the potencial of artificial intelligence and its implications for business”.

(33) Hay otra clasificación de la IA más sofisticada, que se encontraría en estadio posterior que podría llegar cuando una IA sea capaz de mejorarse a sí misma y, como consecuencia, esta versión mejorada podría crear otra aún más inteligente y así sucesivamente. Este tipo de IA que Yudkowsky y Bostrom llaman “IA seminal”, se basa en un automejoramiento recursivo que podría resultar en una explosión de inteligencia que nos lleve al surgimiento de una superinteligencia artificial. Yudkowsky, Eliezer, “Levels of Organization in General Intelligence”, in Goertzel, Ben - Pennachin, Cassio, *Artificial General Intelligence*, Cognitive Technologies, Berlin, Springer, ps. 389-501; Bostrom, Nick, *Superinteligencia*, 2a ed., Tell, 2014, p. 29.

“Inteligencia artificial” puede ser un término problemático, porque sugiere que las máquinas pueden funcionar según los mismos conceptos y reglas que la inteligencia humana. No es así. La inteligencia artificial generalmente optimiza la ejecución de tareas computadorizadas, asignadas por seres humanos, mediante repetición e intentos iterativos. Ver la res. 73/348 de la Asamblea General “Promoción y protección del derecho a la libertad de opinión y expresión”, A/73/348, 29/08/2018, consid. 3º, disponible en: <http://undocs.org/es/A/73/348>.

(34) Corvalán, Juan G., *Prometea. Inteligencia artificial para transformar organizaciones públicas*, Astrea, IMODEV, Universidad del Rosario y DPI Cuántico, Ciudad de Buenos Aires, 2019, [http://dpicuantico.com/libros/prometea\\_oea.pdf](http://dpicuantico.com/libros/prometea_oea.pdf).

Los actores de IA deben respetar el Estado de derecho, los derechos humanos y los valores democráticos a lo largo de todo el ciclo de vida. Cabrol, Marcelo - González A., Natalia - Pombo, Cristina - Sánchez A., Roberto, *Adopción ética y responsable de la inteligencia artificial en América Latina y el Caribe*, Fair Lac BID, p. 14, disponible en: [https://publications.iadb.org/publications/spanish/document/fAIR\\_LAC\\_Adopci%C3%B3n\\_%C3%A9tica\\_y\\_responsable\\_de\\_la\\_inteligencia\\_artificial\\_en\\_Am%C3%A9rica\\_Latina\\_y\\_el\\_Caribe\\_es.pdf](https://publications.iadb.org/publications/spanish/document/fAIR_LAC_Adopci%C3%B3n_%C3%A9tica_y_responsable_de_la_inteligencia_artificial_en_Am%C3%A9rica_Latina_y_el_Caribe_es.pdf).

Consejo de Europa. Carta ética europea sobre el uso de inteligencia artificial en los sistemas de justicia y su entorno, 3-4 de diciembre de 2018 (traducción realizada con Google Translate), disponible en: <https://campusialab.com.ar/wp-content/uploads/2020/07/Carta-e%CC%81tica-europea-sobre-el-uso-de-la-IA-en-los-sistemas-judiciales-.pdf>.

Cuando adquieren o despliegan sistemas o aplicaciones de inteligencia artificial, los Estados deben asegurar que los órganos del sector público actúen de conformidad con los principios de derechos humanos. La res. 73/348 de la Asamblea General

garantizar una intervención humana adecuada sobre todo el ciclo de vida de los sistemas de IA<sup>(35)</sup>.

### 3.2. UN MENÚ DE DEFINICIONES DE “INTELIGENCIA ARTIFICIAL”

Con las bases sentadas por las discusiones hasta el momento, estamos en condiciones de presentar un conjunto de definiciones planteadas desde distintos organismos, cada uno de los cuales ve a la IA como una disciplina que juega o jugará un rol central a nivel mundial. Recorramos, entonces, la selección de una veintena de tales definiciones, que luego analizaremos desde una óptica que nos ayudará a catalogarlas.

#### 3.2.1. ORGANIZACIÓN DE NACIONES UNIDAS (ONU)

*Definición 1:* la ONU define la IA como una constelación de procesos y tecnologías que permiten que las computadoras complementen o reemplacen tareas específicas que de otro modo serían ejecutadas por seres humanos, como tomar decisiones y resolver problemas. (res. 73/348 de la Asamblea General de la ONU<sup>(36)</sup>).

*Definición 2:* la inteligencia artificial es el arte de elaborar de sistemas capaces de resolver problemas y desempeñar tareas mediante la simulación de procesos intelectuales. Se puede enseñar a la inteligencia artificial a resolver un problema, pero ella también puede estudiar el problema y aprender la manera de resolverlo por sí misma sin intervención humana. (res. A-CN.9-960 de la Asamblea General de la ONU<sup>(37)</sup>).

---

“Promoción y protección del derecho a la libertad de opinión y expresión”, A/73/348, 29/08/2018, consid. 62), disponible en: <http://undocs.org/es/A/73/348>.

Nota del director: existe un gran consenso, entre los autores de este *Tratado*, en que la IA debe poseer un enfoque centrado en el ser humano, en los derechos humanos y en los valores democráticos durante todo su ciclo de vida. Específicamente, se ha planteado que en miras de construir un derecho a la inteligencia artificial (DIA), sus principios jurídicos fundamentales podrían inferirse del derecho internacional de los derechos humanos. Ampliar esta postura en: López Oneto, Marcos, “El derecho de la inteligencia artificial (DIA)”, apart. “Para un derecho de la inteligencia artificial (DIA)”, tomo I del presente *Tratado*.

(35) Según la OCDE, las fases del ciclo de vida de la IA son I) la planificación y el diseño, la recabación de datos y su procesamiento, así como la creación de modelos y su interpretación; II) la comprobación y la validación; III) el despliegue y IV) el funcionamiento y el seguimiento. Ver OECD Library, “La inteligencia artificial en la sociedad”, disponible en <https://www.oecd-ilibrary.org/sites/603ce8a2-es/index.html?itemId=/content/component/603ce8a2-es>.

(36) <https://undocs.org/es/A/73/348>.

(37) <https://undocs.org/sp/A/CN.9/960>.

### 3.2.2. PARLAMENTO EUROPEO

*Definición 3:* las modificaciones adoptadas por el Parlamento Europeo sobre la propuesta de Reglamento del Parlamento Europeo y del Consejo, por el que se establecen normas armonizadas sobre inteligencia artificial sostiene, que la inteligencia artificial es un sistema basado en una máquina que está diseñado para operar con diferentes niveles de autonomía y que puede, para objetivos explícitos o implícitos, generar resultados tales como predicciones, recomendaciones o decisiones, que influyan en entornos físicos o virtuales<sup>(38)</sup>.

*Definición 4:* la IA es un término que los encargados de formular políticas utilizan habitualmente al presentar sus estrategias sobre la gobernanza de los desarrollos tecnológicos y los desafíos económicos y sociales asociados, por las empresas que anuncian productos que muestran cierto grado de automatización, y medios de comunicación al presentar al público en general sus comentarios sobre la llamada cuarta revolución industrial.

No existe una noción única de IA y los criterios comúnmente utilizados para calificar ciertas categorías de aplicaciones de IA son amplios y a menudo indeterminados y, por lo tanto, inadecuados como base para adoptar una regulación clara y efectiva.

*IA para investigadores de IA:* la parte computacional de la capacidad de alcanzar objetivos en el mundo, que varían en tipos y grados en personas, animales, así como en algunas máquinas. Esta definición, con algunas variaciones, tiende a ser aceptada entre los investigadores de IA.

*IA como rama de la informática:* una rama de la informática que estudia las propiedades de la inteligencia mediante la síntesis de inteligencia.

### 3.2.3. ORGANIZACIÓN PARA LA COOPERACIÓN Y EL DESARROLLO ECONÓMICO (OCDE)

*Definición 5:* la OCDE sostiene que es posible afirmar que la IA constituye un sistema basado en la máquina que puede hacer predicciones, recomendaciones o tomar decisiones, influyendo en entornos reales o

---

(38) Modificaciones adoptadas por el Parlamento Europeo el 14 de junio de 2023 sobre la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas sobre inteligencia artificial (Artificial Intelligence Act) y se modifican determinados actos legislativos de la Unión, 14/06/2023, [https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236\\_EN.html](https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html) (consultado el 27/07/2023).

virtuales, sobre ciertos objetivos definidos por los humanos (OCDE: “Inteligencia Artificial en la sociedad”<sup>(39)</sup>).

*Definición 6:* la inteligencia artificial (IA) es una tecnología de propósito general que tiene el potencial de mejorar el bienestar y bienestar de las personas, contribuir a una actividad económica global positiva, aumentar la innovación y la productividad y ayudar a responder a los desafíos globales clave. Se implementa en muchos sectores que van desde la producción, las finanzas y el transporte hasta la atención médica y la seguridad (OCDE, “Recomendación del Consejo de Inteligencia Artificial”<sup>(40)</sup>).

*Definición 7:* un sistema basado en la máquina que puede, para un determinado conjunto de objetivos definidos por humanos, hacer predicciones, recomendaciones o decisiones que influyen en entornos reales o virtuales (OCDE, Gobernanza Pública nro. 34 “Estado de la Técnica en el Uso de Tecnologías Emergentes en el Sector Público”<sup>(41)</sup>).

#### 3.2.4. UNIVERSIDAD DE STANFORD

*Definición 8:* la inteligencia artificial (IA) es una ciencia y un conjunto de tecnologías computacionales que se inspiran en las formas en que las personas usan, pero generalmente funcionan de manera bastante diferente sus sistemas nerviosos y cuerpos para sentir, aprender, razonar y actuar. Si bien la tasa de progreso en la IA ha sido irregular e impredecible, ha habido importantes avances desde el inicio del campo hace sesenta años (Universidad de Stanford: “Artificial Intelligence and Life in 2030”<sup>(42)</sup>).

#### 3.2.5. COMISIÓN EUROPEA

*Definición 9:* la inteligencia artificial (IA) se refiere a sistemas que muestran un comportamiento inteligente al analizar su entorno y tomar acciones, con cierto grado de autonomía, para lograr objetivos específicos (Comisión Europea: “Artificial Intelligence for Europe”<sup>(43)</sup>).

---

(39) <https://www.oecd-ilibrary.org/sites/603ce8a2-es/index.html?itemId=/content/component/603ce8a2-es>.

(40) <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

(41) <https://ialab.com.ar/wp-content/uploads/2019/09/OECD-2019-Estado-de-la-t%C3%A9cnica-en-el-uso-de-las-tecnolog%C3%ADas-emergentes-en-el-sector-p%C3%BAblico.pdf>.

(42) [https://ai100.sites.stanford.edu/sites/g/files/sbiybj9861/f/ai100report-10032016fnl\\_singles.pdf](https://ai100.sites.stanford.edu/sites/g/files/sbiybj9861/f/ai100report-10032016fnl_singles.pdf).

(43) <https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe>.

*Definición 10:* la inteligencia artificial (IA) tiene como objetivo estudiar y desarrollar máquinas inteligentes y *software*. La investigación en TIC asociada incluye el desarrollo de *software* que se puede razonar, reunir conocimientos, planificar de forma inteligente, aprender, comunicarse, percibir y manipular objetos (Comisión Europea: “La responsabilidad por las nuevas tecnologías digitales”<sup>(44)</sup>).

*Definición 11:* la inteligencia artificial es una combinación de tecnologías que agrupa datos, algoritmos y capacidad informática. Los avances en computación y la creciente disponibilidad de datos son, por tanto, un motor fundamental en el pronunciado crecimiento actual de la inteligencia artificial (Comisión Europea: “Libro blanco sobre IA”<sup>(45)</sup>).

### 3.2.6. FOROS, CUMBRES Y OTROS EVENTOS NACIONALES E INTERNACIONALES CENTRADOS EN IA

*Definición 12:* poderoso procesamiento informático que utiliza grandes cantidades de información que puede procesar algoritmos complejos y hacer predicciones en tiempo real. (ITU “AI for Good Global Summit”<sup>(46)</sup>).

*Definición 13:* la IA se basa en el desarrollo de algoritmos que permiten a las computadoras procesar datos a una velocidad inusual (tarea que antes requería de varias computadoras y personas), logrando, además, aprendizaje automático. Los algoritmos se nutren de datos y experiencias recientes y se van perfeccionando, habilitando a la máquina con capacidades cognitivas propias de los seres humanos como visión, lenguaje, comprensión, planificación y decisión con base en los nuevos datos (BID: “Industria 4.0”<sup>(47)</sup>).

*Definición 14:* la IA no es un campo nuevo; gran parte de su sustento teórico y tecnológico fue desarrollado en los últimos 70 años por especialistas en ciencias de la computación, como por ejemplo Alan Turing, Marvin Minsky y John McCarthy. En la actualidad, el término se refiere a múltiples tecnologías que pueden combinarse de diferentes maneras

(44) <https://ec.europa.eu/digital-single-market/en/news/european-commission-staff-working-document-liability-emerging-digital-technologies>.

(45) [https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020\\_es.pdf](https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf).

(46) [https://www.itu.int/en/ITU-T/AI/Documents/Report/AI\\_for\\_Good\\_Global\\_Summit\\_Report\\_2017.pdf](https://www.itu.int/en/ITU-T/AI/Documents/Report/AI_for_Good_Global_Summit_Report_2017.pdf).

(47) <https://publications.iadb.org/es/industria-40-fabricando-el-futuro>.

para: sentir, comprender, actuar (Accenture: “Cómo la IA puede generar crecimiento en Sudamérica”<sup>(48)</sup>).

*Definición 15:* no existe una definición única de la IA que es universalmente aceptado por los médicos. Algunos definen AI en términos generales como un sistema computarizado que exhibe un comportamiento que es comúnmente considerado como que requiere inteligencia. Otros definen AI como un sistema capaz de resolver problemas complejos de forma racional o tomar las acciones apropiadas para alcanzar sus objetivos en cualquier circunstancia en la vida real que se encuentra (CONICET: “Preparación para el futuro de la inteligencia artificial”<sup>(49)</sup>).

*Definición 16:* inteligencia artificial es una inteligencia demostrada por las máquinas, en contraste con la inteligencia natural que se muestra por los seres humanos y otros animales. Se aplica el término “inteligencia artificial” sobre todo cuando una máquina imita las funciones cognitivas que se asocian con las mentes humanas, tales como el “aprendizaje” y “resolución de problemas” (UNICRI. INTERPOL: “Inteligencia artificial y robótica. Hacer cumplir la ley”<sup>(50)</sup>).

*Definición 17:* inteligencia artificial (AI). Un conjunto de métodos científicos, teorías y técnicas cuyo objetivo es reproducir, por una máquina, las capacidades cognitivas de los seres humanos. Los desarrollos actuales buscan tener máquinas realizan tareas complejas realizadas previamente por los seres humanos (CEPEJ: “Carta ética europea en el uso de la inteligencia artificial en los sistemas judiciales y su entorno”<sup>(51)</sup>).

*Definición 18:* la inteligencia artificial es un campo de la informática. La inteligencia artificial es la capacidad de una computadora digital o un robot controlado por ordenador para realizar tareas comúnmente asociados con los seres inteligentes. Estas tareas pueden incluir la percepción visual, reconocimiento de voz, toma de decisiones, y la traducción entre lenguas (Foro Europeo de la Discapacidad: “Una perspectiva de la discapacidad en la inteligencia artificial”<sup>(52)</sup>).

---

(48) [https://www.accenture.com/\\_acnmedia/pdf-49/accenture-como-la-ia-puede-generar-crecimiento-en-sudamerica.pdf](https://www.accenture.com/_acnmedia/pdf-49/accenture-como-la-ia-puede-generar-crecimiento-en-sudamerica.pdf).

(49) <https://drive.google.com/drive/u/2/folders/10G3JNpTbhZk10nT40YAJvn8bZixi4rkO>.

(50) <http://213.254.5.198/artificial-intelligence-and-robotics-law-enforcement>.

(51) <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>.

(52) <http://www.edf-feph.org/sites/default/files/edf-emerging-tech-report-accessible.pdf>.

*Definición 19:* la IA no es una tecnología específica, sino que comprende un conjunto de procesos basados en el aprendizaje automático que, a partir de grandes cantidades de datos y algoritmos potentes, desarrolla predicciones sólidas para dotar de una mayor eficacia a los procesos existentes (CIPPEC: “La política de la inteligencia artificial: sus usos en el sector público y sus implicancias regulatorias”<sup>(53)</sup>).

*Definición 20:* tecnología de la información que realiza tareas que normalmente requerirían capacidad intelectual biológica, como dar sentido al lenguaje hablado, aprender comportamientos o resolver problemas (Gobierno de Canadá: “Directiva sobre toma de decisiones automatizada”<sup>(54)</sup>).

Una forma de interpretar estas y otras definiciones de IA es la provista por Russell y Norvig en su clásico libro de texto (conocido por las siglas de su título: “AIMA”)<sup>(55)</sup>. Dichos autores proponen dos dimensiones: una de pensamiento vs. acción, y otra de racionalidad<sup>(56)</sup> vs. humanidad. Así, los autores catalogan cada uno de los cuatro cuadrantes de la siguiente manera:

*Sistemas que actúan de forma humana*, o el abordaje según el Test de Turing. En 1950<sup>(57)</sup>, Alan Turing propuso un test llamado “el juego de la imitación” (posteriormente, pasó a ser conocido por el nombre de creador) para dictaminar si un sistema es inteligente. Su idea fue simple pero poderosa: un sistema es inteligente si puede hacerse pasar por humano interactuando por un sistema de conversación por texto con otra persona. Si bien la definición tiene los méritos de capturar una gran parte de las áreas que hoy siguen teniendo mucho trabajo por realizar (tales como procesamiento del lenguaje natural, representación de conocimiento, razonamiento automatizado, aprendizaje, visión y robótica), las limitaciones de compararse con el comportamiento humano también son claras. En las palabras de Russell y Norvig, la ingeniería aeronáutica no define su disciplina como “la construcción de máquinas que vuelan de forma tan similar a las palomas que hasta pueden engañar a otras palomas”.

(53) <https://www.cippec.org/wp-content/uploads/2019/10/185-DT-Abdala-La-croix-y-Soubie-La-pol%C3%ADtica-de-la-Inteligencia-Artifici...pdf>.

(54) <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592#appA>.

(55) Stuart, Russell - Norvig, Peter, *Artificial intelligence-a modern approach*, 3rd ed., Pearson, 2016.

(56) Término difícil de definir formalmente. Aquí, como volveremos a discutir más adelante, simplemente diremos que un sistema es racional si toma la decisión correcta dado su conocimiento en el momento de tomarla.

(57) Turing, Alan (October 1950), “Computing Machinery and Intelligence”, *Mind*, LIX (236): 433-460.

*Sistemas que piensan de forma humana*, o el abordaje del modelado cognitivo: aquí, el objetivo se centra en desarrollar una teoría de la mente humana capaz que modelar los procesos que rigen los pensamientos de las personas. La disciplina de la ciencia cognitiva busca estudiar la combinación de modelos computacionales con técnicas experimentales de otras disciplinas como la psicología para arribar a teorías sobre la mente humana.

*Sistemas que piensan de manera racional*, o el abordaje según las “leyes del pensamiento”: aquí nos encontramos con la lógica como disciplina que busca modelar las inferencias racionales. La aplicación de la lógica como formalismo de representación e inferencia es poderosa, pero se encuentra rápidamente con obstáculos computacionales (uso de demasiados recursos de tiempo y espacio, como se discutió anteriormente). Otro obstáculo formidable es la representación y manejo racional de conocimiento bajo incertidumbre.

*Sistemas que actúan de manera racional*, o el abordaje por agentes racionales: los agentes son simplemente entidades que actúan en un entorno de manera autónoma. Esta propuesta se centra en el desarrollo de agentes racionales, es decir, aquellos que actúan de manera tal de alcanzar el mejor resultado (o, en caso de incertidumbre, el mejor resultado esperado). Russell y Norvig señalan dos ventajas de este abordaje. Primero, es más general que el basado en las leyes del pensamiento, ya que la realización de inferencias de manera correcta es una de muchas maneras de alcanzar la racionalidad. Por otro lado, es más conducente a desarrollos científicos que los basados en el comportamiento o pensamiento humano.

A continuación, explicaremos las dos grandes ramas en la que típicamente se dividen los esfuerzos científicos e ingenieriles que buscan el desarrollo de herramientas inteligentes. En dicha exploración, volveremos a varios de los conceptos discutidos en las primeras dos secciones para ilustrarlos y profundizar su tratamiento, siempre buscando mantener un tratamiento accesible para no expertos.

#### IV. DOS RAMAS EN LA BUSCA DE LA REALIZACIÓN DE LA INTELIGENCIA ARTIFICIAL

Dada la variedad de definiciones de IA (provenientes de un amplio espectro de orígenes), la gran cantidad de disciplinas que se combinan en ella y el rango de aplicaciones que tienen sus resultados, no sorprenderá al lector saber que su desarrollo no ha sido uniforme a lo largo de su historia relativamente corta. En esta sección, daremos un panorama de este desarrollo estructurado alrededor de dos grandes ramas. Comenzaremos por la basada en conocimiento y luego presentaremos la basada en datos

(como veremos, existen otros nombres para cada una), para luego concluir con una visión —personal pero compartida por muchos expertos— que apunta a la necesidad de combinarlas para sacar el mayor provecho posible de las fortalezas de cada una.

Al igual que el resto de este artículo, esta presentación no busca ser completa (se requeriría un tomo completo para siquiera intentar esto), sino brindarle al lector una introducción suave a la disciplina. En relación con esto, dada la verdadera explosión de popularidad de la rama del *machine learning*, dedicaremos mayor tiempo a esta para aclarar la variedad de conceptos con los que el lector probablemente se encuentre en sus exploraciones individuales de la disciplina.

#### 4.1. INTELIGENCIA ARTIFICIAL BASADA EN CONOCIMIENTO

La primera forma de abordar el desarrollo de herramientas inteligentes que trataremos es la llamada IA basada en conocimiento, popularmente conocida como IA simbólica, representación del conocimiento, razonamiento automatizado, o por su nombre en inglés: *knowledge representation and reasoning*; para abreviar, utilizaremos entonces el acrónimo “KR” para referirnos a esta rama.

Como vimos en la discusión sobre las posibles definiciones, la tradición en IA en considerar al conocimiento como una de las piezas clave para lograr herramientas inteligentes. En su artículo de 1980<sup>(58)</sup>, Allen Newell hizo un resumen del rol que tiene el conocimiento en este tipo de investigación y desarrollo, proponiendo un nuevo nivel de abstracción para darle a este un rol central. A continuación, haremos un breve resumen de esta propuesta, a modo de presentación de los fundamentos de esta rama de la IA.

Los sistemas de cómputo típicamente<sup>(59)</sup> pueden verse a través de una metáfora de organización en *niveles* (también a veces llamadas *capas*), los cuales van abstrayendo las funciones más básicas de las partes que los implementan para alcanzar los comportamientos que proveen las funcionalidades para las que fueron construidos. Cada nivel tiene entonces un *medio* que es el que se procesa, *componentes* que proveen las primitivas para el procesamiento, *leyes de composición* que permiten que las com-

---

(58) Ver Newell, Allen, “The knowledge level”, *Artificial Intelligence*, 18.1 (1982): 87-127. Este artículo está basado en el primer discurso presidencial de la AAAI, una de las principales asociaciones de IA del mundo (al momento de su creación en 1979, las siglas significaban “American Association for Artificial Intelligence”, pero en 2007 fueron cambiadas a “Association for the Advancement of Artificial Intelligence” para reflejar su alcance mundial).

(59) Al menos los clásicos que obedecen al modelo de von Neumann.

ponentes sean ensambladas en sistemas y *leyes de comportamiento* que determinan cómo el comportamiento del sistema depende del comportamiento de las componentes y la estructura del sistema. Así, el nivel más básico se llama de Dispositivo (donde el medio está formado por electrones y dominios magnéticos, por ejemplo), el que sigue es de Circuitos (donde el medio es voltaje y corriente), sigue el de Lógica (que puede dividirse en subniveles de lógica de circuitos y de registros), donde el medio ya son los conocidos bits (0s y 1s) y el más abstracto es el Símbolos (donde tenemos expresiones simbólicas legibles más cómodamente por las personas, y residen los programas en lenguajes de programación como Java, C o Python). Lo interesante de esta metáfora es que no es simplemente una serie de abstracciones, sino que cada nivel es una especialización de la clase de sistemas capaces de ser descriptos en el nivel próximo hacia abajo. Otro aspecto importante es que los niveles son *aproximaciones*; un error en niveles inferiores se propaga a los superiores, generando manifestaciones que no pueden ser explicadas fuera del nivel donde se originó el error<sup>(60)</sup>.

El nuevo nivel propuesto por Newell es el nivel de Conocimiento, donde el sistema que reside en él se denomina *agente* (concepto que ya nos encontramos anteriormente), las componentes son *metas, acciones y conjuntos*, el medio es el conocimiento en sí y la ley de comportamiento es el *principio de la racionalidad* (también ya visto). Por lo tanto, tratar un sistema en el nivel de conocimiento es tratarlo como poseedor de conjuntos de conocimiento y metas, y creyendo que actuará dentro de sus posibilidades (es decir, dentro de sus limitaciones en cuanto a poder de acción y las limitaciones de su conocimiento) para alcanzar sus metas. Este nivel yace por encima del nivel de Símbolos; sus componentes y medio, por lo tanto, pueden definirse en términos de sistemas en el nivel de Símbolos. El autor señala algunos puntos acerca de la naturaleza del conocimiento en este entorno:

- El conocimiento está estrechamente relacionado con la racionalidad. Si un sistema puede ser visto como racional, entonces, puede ser visto como poseedor de conocimiento.
- El conocimiento es una noción similar a una *competencia*; es decir, tiene potencial para generar acción.
- El nivel de Conocimiento es una *aproximación* (en grado y alcance, como todos los niveles de sistemas). No hay garantías acerca

---

(60) Quien haya tenido que lidiar con una computadora que actúa erráticamente porque un banco de memoria tiene una falla, o porque algún componente se está calentando, sabe perfectamente a lo que nos estamos refiriendo.

de qué parte del comportamiento puede verse como parte del nivel de Conocimiento.

- Las *representaciones* del conocimiento existen en el nivel de Símbolos como sistemas (tales como estructuras y bases de datos, o procesos) que implementan un conjunto de conocimiento en el nivel superior.
- Las *lógicas* son tan solo una clase de representaciones entre muchas, aunque tienen capacidades únicas para el análisis del conocimiento y su representación.

Como ampliación de la definición informal de conocimiento vista al principio de este artículo, podemos utilizar la definición brindada por Newell para aclarar un poco más de qué se trata el medio que reside en este nivel:

Conocimiento: *todo aquello que puede ser adscrito a un agente, tal que su comportamiento puede ser computado de acuerdo al principio de la racionalidad.*

Recordemos que el principio de la racionalidad simplemente afirma que las acciones se eligen “de la mejor manera posible” de acuerdo con las metas y el conocimiento. Por lo tanto, esta caracterización del conocimiento es puramente *funcional*; es decir, esta concepción es diferente a la usualmente manejada por los filósofos, quienes plantean que el conocimiento está compuesto por creencias verdaderas justificadas<sup>(61)</sup>.

Una consecuencia casi directa de este tratamiento es que el conocimiento típicamente *no será finito*; para observar esto, basta con considerar todas las proposiciones que pueden hacerse acerca de las posiciones actuales de los 22 jugadores en un partido de fútbol (o, para tomar prestado el ejemplo de Newell en un entorno más estático, las posiciones actuales de las piezas en un tablero de ajedrez). Si bien uno podría objetar que tanto el agente como las observaciones que pueden hacerse acerca de las posiciones de los jugadores o piezas son finitas, el poder del nivel de Conocimiento radica en *abstraerse* de los procesos simbólicos, y, por lo tanto, no se puede fijar una cota sobre estos. A continuación, brindaremos un ejemplo concreto (muy simple y mundano) de cómo surge esto en la práctica. Por lo tanto, un sistema computacional genera elecciones de acciones para satisfacer metas, condicionadas en los estados del mundo; cada relación medios-fines de este tipo, entonces, puede verse como un “elemento” de conocimiento. Si bien podríamos ver a todo el conocimiento como una gran tabla de estos elementos, sería una tabla infinita y,

(61) Ver <https://plato.stanford.edu/entries/knowledge-analysis/>.

por lo tanto, no computable, por lo que debe crearse de manera dinámica por demanda. Aquí es donde radica la parte “inteligente” de las herramientas de IA basadas en conocimiento.

#### 4.1.1. UN EJEMPLO INTEGRADOR

Supongamos que queremos desarrollar una aplicación para que un gran número de usuarios (a nivel global a través de Internet) puedan almacenar sus datos de genealogía<sup>(62)</sup> y realizar consultas pertinentes a este dominio. Esta simple aplicación basta para ilustrar los principales conceptos que hemos presentado anteriormente. Como hicimos en la sección II, los símbolos de predicado *persona*, *padre* y *madre* se utilizan para representar datos acerca de qué personas “conoce”<sup>(63)</sup> el sistema, y quiénes son sus progenitores. Así, podemos tener la siguiente base de datos:

*persona*(Diego)

*persona*(Chitoro)

*persona*(Tota)

*persona*(Dalma)

*persona*(Gianinna)

*persona*(Claudia)

*padre*(Chitoro, Diego)

*madre*(Tota, Diego)

*padre*(Diego, Dalma)

*madre*(Claudia, Dalma)

*padre*(Diego, Gianinna)

*madre*(Claudia, Gianinna)

Estos datos codifican parte de la ascendencia inmediata ascendente y descendiente de una persona llamada Diego. Para mantener el ejemplo simple, solo se codifican los hechos de “ser persona”, “ser padre” y “ser madre”; es claro que con facilidad se podrían almacenar muchos datos más, tales como fecha de nacimiento y defunción, domicilio, teléfono, número de identificación, lugar de trabajo, etc. Con estos datos, un usuario podría realizar algunas consultas al sistema tales como: “¿Quién es el padre de Diego?”, “¿Quién es la madre de Gianinna?”, etc. Si bien esto puede ser interesante en casos de bases de datos más pobladas, se limitan a lo que puede verse como una simple búsqueda en una tabla.

---

(62) El término “genealogía” en general hace referencia a paternidad y maternidad biológica, pero también existen proyectos que mantienen la *genealogía académica* (quién fue director de quién). En particular, el *Mathematics Genealogy Project* (<https://www.mathgenealogy.org/index.php>) mantiene la información ancestral de los matemáticos (incluyendo a los informáticos allegados).

(63) Para evitar sobrecargar al lector, omitiremos las comillas de este tipo en el resto del ejemplo.

Consideremos ahora el agregado de conocimiento un poco más complejo; esto puede hacerse mediante las siguientes fórmulas (llamadas reglas existenciales), similares a las vistas en la sección II:

$$persona(X) \rightarrow \exists Y persona(Y), madre(Y,X)$$

$$persona(X) \rightarrow \exists Y persona(Y), padre(Y,X)$$

Esto codifica un conocimiento que todos albergamos: “toda persona tiene una madre y un padre, que son a su vez personas”; nótese que solo se establece la *existencia* de dichas personas y su relación de ascendencia, lo cual no necesariamente significa que se las tiene *identificadas*. Claramente, esto lleva a inferencias infinitas, ya que la aplicación de estas reglas a un dato como *persona(Chitoro)* nos lleva a obtener:

$$persona(z_1)^{(64)}, madre(z_1, Chitoro) \text{ y } padre(z_1, Chitoro).$$

Pero ahora tenemos una nueva persona ( $z_1$ ), por lo que al aplicar nuevamente la regla tendremos a su vez:

$$persona(z_2), madre(z_2, z_1) \text{ y } padre(z_2, z_1),$$

y así sucesivamente. Aquí vemos con abrumadora claridad la necesidad de codificar *estructuras potencialmente infinitas* ante conocimiento prácticamente del más básico que uno puede imaginar. Completemos el ejemplo para concluir esta discusión; dos fórmulas que podrían agregarse para completar la relación de ascendencia con la de *descendencia* (la relación dual) serían las siguientes:

$$madre(X,Y) \rightarrow hijo(Y,X)$$

$$padre(X,Y) \rightarrow hijo(Y,X)$$

Así, el sistema se ha enriquecido considerablemente, ya que ahora puede contestar preguntas que antes no podía, tales como: “¿tiene hijos Claudia?” y “¿tiene madre Chitoro?”. Esta última pregunta es especialmente interesante, dado que la base de datos no tiene conocimiento *explícito* de quién es la madre de Chitoro. Sin embargo, aplicando la regla, sabemos que esta persona *existe*, por lo que la pregunta tendrá respuesta

(64) Los símbolos de la forma *zi* se llaman *nulos*, y representan un valor desconocido hasta el momento. El procedimiento que aquí ilustramos mediante ejemplos recibe el nombre *chase* (del verbo en inglés que significa perseguir; en este caso, se “persiguen” las consecuencias de los datos y reglas disponibles) desarrollado en la comunidad de Teoría de Bases de Datos. Los artículos fundacionales fueron Aho, Alfred V. - Beeri, Catriel - Ullman, Jeffrey D., “The Theory of Joins in Relational Databases”, *ACM Trans. Datab. Syst.* 4(3): 297-314, 1979 y Maier, David - Mendelzon, Alberto O. - Sagiv, Yehoshua, “Testing Implications of Data Dependencies”, *ACM Trans. Datab. Syst.* 4(4): 455-469, 1979.

positiva; por otra parte, si se realiza la pregunta más particular “¿Quién es la madre de Chitoro?”, esta deberá responderse indicando desconocimiento. Así, podrían plantearse preguntas más complejas combinando estas relaciones de manera tan larga como uno desee: “¿tiene padre la madre de la madre del padre de la madre de Diego?” (claramente, la respuesta es afirmativa). Con este ejemplo vemos la forma en que esta estructura potencialmente infinita puede “domarse”, ya que para contestar esa pregunta basta con aplicar las reglas de paternidad/maternidad cinco veces para tener la información buscada. El hecho de *no poder plantear una pregunta infinita* nos lleva, entonces, a que *dada una pregunta* el procedimiento pueda detenerse luego de una cantidad finita de pasos<sup>(65)</sup>.

Por último, para ilustrar la flexibilidad que se tiene para expandir esta representación de conocimiento, supongamos que queremos incluir la relación de *abuela*; esto lo hacemos mediante dos fórmulas más que codifican las dos formas en que una persona puede ser abuela de otra:

$$\text{madre}(X,Y) \wedge \text{madre}(Y,Z) \rightarrow \text{abuela}(Z,X)$$

$$\text{madre}(X,Y) \wedge \text{padre}(Y,Z) \Rightarrow \text{abuela}(Z,X)$$

Es decir, una abuela es la madre de la madre o la madre del padre de una persona<sup>(66)</sup>.

Esta serie de simples ejemplos busca ilustrar los conceptos de conocimiento, mecanismos de acceso a este y la necesidad de manipular estructuras no acotadas (es decir, potencialmente infinitas). En la próxima sección veremos un abordaje diferente, aunque —como argumentaremos luego— no necesariamente incompatible.

#### 4.2. INTELIGENCIA ARTIFICIAL BASADA EN DATOS

Otra forma de abordar el desarrollo de herramientas inteligentes es la llamada IA basada en datos, popularmente conocida como IA subsimbólica, aprendizaje automatizado, aprendizaje de máquina, o por su nombre en inglés: *machine learning*; para abreviar, utilizaremos, entonces, el acrónimo “ML” para referirnos a esta rama.

---

(65) Esta discusión necesariamente simplificada no contempla otras formas en la que la cantidad de pasos a aplicarse puede resultar no acotada; la idea clave detrás de los lenguajes de ontología como este es restringir las formas que pueden tener las reglas y las consultas para poder garantizar que o bien el procedimiento se detenga luego de una cantidad finita de pasos (lo cual recibe el nombre de *decidibilidad*) o que —no bastando esto— se detenga luego de una cantidad *no demasiado grande* de pasos (lo cual se denomina *tratabilidad computacional*).

(66) Dos reglas análogas codificarían la relación de *abuelo*.

Si bien las técnicas y herramientas que yacen en los cimientos del ML comenzaron a desarrollarse hace unos 70 años atrás<sup>(67)</sup>, en las últimas dos décadas se han producido avances particularmente rápidos en este subconjunto de la IA en el que las máquinas aprovechan los enfoques estadísticos para aprender de datos históricos y realizar predicciones en situaciones nuevas. La madurez de las técnicas de aprendizaje automático, junto con los grandes conjuntos de datos y el aumento del poder computacional están detrás de gran parte de la expansión actual de la IA<sup>(68)</sup>. Antes de avanzar en la descripción de esta aproximación, es importante considerar tres grandes cuestiones asociadas.

*Primero.* Como mencionamos en la sección de KR, las técnicas de IA se basan en construir una base de conocimiento para detectar, predecir y reconocer patrones en los datos<sup>(69)</sup>. Esto se logra a partir de una combinación de computadoras, datos y conexiones a través de internet, algoritmos, lenguajes de programación y otros recursos para resolver problemas o tomar decisiones que antes solo podían ser realizadas por nuestras capacidades cognitivas.

*Segundo.* En los sistemas de IA basados en ML, la modalidad de aprendizaje está dividida entre los sistemas *supervisados* y los *no supervisados*, con matices entre ellos. Los primeros cuentan con una especie de tutor que los guía, mientras que los segundos delegan en sus algoritmos la tarea de descubrir por sí mismos patrones o similitudes que encierran los datos y que no son perceptibles fácilmente. Incluso, ambos abordajes se pueden combinar.

*Tercero.* Los sistemas inteligentes se basan en algoritmos para funcionar; por lo tanto, es útil repasar que la palabra “algoritmo” está asociada a un *conjunto de instrucciones, reglas o una serie metódica de pasos que puede utilizarse para hacer cálculos, resolver problemas y tomar decisio-*

---

(67) Ver, por ejemplo: Samuel, Arthur (1959), “Some Studies in Machine Learning Using the Game of Checkers”, *IBM Journal of Research and Development*, 3 (3): 210-229.

(68) Ver OCDE, “Inteligencia artificial en la sociedad”, 11/06/2019, disponible en: <https://www.oecd-ilibrary.org/sites/603ce8a2-es/index.html?itemId=/content/component/603ce8a2-es>.

(69) Ampliar en Domingos, Pedro, *The master algorithm: how the quest for the ultimate learning machine will remake our world*, Basic Books, New York, 2015, ps. 1 y ss.; Harari, Yuval Noah, *Homo Deus*, Debate, Ciudad Autónoma de Buenos Aires, 2016, ps. 99-107; Palma Méndez, José - Marín Morales, Roque, *Inteligencia artificial*, Mc Graw-Hill, Madrid, 2011, p. 683.

nes<sup>(70)</sup>. Los algoritmos son a la informática lo que los códigos procesales, procedimentales y los protocolos, son al campo jurídico<sup>(71)</sup>.

Como vimos al discutir el *nivel del Conocimiento* en la sección de KR, para resolver un problema mediante el desarrollo de herramientas informáticas, los programadores comienzan por pensar en el proceso de solución que se materializa en un algoritmo a partir del cual escriben su programa o “código fuente” en un lenguaje informático. Es decir, algo así como traducir del algoritmo al programa con base en el lenguaje elegido. Luego, otros programas (por caso compiladores o intérpretes) se encargan de traducirlo a lenguaje de máquina o lenguaje binario<sup>(72)</sup>.

En ML, estos algoritmos denominados “inteligentes” son la base de sistemas con capacidad para ejecutar instrucciones a partir de técnicas de aprendizaje automático. Por eso, funcionan de manera muy diferente de los algoritmos convencionales. Detectan patrones en los datos que los alimentan, lo que permite automatizar tareas, elaborar predicciones o previsiones y realizar detecciones inteligentes<sup>(73)</sup>. Mediante este abordaje, es posible, por ejemplo, resolver problemas que se caracterizan porque su proceso de solución tradicional resulta muy costoso, o incluso no se conoce una forma de resolución directa.

---

(70) Véase Benítez, Raúl - Escudero, Gerard - Kanaan, Samir - Masip Rodó, David, *Inteligencia artificial avanzada*, UOC, Barcelona, 2013, p. 14. Ampliar en Bostrom, Nick, *Superinteligencia*. 2, Tell, España, 2016, p. 29. Ver, además, “un conjunto de instrucciones específicas para realizar un procedimiento o resolver un problema, por lo general con el requisito de que el procedimiento termine en algún momento. A veces los algoritmos específicos son denominados método, procedimiento o técnica. El proceso de aplicar un algoritmo a un aporte para obtener un producto se denomina cálculo” (la res. 72/540 de la Asamblea General “El derecho a la privacidad”, A/72/540, 19/10/2017, consid. 52, disponible en <http://undocs.org/es/A/72/540>).

(71) Ver Corvalán, Juan G., “Inteligencia artificial y proceso judicial. Desafíos concretos de aplicación”, Diario DPI, 09/09/2019, disponible en: <https://dpicuantico.com/2019/09/09/el-impacto-de-la-ia-en-el-derecho-procesal/>; Corvalán, Juan G., “Inteligencia Artificial y proceso judicial. Desafíos concretos de aplicación”, Diario DPI, 30/09/2019, disponible en: <https://dpicuantico.com/sitio/wp-content/uploads/2019/09/Doctrina-Civil-30-09-2019-Parte-II-1.pdf>; Corvalán, Juan G. - Galletta, Diana U., “Intelligenza Artificiale per una Pubblica Amministrazione 4.0? Potenzialità, rischi e sfide della rivoluzione tecnologica in atto”, *federalismi.it*, rivista di diritto pubblico italiano, comparato, europeo, 6 febrero 2019.

(72) El lenguaje informático clásico (es decir, sin considerar los ordenadores cuánticos) se basa en transformar el lenguaje en ceros y unos; esto se basa en la facilidad de detectar voltajes que se encuentran por debajo o por encima de un cierto umbral.

(73) Véase res. 73/348 de la Asamblea General, “Promoción y protección del derecho a la libertad de opinión y expresión”, A/73/348, 29/08/2018, consid. 5º, disponible en <http://undocs.org/es/A/73/348>.

### 4.3. CAJAS NEGRAS Y OTRAS MÁS NEGRAS. DEEP LEARNING

Bajo el lema, “el futuro del pasado es el futuro del futuro” del Instituto Tecnológico de Massachusetts (MIT), emergen las predicciones a partir del uso de redes neuronales artificiales, imitando o copiando ciertos rasgos de los procesos neuronales de los cerebros humanos, que procesan la información a partir de neuronas, sinapsis, dendritas y axones, de las cuales las más actuales y popularizadas por su gran capacidad de respuesta acertiva de problemas en muy diversos dominios, son las de “aprendizaje profundo” o “*deep learning*”.

Las técnicas que utilizan los expertos en IA consisten en desarrollar algoritmos que implementan redes neuronales para reconocer la regularidad de los datos o patrones. Sin embargo, esto no significa que las redes neuronales artificiales funcionen igual a las biológicas.

Imaginemos que hacemos una de las 84.799 búsquedas que se realizan en Google por segundo<sup>(74)</sup>. Escribimos palabras como carteras, autos, flores, bananas, Jennifer Aniston, Luis Miguel, Trump, o el desafío viral “In My Feelings” basado en el popular tema del rapero Drake. Pero no deseamos ver texto sino imágenes. La vinculación entre las palabras y las imágenes que aparecen en la pantalla digital es producto del sistema de inteligencia artificial de Google llamado TensorFlow, que utiliza redes neuronales y aprendizaje automático. Si insertamos la palabra “despacito”, podremos ver en las caras de Luis Fonsi, Daddy Yankee y Justin Bieber, cómo los algoritmos inteligentes reaccionan a videos virales.

Las redes neuronales artificiales transforman sutilmente la frase que inmortalizó William Shakespeare con Hamlet. De “ser o no ser, esa es la cuestión” a “reconocer o no reconocer, esa es la cuestión”. También predicen, asocian, heteroasocian...

Los sistemas de IA que utilizan redes neuronales artificiales (RNA), frecuentemente obtienen mejores resultados para reconocer patrones cuando se trata de analizar cantidades masivas de datos (*big data*<sup>(75)</sup>). Son más eficientes, requieren intervención humana reducida y pueden trabajar información que los informáticos llaman “no estructurada” o no organizada bajo criterios concretos. Un Excel con atributos y datos etiquetados es información estructurada. Lo que la gente publica en las redes sociales es un ejemplo de información no estructurada. Lo no estructurado se explica muy bien cuando se alude a la información que se puede

(74) InternetLive Stats, <http://www.internetlivelivestats.com/> (consultado el 06/08/2020).

(75) Sosa Escudero, Walter, *Big data*, Siglo XXI Editores, 2019.

extraer de instancias que NO fueron previstas para la finalidad tratada. Por caso, en un diálogo, extraer preferencias de música o de películas. La estructurada es la que emerge de instrumentos y procesos preparados para ese fin, como una encuesta, por ejemplo.

Las redes neuronales profundas tienen arquitecturas más complejas que sus antecesoras: sus “capas ocultas”. Aquí se procesan datos y se descubren correlaciones que luego se traducen en resultados más eficientes y que pueden resolver ciertos problemas de una manera inesperada. Incluso, los avances de las nanotecnologías contribuyen a esta explosión de contribuciones de soluciones basados en *deep learning*.

Volvamos a la inteligencia biológica. En nuestro cerebro se disparan y activan los reconocedores de patrones. Si observamos un animal, la detección de un hocico activa una neurona que “avisa” a las demás acerca de que podríamos estar viendo a un perro. Si, además, de un hocico observamos plumas, se inhibe el reconocimiento y luego deberá procesar más información para que pueda decidir si finalmente es o no un perro. Este proceso, simplificado de manera extrema, se da en una fracción de segundo. Las neuronas en nuestros cerebros pueden ser inhibidas o estimuladas en función de ciertos patrones que pueden o no “dispararse”. En otras palabras, alcanzado determinado nivel de estímulo, llamado “umbral de excitabilidad”, la neurona transmite información a todas las demás neuronas a las cuales se encuentra conectada. De este modo, el estímulo recibido es procesado y transmitido a través de las distintas células. Resumidamente, este proceso es lo que explica cómo reconocemos patrones de información para resolver problemas o alcanzar objetivos.

En la inteligencia artificial, cuanto más representativa, precisa y etiquetada sea la muestra de la historia de los datos, es más probable que el sistema acierte más y se equivoque menos. A este fenómeno se lo conoce como mejorar la tasa de acierto y evitar falsos positivos y falsos negativos. En una versión más sofisticada de esta explicación, las 20.000 imágenes de los perros contienen diversos datos e información que la red evalúa y les asigna diferentes pesos o importancia. El primer dato recibido puede ser más determinante que el segundo para lograr la activación, así como el tercero puede ser más relevante que ambos, entre múltiples posibilidades. No todos los datos ingresados son igual de importantes para una red neuronal artificial: algunos pueden ser muy estimulantes, otros poco estimulantes, y otros incluso inhibitorios<sup>(76)</sup>. Luego de transmitir informa-

---

(76) Por otro lado, la neurona artificial presentará una “activación”, resultado de aplicar ese estímulo a un algoritmo. Por ejemplo, el número resultante de la suma de todos los estímulos puede ser incluido en un cálculo cuyo resultado final sea activación o no activación —o, en números, uno o cero—, o activación e inhibición —uno

ción desde la primera capa a la última, las neuronas de la red se activarán ante distintos supuestos: por ejemplo, si en la imagen hay un perro, un gato, un auto, un tractor, entre millones de posibilidades<sup>(77)</sup>. Esta explicación es muy relevante por dos aspectos.

El primero se vincula con el hecho de que las redes neuronales simulan comprensión de causalidad, pero no la realizan en los mismos términos que un humano. Es decir, nosotros experimentamos la realidad a través de los sentidos y también decidimos de forma emocional. Esto también es parte de razonar jurídicamente un fenómeno. En la IA, se correlacionan símbolos que son “muestras” de una realidad que es mucho más rica. Esos símbolos son patrones de datos que se correlacionan bajo formatos probabilísticos. A esto se lo conoce como una aproximación subsimbólica del mundo.

Por otra parte, en este proceso pueden alojarse los caballos de troya para el derecho. Los estímulos dependen de la cantidad y calidad de los datos ingresados, pero también de los pesos y valores que la propia red en sus capas ocultas correlaciona para llegar a un determinado resultado de caja negra. Si la muestra representativa presenta patrones de discriminación, la red, de comportarse como se espera, los reproducirá y puede que también los amplifique.

Ahora bien, dentro del ámbito de las redes neuronales artificiales, hay dos técnicas que se usan frecuentemente. Suponga que recibe muchos *emails* y que todavía no se inventó el *spam*. Usted tiene que entrenar al sistema para que le pueda filtrar los *mails* para saber cuándo está o no en presencia de un *email* que no desea recibir. Se filtra una carpeta para cada grupo, se entrena su red neuronal en estas dos carpetas (*spam/okey*)

o menos uno—, así como todos sus intermedios. En resumen, se ha emulado el funcionamiento de una red neuronal a través de distintos cálculos y algoritmos. Luego, a este proceso se le incorpora la posibilidad de modificar cómo cada neurona individual en la red procesa los distintos estímulos para activarse según si los resultados que obtiene son correctos e incorrectos para obtener un proceso de aprendizaje.

(77) Sobre todas estas cuestiones, ampliar en Winston, Patrick H., *Inteligencia artificial*, Pearson, ps. 477-505; Kurzweil, Ray, *La singularidad está cerca*, Lola Books, ps. 305-307 y 650-654 y el mismo autor en *Cómo crear una mente*, Lola Books, ps. 126-136; García Serrano, Alberto, *Inteligencia artificial*, Lola Books, ps. 208-209; Escolano Ruiz, Francisco - Cazorla Quevedo, Miguel Ángel - Alfonso Galipienso, María Isabel - Colomina Pardo, Otto - Lozano Ortega, Miguel Ángel, *Inteligencia artificial*, Paraninfo, ps. 91-118; Engelbrecht, Andries P., *Computational Intelligence*, Wiley, USA, 2005, ps. 6 y ss. En sentido técnico, los problemas que resuelve los sistemas de RNA son los que se derivan de la clasificación y la regresión. Los clasificadores se conocen como reconocimiento de patrones. Ampliar en Palma Méndez, José T. - Morales Marín, Roque, *Inteligencia artificial*, ps. 649 y ss., especialmente, ps. 651-652; Hofmann, Katja - Croft, W. Bruce, en línea: <https://www.microsoft.com/en-us/research/research-area/artificial-intelligence/>.

y se analizan los patrones de cada uno. De ese modo, se filtran automáticamente los *emails* subsecuentes a las categorías adecuadas y el sistema continúa entrenando con cada *mail* subsecuente<sup>(78)</sup>.

Otra forma se relaciona con una técnica (los modelos de Markov) para enseñar a los algoritmos inteligentes a que puedan reconocer el habla humana y que sirvan para desarrollar un asistente de voz en su dispositivo móvil, estilo Siri de Apple o Alexa de Amazon. En vez de enseñar los fonemas o las secuencias de los fonemas, se permite que el sistema descubra las reglas por sí mismo a partir de la transcripción de datos con miles de horas de habla humana. Estos sistemas generan reglas sutiles de las que los expertos humanos no tienen que necesariamente estar al tanto.

*Aprendizaje supervisado.* Muchos sistemas de IA se basan en desarrollos bajo un enfoque de aprendizaje supervisado, y esto es un punto crítico para la protección de los derechos de las personas. Sin embargo, no debe confundirse con la supervisión humana en todo el ciclo de vida de una IA, que constituye el principio rector para los desarrollos que tengan impacto en los derechos de las personas<sup>(79)</sup>.

Tanto en las cajas blancas como en las cajas negras, se habla de aprendizaje supervisado y no supervisado, para referirse a la forma en que au-

---

(78) Esto es un ejemplo de aplicación de una red bayesiana. Estas técnicas de IA se vinculan con la aplicación de un método trabajado por un matemático llamado Thomas Bayes. Los métodos que sustentan la utilización de estas redes suelen ser eficientes para predecir las probabilidades de sucesos futuros, basándose en hechos similares del pasado. Ampliar en Kurzweil, Ray, *La singularidad está cerca: cuando los humanos trascendemos la tecnología*, Lola Book, ps. 304-305; Escolano Ruiz, Francisco - Cazorla Quevedo, Miguel Ángel - Alfonso Galipienso, María Isabel - Colomina Pardo, Otto - Lozano Ortega, Miguel Ángel, *Inteligencia artificial: modelos, técnicas y áreas de aplicación*, Thomson, 2003, ps. 119-123.

(79) Téngase en cuenta que para garantizar que el sistema es compatible con los derechos, es necesario someter de manera constante a procesos de verificación, validación y supervisión.

Ampliar en “El ciclo de vida de un sistema de información”, disponible en: <http://flanagan.ugr.es/docencia/2005-2006/2/apuntes/ciclovida.pdf>.

Comisión Europea, “Generar confianza en la inteligencia artificial centrada en el ser humano”, Bruselas, 08/04/2019, p. 2, disponible en: <https://ec.europa.eu/transparency/regdoc/rep/1/2019/ES/COM-2019-168-F1-ES-MAIN-PART-1.PDF>.

Más allá de los nuevos principios que nacen a causa de la inteligencia artificial, se ha sostenido que quizás sea necesario realizar ajustes incorporando nuevos derechos relacionados con la tecnología, pero la clave principal pasará por interpretar y aplicar los derechos normativamente existentes desde una mirada poshumanista. Ver Gil Domínguez, Andrés, *Inteligencia artificial y derecho*, Rubinzal-Culzoni, noviembre 2019, cap. 1, “El poshumanismo y los derechos”, disponible en: <https://ialab.com.ar/wp-content/uploads/2020/05/Gil-Domi%CC%81nguez-Andre%CC%81s-Inteligencia-Artificial-y-Derecho-Capi%CC%81tulo-1-1.pdf>.

tomáticamente el sistema aprende. Cuando es supervisado, hay el algoritmo de aprendizaje se “entrena” con un conjunto de datos del cual ya conoce los resultados que la red debería entregarle<sup>(80)</sup>.

En esencia, en el aprendizaje supervisado los aprendices son los algoritmos y sus entrenadores son los datos que se usan para el procesamiento de los programas y técnicas informáticas. El aprendizaje presupone dividir conjuntos de datos que se llaman “*data sets* de entrenamiento” y “*data sets* prueba”, entre otras denominaciones. En el proceso de preparación de los datos, son los humanos quienes lleven adelante el proceso de etiquetar los ejemplos en para que la máquina pueda identificar palabras, imágenes, voz, entre otros, y de esa forma validar los resultados de la detección de los patrones de información que surgen de ese conjunto de datos etiquetados.

Por ejemplo, si se trata de reconocer lenguaje natural y detectar patrones de información en sentencias, denuncias, dictámenes o demandas, hablamos del aprendizaje y la supervisión acerca de las correlaciones entre palabras o frases para que se puedan extraer reglas de correlación sobre la especie o subespecie de decisión concreta, dentro de un grupo de posibilidades jurídicas y fácticas del género. Una vez que el sistema aprende que ciertas combinaciones de palabras siguen ciertas reglas, luego un programa o ciertos procedimientos pueden clasificar nuevos ejemplos en el conjunto de pruebas mediante el análisis de ejemplos que ya han sido aprobados por las personas humanas; es decir, tienen una etiqueta que indica su género, especie y subespecies en un conjunto de datos<sup>(81)</sup>.

*Aprendizaje no supervisado.* En el aprendizaje *no* supervisado o no guiado, el volumen de datos que se maneja no contiene información precisa o expresa ni implícita etiquetada *a priori* por una persona. Si queremos organizar una biblioteca, podemos comenzar por ordenar los libros según las categorías; pero si lo hiciera un algoritmo no supervisado probablemente la organizaría según similitudes que ellos mismos aprenden o detectan de los datos y que no son producto de la intervención humana, es decir, deciden autoorganizarse según criterios subsimbólicos.

Se trata de encontrar patrones o características que sean significativas y que evidencian similitudes en los datos de entrada, ya que no se establece ninguna salida con la que comparar el rendimiento del méto-

---

(80) Massachusetts Institute of Technology, Machine Learning from Data Decisions, 23/04/2019, Módulo 1.

(81) Ver, Learned-Miller, E., *Introduction to Supervised Learning*, Department of Computer Science University of Massachusetts, Amherst, 2014, p. 2, (en línea) <https://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a.pdf>.

do<sup>(82)</sup>. Por ejemplo, puede iniciarse a partir de ciertas categorías de los datos a partir de sus semejanzas. Es decir, aquí no se establece una salida deseada y tampoco el objetivo es encontrar un mapeo entrada-salida. La esencia de un sistema de aprendizaje no supervisado es su capacidad autoorganizativa.

Ahora bien, estas categorías de supervisado y no supervisado (entre otras como el aprendizaje recursivo<sup>(83)</sup>) no se presentan de manera aislada cuando se entrena a un sistema de IA. Por el contrario, la idea es tratar de mezclar técnicas y tácticas algorítmicas que más se ajusten a los problemas concretos que se intenta resolver<sup>(84)</sup>.

#### 4.4. APRENDIZAJE PROFUNDO (DEEP LEARNING) MÁS AUTOAPRENDIZAJE AUTÓNOMO

Desde el famoso duelo entre Deep Blue de IBM y Kasparov en 1997, se han llevado a cabo cientos de desafíos entre máquinas y humanos. Hace algunos años se realizó una competencia entre una IA y un campeón humano en el famoso juego creado en China hace más de dos mil quinientos años: el “Go”. En este juego, hay un tablero para dos jugadores. El objetivo es que uno de los jugadores rodee con piedras un área mayor en el tablero que su oponente. Al final del juego, se puntúa y el jugador que tenga mayor territorio gana la partida. En el ajedrez, normalmente se pueden realizar unos 37 movimientos de media. En el juego Go, una partida profesional en el tablero más grande suele oscilar entre 150 y 250 posibilidades. Aunque las reglas son simples, la estrategia es muy compleja y hay que equilibrar muchos requisitos, algunos contradictorios. Por ejemplo, ubicar piedras juntas ayuda a mantenerlas conectadas. Por otro lado, colocarlas separadas hace que se tenga influencia sobre una mayor porción del tablero y eso abre la posibilidad de apropiarse de un territorio más extenso.

En marzo de 2016 se batieron a duelo un campeón del mundo humano con una IA: Ke Jie vs. *AlphaGo* de Google. La victoria fue para AlphaGo

---

(82) Existen tres grandes grupos de métodos de aprendizajes no supervisados en el ámbito de las redes neuronales artificiales. Los que se basan en las reglas de HEB, los competitivos y los modelos basados en la teoría de la información. En estos últimos, se trata de maximizar la cantidad de información que se conserva en el procesamiento de los datos. Por un lado, también múltiples métodos de aprendizaje supervisado (redes de neuronas de una capa, redes de base radial, aprendizaje adaptativo, de segundo orden, entre muchos otros). Palma Méndez, José T. - Morales Marín, Roque, *Inteligencia artificial*, McGraw-Hill, ps. 652-683.

(83) Kurzweil, Ray, *Cómo crear una mente*, Lola Books, p. 61.

(84) Véase, Abu-Mostafa, Yaser, *Técnicas de aprendizaje automático*, especial *Inteligencia artificial*, *Investigación y Ciencia*, abril 2003, p. 52. Ver: <https://dialnet.unirioja.es/servlet/articulo?codigo=4192606>.

4 a 1. Y al igual que acontece en otros ejemplos, el aprendizaje de esta IA se basa en la utilización de una base de datos de alrededor de 30 millones de movimientos. Se intenta imitar el juego humano, tratando de emular los movimientos de los jugadores expertos de juegos históricos registrados<sup>(85)</sup> y continuar aprendiendo a partir de las nuevas experiencias.

Hasta acá, nada nuevo bajo el sol del aprendizaje profundo o del llamado *deep learning*. Sin embargo, en el 2017 crearon a Alpha Go Zero que superó a su versión previa (AlphaGo) 100 a 0<sup>(86)</sup>. Mientras las versiones anteriores se entrenaron a partir de cientos de jugadas de seres humanos expertos en el juego Go, a Zero solo se le dieron las reglas del Go y una retroalimentación respecto de la posición de las distintas piedras del tablero y cómo transcurría la jugada. En otras palabras, comenzó como una hoja en blanco, sin ninguna idea de posibles jugadas. A partir de esto, el proceso de aprendizaje se logró porque Zero jugó miles o millones de veces contra sí misma. A pesar de que esta IA comenzó simplemente colocando piedras al azar en el tablero, luego de 3 horas ya jugaba como un ser humano principiante y, en tres días, había derrotado a sus predecesores que son las que habían logrado la proeza de derrotar a los expertos humanos. Para ponerlo en cifras, jugó contra sí misma unas *cuatro millones novecientas mil partidas*, que le permitieron derrotar a su versión anterior, en 72 horas<sup>(87)</sup>.

En conclusión, AlphaGo Zero derrotó a su anterior versión entrenada bajo aprendizaje supervisado y que había derrotado a más de 60 expertos humanos en juegos online<sup>(88)</sup>. Según Elon Musk<sup>(89)</sup>, Alpha GoZero se

(85) En octubre de 2015 se convirtió en la primera máquina de Go en ganar a un jugador profesional de Go sin emplear piedras de *handicap* en un tablero de 19x19. Véase, Silver, D. *et al.*, "Mastering the game of Go with deep neural networks and tree search", *Nature*, 529, 484-489 (2016); <http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html> (consultado 20/10/2017).

(86) La publicación realizada sobre AlphaGo Zero puede ser obtenida en: Silver *et al.*, "Mastering the game of Go without human knowledge", disponible en: <https://www.nature.com/nature/journal/v550/n7676/full/nature24270.html> (consultado 06/06/2018).

(87) Estos tiempos pudieron lograrse con una capacidad de cómputo exponencial. En algunos años, tal vez dichos tiempos podrán lograrse con máquinas personales.

(88) Ampliar en blog oficial de DeepMind en <https://deepmind.com/blog/alpha-go-zero-learning-scratch/> (consultado 06/06/2018). Los partidos que AlphaGo Master jugó contra humanos pueden verse en <https://deepmind.com/research/AlphaGo/match-archive/master/> (consultado 06/06/2018).

(89) Le, Quoc - Zoph, Barret, "Using Machine Learning to Explore Neural Network Architecture", 2017. Disponible en: <https://ai.googleblog.com/2017/05/using-machine-learning-to-explore.html> (consultado 05/06/2018). También, Zoph, Barret - Vasudevan, Vijay - Shlens, Jonathon - Le, Quoc, "AutoML for large scale image clas-

puede autoentrenar con las reglas de cualquier juego y ganar a cualquier humano<sup>(90)</sup>. ¿Inteligencia artificial tan débil?

Zero es un ejemplo de una caja negra que se vuelve más negra. Es decir, es un ejemplo de IA que puede “independizarse” de los humanos<sup>(91)</sup>. Estas clases de IA que seguirán escalando, profundizan radicalmente la problemática para la disciplina y plantean dificultades sistémicas previas a cualquier consideración jurídica.

La primera se vincula con la imposibilidad de pronosticar el grado de avance de la IA. Los que saben más del tema, los que están a la vanguardia, suelen fallar en los pronósticos; tanto para sobreestimar las capacidades de las IA hoy en día, como subestimar su actualidad y potencialidad. Era imposible imaginar que 20 años después de ganar al ajedrez, la IA aprendería sin ninguna intervención, vencer a los mejores humanos en los juegos que se proponga. Ni Ray Kurzweil previó exactamente esto, aunque suele ser el oráculo humano más preciso para pronosticar el avance de las tecnologías de la información y de la comunicación (TIC)<sup>(92)</sup>.

La segunda es más preocupante aún. Quién está a la vanguardia en temas de IA, nos dice: “La IA es capaz de mucho más de lo que casi nadie sabe y la tasa de mejora es exponencial”. También lanzan esta advertencia: “La IA es mucho más peligrosa que las armas nucleares”. Lo afirma Elon Musk, uno de los actores industriales centrales a nivel mundial, que, además, basa todo su modelo de negocios en la IA<sup>(93)</sup>.

---

sification and object detection”, <https://ai.googleblog.com/2017/11/automl-for-large-scale-image.html> (consultado 05/06/2018).

(90) Entrevista disponible en YouTube titulada “Elon Musk Answers Your Questions”, publicada en el canal de SXSW.

(91) Pero también hay proyectos concretos en donde es la propia IA que crea a otro sistema de inteligencia artificial. El resultado, al que inicialmente llamaron “AutoML”, fue una red neuronal artificial cuyo producto era otra red neuronal artificial. Es decir, entre la red controladora y la red controlada, fluye la información para el aprendizaje. Cuando una inteligencia artificial crea a otra no debe esperar nuevos meses. En realidad, son ciclos que pueden durar horas o algunos pocos días. Es más, los expertos humanos, no pueden comprender en su totalidad cómo es el proceso de creación y entrenamiento.

(92) Véase el análisis de sus propias predicciones de sus cuatro libros. En *La era de las máquinas espirituales* repasa el grado de acierto de las predicciones que realizó en *La era de las máquinas inteligentes*. Luego hizo lo mismo en *La singularidad está cerca*. En su último libro publicado en el 2013, *Cómo crear una mente*, hizo un repaso más detallado de sus predicciones.

(93) Entrevista disponible en YouTube titulada “Elon Musk Answers Your Questions”, publicada en el canal de SXSW, véase minuto 35 y ss.

En la primera edición de este *Tratado*, no había tantas voces que advertían acerca de los riesgos de esta tecnología<sup>(94)</sup>. Con la llegada de los grandes modelos de lenguaje tipo ChatGPT y la evolución de la IA generativa<sup>(95)</sup>, llegó la suscripción de una carta abierta para pausar los proyectos de IA, que ha sido firmada por más de 30.000 investigadores, empresarios, profesores y profesoras de las más prestigiosas universidades del mundo. Este llamado de casi dos carillas de extensión, busca pausar durante seis meses los sistemas de IA más avanzados que GPT-4<sup>(96)</sup>.

La tercera cuestión se relaciona con un combo de medidas que los Estados deben adoptar. Por un lado, hay que focalizar y matizar los diferentes riesgos. En el caso de las IA que autoaprenden, hay que desarrollar estrategias estatales que prioricen los riesgos asociados su desarrollo<sup>(97)</sup>. Por ejemplo, los que son similares a AlphaGo Zero, que deben ser identificados para que se puedan aplicar estrictas medidas de control y seguridad<sup>(98)</sup>.

Por todo ello, es clave adoptar un enfoque basado en el principio de precaución o prevención que muchos países usan para gestionar potenciales perjuicios o daños que se pueden causar a las personas o al ambiente. Es una tarea muy difícil, porque hay que garantizar un equilibrio dinámico entre “no matar la innovación”, ya que la IA es una aliada del

(94) Nota del director: sobre los riesgos de la IA y las nuevas tecnologías ver: Martino, Antonio A., “Ética y sistemas inteligentes”, tomo I del presente *Tratado*.

(95) Nota del director: sobre la IA generativa ver: Corvalán, Juan G., “Qué trae de nuevo la IA generativa”, tomo I del presente *Tratado*.

(96) Al 28 de julio de 2023, la declaración fue firmada por un total de 33.002 personas, incluidos el empresario Elon Musk, el cofundador de Apple, Steve Wozniak, Yuval Nova Harari, reconocido profesor y escritor de inteligencia artificial, el director ejecutivo de la firma Stability AI, Emad Mostaque, además de investigadores de la firma DeepMind. Ver más en “La carta en la que más de 1.000 expertos piden frenar la inteligencia artificial por ser una amenaza para la humanidad”, *BBC News Mundo*, 29/03/2023, disponible en: <https://www.bbc.com/mundo/noticias-65117146> (consultado el 28/07/2023) Ver también <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> (consultado el 28/07/2023).

(97) Nota del director: sobre las recomendaciones y propuestas de regulación de inteligencia artificial por parte de distintos estados de la región y Europa ver: Corvalán, Juan G. (dir.) - Sánchez Caparrós, Mariana - Rabán Melisa (coord.) - Heleg, Gislle - Papini, Carina M. - Stringhini, Antonella - Bonato, Valentín (equipo de trabajo e investigación), “Propuestas de regulación y recomendaciones de inteligencia artificial en el mundo”, tomo I del presente *Tratado*.

(98) Véase [https://www.elconfidencial.com/tecnologia/2017-02-02/inteligencia-artificial-elon-musk-stephen-hawking-ia\\_1325057/](https://www.elconfidencial.com/tecnologia/2017-02-02/inteligencia-artificial-elon-musk-stephen-hawking-ia_1325057/). El Principio 22 de Asilomar establece que los sistemas de IA diseñados para automejorarse recursivamente o autorreplicarse de una forma que pudiera llevar al rápido incremento en su calidad o cantidad deben estar sujetos a unas estrictas medidas de control y seguridad.

desarrollo sostenible, y al mismo tiempo aplicar el principio precautorio o de precaución<sup>(99)</sup> para contrarrestar riesgos, mitigar daños y proteger derechos.

#### 4.5. APRENDIZAJE AUTOMÁTICO Y CAJAS BLANCAS

Dentro de esta IA débil, blanda, estrecha o restringida, hay otro “mundo” de sistemas de IA que son de “caja blanca” y se basan en un conjunto de técnicas<sup>(100)</sup> que se utilizan para obtener predicciones, automatizaciones, clasificaciones o detecciones inteligentes. Gracias a las cajas blancas, los resultados a los que se arriba son auditables, trazables, explicables e interpretables, y ello resulta muy beneficioso para comprender la dinámica del tratamiento automatizado cuando se usan estas técnicas. Y esto redundará en enormes beneficios para el campo jurídico, optimizar la tarea judicial y la transformación digital de las organizaciones.

En este ecosistema de *machine learning* o aprendizaje automático de caja blanca, existen dos grandes técnicas que pueden usarse para realizar predicciones. Nos referimos a las técnicas de “Regresión” y “Clasificación” y, en esta última, se encuentra una subespecie llamada “Topic Model”<sup>(101)</sup>. Estas técnicas son categorizaciones de algoritmos supervisa-

---

(99) Los principios de Asilomar se desarrollaron después de que el Instituto Future of Life reuniera a docenas de expertos quienes consideraron la necesidad de crearlos para guiar el desarrollo de la IA en una dirección productiva, ética y segura; los mismos han sido apoyados por más de 1200 figuras relacionadas con la innovación tecnológica y científica como Stephen Hawking y Elon Musk, véase [https://www.elconfidencial.com/tecnologia/2017-02-02/inteligencia-artificial-elon-musk-stephen-hawking-ia\\_1325057/](https://www.elconfidencial.com/tecnologia/2017-02-02/inteligencia-artificial-elon-musk-stephen-hawking-ia_1325057/) y Véase Principios 19, 20 y 21 de Asilomar, <http://puente.digital/blog/blog/inteligencia-artificial-segun-stephen-hawking-y-elon-musk/>; el principio 19 establece la Capacidad de Precaución: Al no haber consenso, deberíamos evitar las asunciones sobre los límites superiores de las futuras capacidades de la IA; el principio 20 destaca que la IA avanzada podría representar un profundo cambio en la historia de la vida en la Tierra, y debería ser planificada y gestionada con el cuidado y los recursos adecuados; mientras que el 21 se refiere a que los riesgos asociados a los sistemas de IA, especialmente los catastróficos o existenciales, los cuales deben estar sujetos a planificación y esfuerzos de mitigación equiparables a su impacto esperado.

(100) Corvalán, Juan G., *Perfiles digitales humanos*, Thomson Reuters, 2020; Corvalán Juan G., *Prometea. Inteligencia Artificial para transformar organizaciones públicas*, Astrea, DPI Cuántico, Universidad del Rosario, IMODEV, 2018 disponible en: [https://dpicuantico.com/libros/prometea\\_oea.pdf](https://dpicuantico.com/libros/prometea_oea.pdf).

(101) El *Topic Model* es una herramienta estadística utilizada en *machine learning* y en aplicaciones de lenguaje natural que permite identificar temáticas en grupos de documentos de textos. El uso de *topic models* permite aplicar el análisis de *clustering* a conjuntos de datos no estructurados superando algunas de las limitaciones que presenta la herramienta de K-medias. Lo que hacen los *topic model* es tomar un texto no estructurado y aplicarlo a una dimensión más estructurada. Ver Instituto Tecno-

dos de aprendizaje automático que se obtienen mediante la diferenciación con respecto al tipo cuantitativo o cualitativo de la variable de salida involucrada en el problema. Es decir, si la salida de un problema es cuantitativa o cualitativa, nos referimos al problema como regresión o clasificación. Regresión significa que el resultado (variable dependiente) es cuantitativo y clasificación significa que el resultado es cualitativo. No importa si las entradas (variables independientes) son cuantitativas o cualitativas. Las técnicas que se usan tienen que ver con las salidas<sup>(102)</sup>. Por ejemplo, la regresión se usa para predecir la evolución de los precios de las propiedades en un determinado territorio.

En cambio, la clasificación en el ámbito del aprendizaje de máquina puede usarse en el ámbito del derecho para tratar de establecer correlaciones entre palabras o frases, y correlacionarlas con decisiones e hipótesis fácticas que están presentes en decisiones judiciales. A partir de esta técnica, en el año 2017 logramos predecir con una tasa de acierto superior al 96%, sobre ciertos casos judiciales en los que estaba en juego el derecho a la vivienda y otras materias en el ámbito del contencioso administrativo y tributario de la Ciudad Autónoma de Buenos Aires. Por ejemplo, desde el Laboratorio Innovación e Inteligencia Artificial de la UBA (IALAB), hemos combinado estas técnicas para detectar en segundos, si el vínculo causal entre un evento y el daño sostenido asociado a un caso de accidente de tránsito se ha fracturado<sup>(103)</sup>, con una tasa de acierto de más de un 95%.

El proceso de entrenamiento para su desarrollo constó de tres etapas: en la primera se utilizó un *data set* de sentencias que fue provisto por la jueza Gabriela Gil y por el secretario de Cámara Hernán Quadri. A partir de ahí, se elaboró un primer etiquetado manual, para que luego comience el proceso de aprendizaje de máquina de caja blanca “Clasificación”. En esta primera etapa, se realizó un análisis jurídico para extraer patrones comunes, que permitió identificar en las sentencias la existencia o interrupción del nexa causal en cada caso en particular. Luego, se

---

lógico de Massachusetts, *Programa en línea de Machine Learning*, 23 de abril de 2019 al 18 de junio de 2019.

(102) Ver, Lindholm, A. - Wahlström, N. - Lindsten, F. - Schön, T., “Supervised Machine Learning. Lecture notes for the Statistical Machine Learning course”, Department of Information Technology, Uppsala University, 2019, p. 8 (en línea), [http://www.it.uu.se/edu/course/homepage/sml/literature/lecture\\_notes.pdf](http://www.it.uu.se/edu/course/homepage/sml/literature/lecture_notes.pdf).

(103) Ver Corvalán, Juan G. - Le Fevre Cervini, Enzo M., “Inteligencia artificial en accidentes de tránsito: primera aplicación predictiva en el mundo para la justicia civil”, 2019, disponible en: <https://dpicuantico.com/sitio/wp-content/uploads/2020/08/Corval%C3%A1n-y-LeFevre.pdf>.

segmentó este grupo de sentencias en grupos y subgrupos, según las distintas hipótesis fácticas. Estos grupos y subgrupos fueron asociados a una determinada solución jurídica: la posibilidad de atribuir o no la responsabilidad por los daños a la parte demandada. La segunda etapa implicó transformar lo analizado a un lenguaje de programación. Por ejemplo, para la máquina es irrelevante muchas palabras que se usan para conectar frases (“y”, “o”, “que”, entre muchas otras).

Una vez finalizado el diseño de programación, se realizó la primera prueba predictiva sobre el universo total de sentencias. La tasa de acierto alcanzada fue del 81,4%, lo que obligó a analizar humanamente otro data set de sentencias, para intentar mejorar esa tasa. De esta forma, se inició la tercera y última etapa: el refinamiento de los patrones de información jurídica y de toda técnica que mejore el rendimiento predictivo. Es un trabajo conjunto entre los operadores jurídicos especializados y los expertos en inteligencia artificial, a efectos de lograr sinergia entre contenido legal y poder de máquina. La mejora fue impactante, la tasa de acierto alcanzó el 96,5% (83 aciertos sobre un total de 86 casos).

Este desarrollo actualmente se utiliza en un Juzgado Civil de la Provincia de Buenos Aires. Gabriela Gil, la jueza con la que entremos el sistema basado en la experiencia Prometea, lo usa bajo un enfoque de “control” de proyectos que se realizan en su juzgado. Verifica si la sentencia proyectada coincide con el resultado del predictivo que se entrenó con más de 400 sentencias de la Cámara Civil de Morón en la Provincia de Buenos Aires. La tarea de control dura segundos, ya que se ingresa al agente conversacional la sentencia y en pocos segundos se elabora un informe de predicción. El sistema es autoexplicable, ya que ofrece al usuario los métodos, las tasas de acierto y los *data sets* utilizados. Como veremos con mayor detalle al analizar el sistema PretorIA<sup>(104)</sup>, esta lógica de diseño y entrenamiento es plenamente aplicable a múltiples ramas y áreas del derecho, más allá de la justicia.

---

(104) Nota del director: sobre los principios que cumple el sistema PretorIA, entre ellos la trazabilidad, explicabilidad, interpretabilidad, ver Corvalán, Juan G. - Montenegro, Ana Paula - Papini, Carina M. - Puig, Matías - Pitrau, Osvaldo - Quartesan, Verónica - Stringhini, Antonella, “Protocolo sobre gobernanza algorítmica. Trazabilidad, explicabilidad, interpretabilidad y fiabilidad. Diseño, desarrollo y despliegue. Basado en el sistema PretorIA, desarrollado por UBA IALAB para la Corte Constitucional de Colombia”, tomo III del presente *Tratado*. Asimismo, ver Corvalán, Juan G. - Papini, Carina M. - Palacios, Mariela - Palumbo, Julián - Pitrau, Osvaldo - Quartesan, Verónica, “Instructivo de gobernanza de datos para sistemas predictivos. Basado en el sistema PretorIA, desarrollado por UBA IALAB para la Corte Constitucional de Colombia”, tomo III del presente *Tratado*.

## V. ¿CÓMO SURGEN LOS ORÁCULOS DE LA IA?

Desde hace muchos siglos los humanos han tomado decisiones o se han apoyado en oráculos predictivos para adoptarlas<sup>(105)</sup>. Por ejemplo, Delfos, pitonisas y profetas emitían consejos, augurios, afirmaciones ambiguas, vagas o metafóricas, aunque las predicciones se comunicaban en forma de respuesta, dictamen o sentencia (*oraculum-chresmos*), sin ofrecer una interpretación o explicación de las razones o fundamentos. Ambas tareas, eran asuntos de quien preguntaba o consultaba al oráculo<sup>(106)</sup>.

El siglo XXI nos trae nuevos oráculos artificiales que, paradójicamente, muchas veces usan técnicas de caja negra que se asemejan mucho a sus colegas griegos y chinos. La frase del MIT “el futuro del pasado es el futuro del futuro” encierra varios problemas y refleja la gran paradoja que esta clase de sistemas conlleva para los siguientes principios y categorías: transparencia, acceso a la información, seguridad jurídica, voluntad, competencia, motivación y racionalidad argumentativa.

Qué datos, cómo se seleccionan, cuán representativa es la muestra, qué valores subyacen a los elegidos y cómo se avanza en el proceso de supervisión de las entradas y salidas, condicionan la razonabilidad y legitimidad de las predicciones que realizan acerca del futuro. Y aunque esta

(105) Sobre los oráculos chinos en el considerado primer libro de la historia humana, véase, Wilhelm, Richard, I Ching, *El libro de las mutaciones*, 22a ed., Sudamericana, Buenos Aires, 2013, ps. 61-62 y 454-457. Sobre oráculos griegos, véase Mumford, Lewis, *La ciudad en la historia*, ps. 99 y 199, véase en: [https://istoriamundial.files.wordpress.com/2013/11/la-ciudad-en-la-historia\\_lewis-mumford.pdf](https://istoriamundial.files.wordpress.com/2013/11/la-ciudad-en-la-historia_lewis-mumford.pdf). Jünger, Friedrich G., *Mitos griegos*, Herder Editorial, ps. 230-236; Ordóñez Burgos, Jorge A., “La adivinación en Egipto: praxis política imperial”, *Revistas Espacio, Tiempo y Forma, Serie II, Historia Antigua*, <http://dx.doi.org/10.5944/etfi.22.2009.1753> (consultado el 31/07/2017). En el caso del libro de las mutaciones, usualmente las respuestas se limitaban a una fórmula binaria: sí y no. Para el *sí* se utilizaba un trazo entero (-) y para el *no* un trazo quebrado (--). En esencia, el I Ching reemplazó al oráculo antiguo que utilizaba el caparazón de una tortuga (Wilhelm, Richard, I Ching, *El libro de las mutaciones*, p. 457).

(106) En la mitología griega, por ejemplo, podemos encontrar a Falanto. A veces, quien decidía atacar, o no, en función de la predicción del Oráculo de Delfos para conocer qué le deparaba el destino. Según narra el mito, en una oportunidad, el oráculo dictaminó que solo conquistaría una ciudad “cuando la lluvia cayera de un cielo limpio y sereno”. Una noche, su esposa despertó a su esposo llorando y mojándolo con sus lágrimas. Como ella se llamaba Etra, que significa “Cielo Sereno”, Falanto interpretó que se había cumplido la predicción del oráculo, ya que para él “había llovido desde el cielo sereno. Falanto ganó la batalla, aunque no se sabe si el oráculo influyó para que ello ocurra”. Véase: “Los orígenes de Taranto: la leyenda de Falanto”, publicado en sitio Taranto Magna, página consultada el día 08/08/2020, cita online: <http://www.tarantomagna.it/storia-taranto/origini-taranto-leggenda-falanto/>; “La fundación de Tarento, una colonia espartana”, publicado por Teo Palacios, página consultada el día 08/08/2020, cita online: <https://teopalacios.com/la-fundacion-de-tarento/>.

problemática es anterior a la IA, esta tecnología disruptiva lleva las cosas a otro nivel de complejidad. Veamos.

El lado oscuro del aprendizaje profundo basado en redes neuronales complejas<sup>(107)</sup> se configura por la existencia de un déficit estructural asociado a la propia dinámica del funcionamiento de esta clase de redes: no es posible explicar en un 100%, el paso a paso que permita interpretar o explicar en lenguaje humano, cómo sopesa o valora los atributos y la importancia que le asigna a cada dato e información para llegar a un determinado resultado<sup>(108)</sup>.

Los millones de correlaciones que se procesan en las capas ocultas de la red no pueden ser totalmente explicitadas, en el sentido de que se pueda ofrecer una explicación detallada de lo que ocurrió allí. Como el Estado debe poder justificar, motivar y explicar sus decisiones<sup>(109)</sup>, es indispensable explicar íntegramente la correlación entre los datos, su procesamiento y los resultados, en todo el ciclo de vida de la IA.

Como no es posible determinar el paso a paso de la lógica del procesamiento de datos que sucede en el interior del sistema, lo que pasa en las capas ocultas de la red<sup>(110)</sup>, en términos jurídicos impide desarrollar la motivación, fundamentación y explicabilidad en cuanto a sus resultados<sup>(111)</sup>. Por ejemplo, cuando se realiza una traducción automática en el traductor de Google, se trata de una predicción que no puede ser expli-

---

(107) Aunque también existen otros métodos de *machine learning* que podrían presentar problemas similares en términos de caja negra.

(108) Parlamento Europeo, “El impacto del Reglamento General de Protección de Datos (GDPR) en la inteligencia artificial”, 25/06/2020, disponible en: [https://www.europarl.europa.eu/stoa/en/document/EPRS\\_STU%282020%29641530](https://www.europarl.europa.eu/stoa/en/document/EPRS_STU%282020%29641530).

(109) Corvalán, Juan G., *Prometea. Inteligencia artificial para transformar organizaciones públicas*, Astrea, DPI Cuántico, Universidad del Rosario, IMODEV, 2018 disponible en: [https://dpicuantico.com/libros/prometea\\_oea.pdf](https://dpicuantico.com/libros/prometea_oea.pdf). Existe también un riesgo evidente en el hecho de que los modelos de *deep learning* simplemente realicen correlaciones y determinen resultados a través de análisis lineares que no son del todo compatibles con la estructura del derecho. Ver Amunategui, Carlos, “Sesgo e inferencia en redes neuronales ante el derecho”, 2020, ps. 18-19, disponible en: [https://www.researchgate.net/publication/342667125\\_Sesgo\\_e\\_inferencia\\_en\\_Red\\_Neuronales\\_ante\\_el\\_Derecho](https://www.researchgate.net/publication/342667125_Sesgo_e_inferencia_en_Red_Neuronales_ante_el_Derecho).

(110) “La IA moderna es, básicamente, una caja negra, que logra un desempeño superior al humano sin que las personas comprendan cabalmente cómo se obtiene ese resultado” (Comisión Económica para América Latina y el Caribe, CEPAL, “Datos, algoritmos y políticas: la redefinición del mundo digital”, LC/CMSI.6/4, Santiago de Chile, 2018, p. 171, disponible en [https://repositorio.cepal.org/bitstream/handle/11362/43477/7/S1800053\\_es.pdf](https://repositorio.cepal.org/bitstream/handle/11362/43477/7/S1800053_es.pdf)).

(111) Ver Corvalán, Juan G., *Perfiles digitales humanos*, Thomson Reuters, Buenos Aires, 2020.

cada, paso a paso, desde un punto de vista gramatical o de sintaxis. En cambio, un traductor humano puede explicar cuál método utilizó y cuáles fueron las razones por las que eligió ciertas palabras, giros o frases en vez de otras para realizar la traducción.

Aunque no sabemos exactamente qué hay detrás del método que utiliza Google Translate, lo cierto es que sus técnicas se basan en comparaciones que correlacionan por proximidad, millones de patrones de información por segundo a partir de técnicas de caja negra ¿Cuáles son las razones concretas y específicas por las cuales eligió ciertas palabras, y no otras para proponer la traducción? La respuesta que se daba acerca del modo en que aprendía Google Translate es que trabajaban de manera enfocada en el aprendizaje automático. En el año 2016, Google anunció la transición a una premisa de traducción automática neural, una práctica de “aprendizaje profundo” que permitía al servicio comparar frases enteras a la vez a partir de una gama más amplia de fuentes lingüísticas. Esto aseguró una mayor precisión al dar el contexto completo en lugar de solo cláusulas de oración aisladas. Sin embargo, no explican el motivo específico por el cual Google llega a un determinado resultado o cuáles son las fuentes por las que, a una palabra, de acuerdo con un determinado contexto, se le atribuye esa traducción<sup>(112)</sup>. Ni siquiera los programadores que diseñaron y entrenan al sistema pueden conocer millones de fuentes, por medio de las cuales Google aprende.

En síntesis, las cajas negras de la IA (como las redes neuronales) pueden ser de diferentes tipos. Las redes neuronales que se usan con mayor frecuencia son: perceptrón multicapa, redes convolucionales, redes recurrentes, redes LSTM, redes de creencia profunda, redes generativas adversariales y *capsule networks*. En todas estas técnicas subyace el problema asociado a las cajas negras: no se puede, al menos en parte, interpretar, explicar, trazar y auditar el modo en que se procesan los datos y la información para conectar lo que ingresa y lo que egresa del sistema. *Podemos hablar aquí de un problema estructural intrínseco que condiciona fatalmente la explicabilidad completa entre las correlaciones de patrones de información con los resultados a los que arriba el sistema.* Y este problema agrava aún más a otro que ha sido una de las grandes preocupaciones de la filosofía del derecho a lo largo del tiempo: cómo pueden fundamentarse las decisiones jurídicas.

(112) Norvig, Peter, “Una mirada dentro de la tecnología de Google Translate”, Google, <https://latam.googleblog.com/2011/11/una-mirada-dentro-de-la-tecnologia-de.html>; “¿Cómo funciona realmente el traductor de Google?”, Redacción PressDigital, 2019, disponible en: <https://www.pressdigital.es/texto-diario/mostrar/1116921/como-funciona-realmente-traductor-google>.

## VI. CORRELACIONES, CAUSALIDAD Y PREDICCIONES DE IA

Existen dos planos que coexisten y se retroalimentan. Por un lado, las teorías de la causalidad que sirven para establecer vínculos entre dos hechos<sup>(113)</sup>. Por otro, darle *sentido jurídico* a las correlaciones causales a través de métodos de argumentación, interpretación o ponderación. Al derecho le interesan las causas jurídicamente relevantes que causen resultados “captados” por normas jurídicas.

En una simplificación extrema, la IA puede detectar ciertas relaciones de causa y efecto, pero resulta mucho más difícil entrenarla para que pueda descubrir (no entender ni razonar en términos jurídicos humanos) y predecir<sup>(114)</sup> las inferencias causales con relevancia jurídica para una amplísima y vasta gama de posibilidades que se dan en cada área del derecho<sup>(115)</sup>.

---

(113) El análisis de la causalidad puede descomponerse en dos. Por un lado, la causalidad general, según la cual la producción de un evento de cierto tipo hace más probable la producción de otro porque existe una ley de la naturaleza que da cuenta de esa probabilidad en un grado relevante. El segundo es la causalidad individual, la que permite afirmar que, en una específica y concreta situación, un hecho particular ha sido causado por la ocurrencia de otro hecho específico y determinado (Taruffo, Michele, *La prueba de los hechos*, Trotta, Madrid, 2008, p. 256).

Entre ambos conceptos media una relación lógica. Lo que implica que, si se quiere probar el nexo causal entre dos hechos individuales, debe probarse primero la existencia de una ley general para después subsumir la causalidad individual en la generalización (Ferrer Beltrán, *Prueba y verdad en el derecho*, Marcial Pons, Barcelona, 2014, p. 223). De este modo, para que pueda sostenerse que existe una relación de causalidad individual entre dos eventos, debe existir una ley de cobertura (Ferrer Beltrán, *Prueba y verdad en el derecho*, Marcial Pons, 2008, p. 257).

(114) La forma en la que usamos el término de “predicción” en este artículo es idéntica al que usa la Comisión Europea cuando habla de “previsión”. Para este organismo, la predicción es el acto de anunciar lo que sucederá (antes de dictar) antes de los eventos futuros (por inspiración sobrenatural, por clarividencia o premonición). La previsión, por el contrario, es el resultado de observar (apuntar, ver) un conjunto de datos para prever una situación futura. Según la Comisión Europea, este abuso del lenguaje y su difusión parece explicarse por una transferencia del término de las ciencias “duras”, donde se refiere a una variedad de técnicas de ciencia de datos derivadas de las matemáticas, estadística y teoría de juegos que analizan hechos presentes y pasados para formular hipótesis sobre el contenido de eventos futuros. Ver: European Commission for the Efficiency of Justice (CEPEJ), “European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment”, Council of Europe, 3-4 December, 2018.

(115) Nota del director: sobre la limitación para realizar inferencias causales de la inteligencia artificial en el ámbito del derecho ver: Corvalán, Juan G. - Carro, María Victoria, “Más allá del juego de la imitación: las principales limitaciones de la inteligencia artificial”, tomo I del presente *Tratado*.

Por ejemplo, supongamos que diseñamos y entrenamos el modelo predictivo con 400 sentencias emitidas por diez juezas y jueces para que una IA como Prometea o PretorIA correlacione datos y patrones de información en sentencias. La máquina podría aprender a detectar con una alta tasa de acierto, elementos como la inexistencia de dolo o la interrupción del nexo causal, según la teoría que se aplique para resolver casos.

Sin embargo, en este proceso no ejecuta técnicas de argumentación vinculadas a teorías jurídicas. La IA, no sabe nada de derecho ni de ética ni de campo alguno del conocimiento!, trabaja subsimbólicamente y es capaz de dar respuestas asertivas si el modelo funciona adecuadamente. Reconoce patrones y correlaciones de palabras, frases o símbolos, para luego agruparlos en función de diversos criterios o según un índice de pesos o reglas de inferencia, que no son jurídicas, aunque luego pueden ser útiles para realizar argumentaciones racionales, interpretadas a la luz del pensamiento humano.

Sin embargo, aunque la IA no sepa derecho, la tarea de correlacionar patrones de información históricos jurisprudenciales puede ser de gran ayuda<sup>(116)</sup>. Por ejemplo: *i*) es útil para ajustar o mejorar los análisis acerca de las causas relevantes jurídicamente que causen resultados “captados” por normas jurídicas y aplicadas por humanos. Esto, lo hemos podido comprobar a partir de desarrollar y aplicar un predictivo que puede detectar o no la interrupción del nexo causal, en el ámbito de los casos judiciales vinculados a los accidentes de tránsito; *ii*) también contribuye a corregir, prevenir o mitigar sesgos o patrones de discriminación en los datos<sup>(117)</sup>.

---

(116) El año pasado, Francia emitió una norma muy discutible para nuestro sistema jurídico. Aunque no prohíbe usar técnicas de IA sobre fallos judiciales, sí establece la prohibición y criminalización para usar técnicas (no aclara cuáles) para evaluar, analizar, comparar o predecir las prácticas de un juez, basándose en comparar su identidad y de qué tribunal es miembro. La prohibición francesa, en nuestro sistema, sería manifiestamente inconstitucional y contraria a tratados internacionales (principio de publicidad, transparencia, entre otros). Su aplicación literal, llevaría al absurdo de que se pueda criminalizar con una pena de hasta 5 años, a quien “evalúe”, “analice” y “compare” (usando palabras textuales de la ley) en un Excel los fallos judiciales vinculados a la identidad del juez/a y respecto de qué tribunal integra. Para acceder a la ley consulta a <https://bit.ly/2MY0fx8>. Ver especialmente el artículo 33 que reforma el artículo L. 153-1 y el L. 10.

(117) En el ámbito de la medicina, las predicciones de IA se basan en anticipar un resultado a partir de lo que se conoce “una intervención”. Por ejemplo, un tratamiento asignado por un médico que cambiará la condición cardíaca de un paciente es una intervención. Predecir el cambio en la condición del paciente es una tarea de inferencia causal. En general, una intervención es una acción realizada por un agente externo que cambia los valores originales o las distribuciones de probabilidad, de algunas de las variables en el sistema. Además de predecir los resultados de las accio-

En síntesis, en las predicciones de IA uno o varios algoritmos revelan una enorme cantidad de datos a fin de establecer patrones que se traducen en predicciones, sobre la base de algún criterio *a priori*, propio del modelo de IA. Se trata de un entrenamiento con base en patrones identificados en casos previos. En el ámbito del lenguaje natural, a través de la identificación de patrones y palabras o conjuntos de palabras o frases clave (*keywords*) sobre determinados documentos o información, en la predicción lee el documento que se está trabajando en la actualidad y lo compara contra documentos elaborados en el pasado. Cuando logra correlacionar (“machear”) los patrones de información bajo los términos con los que fue entrenada una respuesta o solución, luego es posible automatizarlas.

Por ejemplo, intenta detectar hipótesis fácticas similares, para indicar que correspondería la solución específica que se adoptó cuando aquellas se verificaron en el pasado. También podría predecir la existencia de ciertos criterios jurídicos que están presentes en el caso que se examina (siempre en relación con ejemplos resueltos en el pasado). En ambos casos, si se realizan con técnicas de caja blanca, es posible trazar, explicar y transparentar los datos y la forma de procesarlos para arribar a los resultados tal como lo hacemos desde el IALAB de la UBA. Además, las personas que lo diseñan y entrenan, nunca pierden el control, su posibilidad de intervenir, de auditar, de interpretar. En esta clase de IA, no existe el déficit estructural asociado a la falta de explicabilidad intrínseca que se presenta en las redes neuronales y en otros sistemas opacos.

Por último, cuando se trata de predicciones de caja blanca, estas contribuyen a garantizar el principio de no discriminación algorítmica. Por un lado, nos ayudan a realizar un ejercicio retrospectivo acerca de posibles correlaciones entre decisiones. Por otro, es posible corregir, mitigar o eliminar esquemas decisionales o bien, “curar” o “limpiar” en los datos de entrenamiento, los sesgos negativos o las distinciones basadas en motivos de raza, color, sexo, idioma, religión, opinión política o de otra índole, origen nacional o social, posición económica, nacimiento o cualquier otra condición social<sup>(118)</sup>.

---

nes, la inferencia causal también tiene que ver con la explicación: identificar cuáles fueron las causas de un evento particular que sucedió en el pasado. Véase Silva, R., *Casuality*, Centre for Computational Statistics and Machine Learning, University College London, 2014, p. 1, (en línea), <http://www.homepages.ucl.ac.uk/~ucgtrbd/papers/causality.pdf>.

(118) Corvalán, Juan G., “Inteligencia artificial y derechos humanos”, *Parte II, DPI Cuántico*, 10/07/2020, disponible en: <https://dpicuantico.com/sitio/wp-content/uploads/2017/07/Juan-Gustavo-Corvalan-Constitucional-10.07.2017.pdf>.

# FUNDAMENTOS ANTROPOLÓGICOS, ÉTICOS, FILOSÓFICOS, HISTÓRICOS, SOCIOLÓGICOS Y JURÍDICOS PARA LA CONSTITUCIÓN UNIVERSAL DE UN DERECHO DE LA INTELIGENCIA ARTIFICIAL (DIA)

MARCOS LÓPEZ ONETO

## I. INTRODUCCIÓN

La posibilidad de que la inteligencia artificial (IA) llegue a superar la capacidad intelectual humana dejó de ser un tema de ciencia ficción<sup>(1)</sup>. La IA es una tecnología que se está desarrollando rápidamente, por momentos, incluso a ritmos vertiginosamente exponenciales<sup>(2)</sup>. Y así como

(1) Iremos viendo a lo largo de este trabajo que, a nivel gubernamental, existe una variada gama de documentos oficiales que advierten sobre el potencial disruptivo de la IA, mencionando expresamente que el tema ya no admite ser comprendido como uno de ciencia ficción. Entre ellos puede mencionarse la resolución del Parlamento Europeo (P8\_TA [2017] 0051), de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre Normas de Derecho Civil sobre Robótica (2015/2103 [INL]), que admite expresamente esa posibilidad y describe los peligros que entraña, proponiendo regulaciones para evitarlos. Véase Parlamento Europeo (2017). Pero no puede perderse de vista que para algunos investigadores la IA en muchos ámbitos ya es sobrehumana (tratándose de los sistemas expertos, por ejemplo), lo cual, en todo caso, no justificaría una proyección distópica de sus capacidades pues aun cuando esas capacidades “han sido disruptivas para las vidas humanas en facetas como el mercado laboral, en ningún caso han dotado a las máquinas de ambición” (Bryson, Joanna, “La última década y el futuro del impacto de la IA en la sociedad”, en *¿Hacia una nueva ilustración? Una década trascendente*, BBVA, Madrid, BBVA/OpenMind, p. 133).

(2) Una forma de apreciar la velocidad del crecimiento y desarrollo de la IA es mediante la observación de la denominada Ley de Moore. Dice la citada ley que cada dos años se duplica el número de transistores que pueden ser instalados en un microprocesador de tamaño fijo. El resultado de este fenómeno es que los electrones tienen que viajar cada vez menos distancia y por eso los circuitos también funcionan más rápido, lo cual, ciertamente, “da un empuje adicional a la capacidad de computación en general [...] El resultado de todo esto es un crecimiento exponencial en el rendimiento por unidad de coste de la computación” (Kurzweil, Raymond, *La singularidad está cerca. Cuando los humanos trascendamos la biología*, Lola Books GbR, Berlín, 2012, e-book, loc. 1436). Gordon Moore, formulador de la ley que lleva su nombre, quien fuera además cofundador y *chairman emeritus* de Intel, la dio a conocer en un *paper* de circulación interna que escribió en 1965 para *Fairchild Semiconductor In-*

promete modificar la forma en que vivimos, trabajamos, aprendemos, descubrimos y nos comunicamos, incorporando extraordinarios beneficios a la vida humana, así también, y como ha sido advertido por connotadas voces científicas, tendría el potencial de contribuir al fin de la humanidad, mediante la inducción de un cambio evolutivo radical en la línea del *homo sapiens* que lo conduciría a su reemplazo por una nueva especie; un híbrido ciberbiológico; incluso hasta un ser meramente virtual, que de humano, como ha dicho un conocido transhumanista, solo tendría la tendencia inherente a expandir “su alcance físico y mental más allá de sus limitaciones”<sup>(3)</sup>.

Y es que, si la historia de la tecnología nos muestra que la revolución industrial del siglo XIX fue una revolución en la mecánica que sirvió para automatizar el trabajo físico, localizando principalmente en ese ámbito su efectos, la actual del siglo XXI, conocida como cuarta revolución industrial (4RI), está dando muestras de ser una revolución de carácter existen-

*ternational, Inc.* El artículo se tituló “The Future of Integrated Electronics”. Posteriormente, el escrito fue publicado en la revista *Electronics* bajo el título “Cramming more components onto integrated circuits”. Originalmente Moore advirtió una duplicación anual (Moore, Gordon E., “Cramming more components onto integred circuits”, *Electronics*, 1965, 38 [8]:114-117). Diez años después revisó su afirmación ampliando el enunciado de duplicación del número de transistores a dos años (Moore, Gordon, “Progress in Digital Integrated Electronics”, *IEEE Solid-State Circuits Society Newsletter*, 2006, 11[3]:36-37. DOI: 10.1109/N-SSC.2006.4804410). La tendencia descrita por Moore se ha ido cumpliendo hasta ahora. Recordemos que en 1971 un *chip* podía contener hasta 2.300 transistores y hoy día ya existen *chips* del tamaño de una uña humana que contienen 30 billones de transistores de 5 nanómetros cada uno. IBM dice que con dichos *chips* se logra un aumento del rendimiento del 40%, o un 75% de eficiencia energética con el mismo rendimiento, en comparación con los *chips* de 10 nanómetros de la generación anterior (véase, Bu, Huiming, “5 nanometer transistors inching their way into chips”, *IBM Think Blog*, 2017, disponible en <https://www.ibm.com/blogs/think/2017/06/5-nanometer-transistors/> [consultado el 30/04/2021]). Sin embargo, hay límites que vienen impuestos por el soporte físico de los *chips* actuales, el silicio, que toleraría eficientemente solo hasta cierta cantidad de transistores de escala nanométrica. Se conjetura que llegado ese límite físico (o también de eficiencia y rentabilidad económica), la tendencia de desarrollo exponencial se podrá sostener gracias al reemplazo del silicio por nanotubos de carbono, computación molecular, autoensamblado en circuitos de nanotubos, sistemas biológicos que emulan circuitos de ensamblado, computación con ADN, *spintronics* (computación con el *spin* de los electrones), computación con la luz y computación cuántica; todo lo cual crearía el puente hacia la computación molecular en 3-D que, como observa Kurzweil, sería el sexto paradigma de la computación: el paradigma capaz de producir la fusión entre inteligencia artificial y biológica (Kurzweil, Raymond, *La singularidad está cerca. Cuando los humanos trascendamos la biología*, Lola Books GbR, Berlín, 2012, e-book, loc. 2260).

(3) Kurzweil, Raymond, *La singularidad está cerca. Cuando los humanos trascendamos la biología*, Lola Books GbR, Berlín, 2012, e-book, loc. 352.

cial, porque impacta (e impactará cada vez más) la constitución integral del ser humano (ontológica, epistemológica, estética, ética y poética).

Así las cosas, no está de más apuntar que si tuviéramos que elegir cuál visión distópica es la que más se asemejaría a la derivada de la IA, si la de Huxley o la de Orwell; la de Huxley parecería más cercana, por cuanto está basada en el deseo, a diferencia de la orweliana que se sostiene más en el miedo. Huxley comprendió que en el futuro íbamos a ser controlados no través de la fuerza o el miedo, sino que mediante la manipulación de nuestros deseos. Y en efecto, la IA seduce y erotiza al ser humano, por cuanto sintoniza fuertemente con el deseo, a nivel de inconsciente individual y colectivo, de superación de los límites materiales que la condición humana supone, mediante la promesa del advenimiento de los paraísos artificiales y la inmortalidad que la hibridación permitiría.

El tema es preocupante. La potencialidad de un final apocalíptico para la humanidad se resume en la categoría riesgo existencial acuñada por el filósofo oxoniano de la inteligencia artificial, Nick Bostrom, quien escribe: “[...] un riesgo existencial es el que amenaza con causar la extinción de la vida inteligente de origen terrestre o con destruir de forma permanente y drástica sus posibilidades de desarrollarse en el futuro”<sup>(4)</sup>. Así, la caída de un meteorito sobre la tierra es un riesgo existencial natural. La vida sobre la Tierra ha sobrevivido a lo largo de su trayectoria a una variada gama de riesgos existenciales naturales, como impactos de asteroides, pandemias, erupciones supervolcánicas, terremotos, explosiones de rayos gamma. Sin embargo, las distribuciones empíricas de impacto y los modelos científicos sugieren que la probabilidad de extinción debido a este tipo de riesgo es extremadamente pequeña en una escala de tiempo de aproximadamente un siglo<sup>(5)</sup>. Por el contrario, los análisis de los especialistas que han examinado la fenomenología de los riesgos existenciales, sugieren que, en una escala de tiempo mucho más acotada que la de la naturaleza, existen más posibilidades de ocurrencia de riesgos existenciales producidos por la humanidad (piénsese solamente en los riesgos existenciales derivados de una conflagración nuclear mundial).

(4) Bostrom, Nick, *Superinteligencia. Caminos, peligros, estrategias*, Teel Editorial, Madrid, 2016, p. 115; “Existential Risk Prevention as Global Priority”, *Global Policy*, vol. 4, p. 15.

(5) Bostrom, Nick, “Existential Risk Prevention as Global Priority”, *Global Policy*, vol. 4, ps. 15 y ss.

Nota del director: además, Bostrom considera que la dignidad humana y posthumana son compatibles y complementarias. Sobre esta cuestión ver Gil Domínguez, Andrés, “Transhumanismo y derechos ¿humanos?”, tomo I del presente *Tratado*.

Se trata de los denominados riesgos existenciales antropogénicos<sup>(6)</sup>. La mayoría de los riesgos existenciales antropogénicos pueden ser relacionados con los avances tecnológicos futuros que le otorgarían a la humanidad el poder de manipular radicalmente la naturaleza, incluida la propia biología del ser humano, todo esto en el contexto geológico más general del denominado Antropoceno<sup>(7)</sup>.

Pues bien: al día de hoy, estos riesgos existenciales antropogénicos derivan principalmente del desarrollo sin controles éticos y jurídicos de las denominadas ciencias y tecnologías NBIC, a saber: (N) nanotecnología y las nanociencias, (B) biotecnología, biomedicina, incluyendo ingeniería genética, (I), tecnologías de la información (TIC), incluye computación, inteligencia artificial y comunicaciones, y (C), ciencias cognitivas, que incluye, por cierto, a las neurociencias. La característica de las ciencias y tecnologías NBIC es que convergen sinérgicamente y a escala manométrica<sup>(8)</sup>.

Pero, aun cuando la circunstancia que la mayor parte de los riesgos existenciales antropogénicos del presente siglo, derivados de las ciencias NBIC, pudieran ser meramente especulativos, es decir, no susceptibles de precisión probabilística a través de algún método estadístico o científico riguroso, el imperativo moral de preservar la vida futura de la humanidad impone el deber de actuar con precaución y cautela si no se tiene la cer-

(6) Bostrom, Nick, "Existential Risk Prevention as Global Priority", *Global Policy*, vol. 4, p. 16.

(7) El Antropoceno (de griego *ἄνθρωπος* *anthropos*, "ser humano" y *καινός* *kainos*, "nuevo") es una época geológica propuesta por una parte de la comunidad científica para suceder o reemplazar al denominado Holoceno, la época actual del período cuaternario en la historia terrestre, debido al significativo impacto global que las actividades humanas han tenido sobre los ecosistemas terrestres (especialmente ilustradas por la denominada "extinción masiva del Holoceno"). No hay un acuerdo común respecto a la fecha precisa de su comienzo; algunos lo consideran junto con el inicio de la Revolución Industria (a finales del siglo XVIII), mientras que otros investigadores remontan su inicio al comienzo de la agricultura, solapando enteramente al Holoceno. El término Antropoceno fue usado en el año 2000 por el ganador del premio Nobel de Química Paul Crutzen, quien considera que la influencia del comportamiento humano sobre la Tierra en las recientes centurias ha sido significativa, y ha constituido una nueva era geológica. La propuesta del uso de este término como concepto geológico oficial ha ganado fuerza desde el 2008 con la publicación de nuevos artículos que apoyan esta tesis. Sin embargo, para que se convierta en oficial se requiere la aprobación de la Comisión Internacional de Estratigrafía. La bibliografía en la materia es extensa, entre otros, puede consultarse: Trischler, Helmuth, "El Antropoceno, ¿un concepto geológico o cultural, o ambos?", *Desacatos* 54, 2017, ps. 40-57.

(8) Volveremos sobre este importante asunto en la sección I.

Nota del director: sobre las tecnologías NBIC el mismo enfoque se presenta en Gil Domínguez, Andrés, "Transhumanismo y derechos ¿humanos?", tomo I del presente *Tratado*.

teza que la acción u omisión no producirá efectos dañinos en contra de la humanidad.

Aquí entra al juego de la política jurídica el principio de responsabilidad bajo la formulación del filósofo alemán de origen judío Hans Jonas, quien argumentó que la ciencia y la tecnología deben orientar su desarrollo al bien de la humanidad y nunca apostar, en las apuestas de la acción, la existencia o la esencia del hombre en su totalidad. Porque:

“[...] dada sobre todo la irreversibilidad de los procesos puestos en marcha, hasta que no podamos hacer proyecciones seguras no habrá mayor valentía que la cautela, que en todo caso será un mandamiento de la responsabilidad [...] La inseguridad puede ser nuestro destino permanente, lo cual tiene consecuencias morales”<sup>(9)</sup>.

Sin embargo, y esto es lo alarmante, los riesgos existenciales derivados del desarrollo de la IA, como ya he sugerido, son más que especulativos y al día hoy existe consenso a nivel mundial que su desarrollo sin controles éticos y jurídicos pueden atentar en contra la existencia de la humanidad. En parte este consenso se funda en la circunstancia que la IA ya está asediando el derecho a la privacidad, a la libertad y a la igualdad. Y en perspectiva, ya se proyecta científicamente que podría llegar a afectar la vida humana en su totalidad.

En ese orden de razonamiento, entonces, Bostrom se pregunta qué sucedería si alguna vez se llegase a construir una máquina superinteligente, ¿sobreviviríamos a eso?<sup>(10)</sup>. Si la superinteligencia cibernética (un sistema de algoritmos evolutivos cuánticos, por ejemplo) decidiera que ya no le interesa convivir con los humanos, ¿podría la humanidad, que sería menos inteligente que la máquina superinteligente, ganarle; podría asestar un primer golpe decisivo que la llevase a la victoria? ¿Es un proyecto responsable éticamente para con la humanidad, la construcción de una superinteligencia artificial?

Por supuesto que la pregunta de Bostrom no constituye una excentricidad filosófica. El potencial distópico que entraña la IA ha sido puesto de relieve por científicos de la trascendencia de un Stephen Hawking, quien, meses antes de fallecer, advirtió a través de una teleconferencia en

(9) Jonas, Hans, *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*, Herder, Madrid, 1995, ps. 309-310.

(10) Bostrom, Nick, *Superinteligencia. Caminos, peligros, estrategias*, Teel Editorial, Madrid, 2016, p. 115; “Existential Risk Prevention as Global Priority”, *Global Policy*, vol. 4, ps. 115-126.

Nota del director: sobre superinteligencia o inteligencia artificial fuerte ver Corvalán, Juan G. - Díaz Dávila, Laura - Simari, Gerardo, “Inteligencia artificial: bases conceptuales para una aproximación interdisciplinar”, tomo I del presente *Tratado*.

el inicio de la *Web Summit* en Lisboa que “el desarrollo de la IA podría ser lo peor o lo mejor que le ha pasado a la humanidad”. Y destacó: “Tenemos que llevar el conocimiento más allá de la discusión teórica respecto a cómo debería ser la IA y tomar decisiones para asegurarnos que planeemos como puede ser”<sup>(11)</sup>. Anteriormente, en una entrevista concedida a la BBC de Londres había señalado —derechamente— que el desarrollo de la IA podría significar el fin de la especie humana<sup>(12)</sup>.

Y por cierto que Hawking no es el único científico que ha advertido sobre los catastróficos peligros que, para el futuro de la humanidad, implica el desarrollo sin controles éticos de la IA. La lista es larga<sup>(13)</sup>. El mismo Hawking, junto al físico del *Massachusetts Institute of Technology* (MIT) Max Tegmark, el ganador del premio de física Nobel Frank Wilczek y el reconocido teórico y profesor de IA de Berkeley Stuart Russell, publicaron un inquietante artículo de opinión en *The Independent*, en el cual decían:

“Solo podemos imaginar una [la IA] más inteligente que los mercados financieros, que supera en creaciones a los investigadores humanos, manipula a los líderes humanos y desarrolla armas que ni siquiera podemos entender. Aunque el impacto a corto plazo de la IA depende de quién la controle, el impacto a largo plazo depende de si puede controlarse en algún grado”<sup>(14)</sup>.

Por otra parte, también importantes visionarios y emprendedores del mundo de la IA como Bill Gates (Microsoft) y Elon Musk (Tesla & Space X) han alzado sus voces previniendo a la humanidad sobre el potencial destructivo de la IA<sup>(15)</sup>. De hecho, Microsoft es parte del programa de las

---

(11) Reporte de prensa en *Infobae* disponible en <https://www.infobae.com/america/mundo/2017/11/07/stephen-hawking-alerta-sobre-la-inteligencia-artificial-puede-ser-lo-mejor-o-lo-peor-que-le-suceda-a-la-raza-humana/> (01/03/2021). También se encuentra disponible la videoconferencia en YouTube: <https://www.youtube.com/watch?v=E3efC-lpBo0> (consultado el 18/04/2021). Como quedó testimoniado en su libro póstumo *Breves respuestas a las grandes preguntas* (Hawking, 2018), Hawking nos alertó hasta el último de sus días sobre los peligros de la IA.

(12) Reporte de prensa en BBC disponible en <https://www.bbc.com/news/technology-30290540> (consultado el 01/03/2021).

(13) Una lista de científicos de primera línea que llaman a priorizar los beneficios de la IA y prevenir sus dificultades, puede ser consultada en “An Open Letter. Research Priorities For Robust and Beneficial Artificial Intelligence”, publicada y gestionada por Future of Life Institute, disponible en <https://futureoflife.org/ai-open-letter/?cn-reloaded=1> (consultado el 08/03/2021). Hago presente desde ya que todas las traducciones de obras escritas en inglés citadas en este artículo son de mi autoría.

(14) *The Independent*, 01/05/2014, disponible en [https://www.huffpost.com/entry/artificial-intelligence\\_b\\_5174265](https://www.huffpost.com/entry/artificial-intelligence_b_5174265) (consultado el 04/03/2021).

(15) Una muestra de esas opiniones puede ser revisada en el documental de Paine, Chris, “Do you trust this computer”, disponible en YouTube: [https://www.youtube.com/watch?v=ZFRJgS\\_sBCY](https://www.youtube.com/watch?v=ZFRJgS_sBCY) (consultado el 03/03/2021).

Naciones Unidas *AI for Good* cuyo objetivo es orientar el desarrollo de la IA en función de la solución de los problemas globales de la humanidad tales como pobreza, cambio climático, salud, educación, entre otros<sup>(16)</sup>.

Y es que el poder creativo y destructivo de la IA, potencialmente, es paragonable con la energía nuclear, en el sentido que ambas tecnologías, bajo ciertas variables, están en condiciones de provocarle un daño devastador a la especie humana. En el caso de la energía nuclear, por la incineración y destrucción material de la vida; y en el caso de la IA, por la paulatina desaparición material y cultural de la humanidad, que, por haber caído en la fascinación fáustica de la inmortalidad tecnológica, terminará auto extinguiéndose, dejando el mundo a su ciberdescendencia.

Escribo estas líneas, por tanto, bajo la premisa que lo que está en juego a futuro con el desarrollo de la IA sin regulaciones éticas y jurídicas *pro humanitas* es la supervivencia de la especie humana. El desafío para la sociedad mundial<sup>(17)</sup> es y será crear regulaciones que orienten el desarrollo científico y tecnológico hacia la sustentabilidad de la vida como fenómeno general en el planeta Tierra y de la humanidad, en particular. Llevada esa premisa al terreno concreto de la IA, desde el punto de vista jurídico se trata, esencialmente, de un asunto de derechos humanos, es decir, de guiar u orientar el desarrollo de la IA en función del respeto, vigencia y promoción de esa clase derechos, sin los cuales, lisa y llanamente, el futuro de la humanidad no puede asegurarse.

Sistemas de IA que amenazaren en su esencia los derechos humanos, no debieran permitirse, más aún, obviamente, los que derechamente los vulneraren. Y si es muy grave que estos sistemas de IA amenacen o dañen derechos humanos tales como la libertad, la igualdad o la privacidad, más

(16) El programa de *AI for Good* de Microsoft puede ser consultado en <https://www.microsoft.com/en-us/ai/ai-for-good> (consultado el 01/03/2021). El programa *AI for Good*, liderado por la Unión Internacional de Telecomunicaciones (UIT/en inglés: ITU), pone el acento en el desarrollo de la IA para un mundo sustentable. Puede ser consultado en <https://aiforgood.itu.int/> (consultado el 01/03/2021). También se recomienda consultar, "IA para el bien: Sendas para avanzar. El progreso mediante la innovación", *ITU News Magazine*, nro. 3, 2019, disponible en [https://www.itu.int/en/itunews/Documents/2019/2019-03/2019\\_ITUNews03-es.pdf](https://www.itu.int/en/itunews/Documents/2019/2019-03/2019_ITUNews03-es.pdf) (consultado el 07/08/2019).

(17) Me parece que la teoría de la sociedad mundial como "el" sistema social operativamente clausurado en torno a comunicaciones, que incluye a todos los demás sistemas sociales, pues realiza las reducciones de complejidad más generales que permiten sus respectivas constituciones en el mundo moderno, así como ha sido desarrolla por Luhman, ofrece una (auto)descripción plausible de la civilización actual, estructurada y organizada en base a las tecnologías de la información. Véase: Luhman, Niklas, *La sociedad de la sociedad*, trad. Javier Torres Nafarrate, Herder, México, 2007.

grave aún lo serán aquellos que potencialmente pudieran inducir, paulatina o bruscamente, una transformación radical del ser humano, en términos de resultar imposible identificar como humano, al ente (producto) de ese proceso de trasmutación ontológica.

El tema causa vértigo y quizás para algunos todavía podría parecer de ciencia ficción. Sin embargo, esa no es la opinión de los principales líderes civilizaciones del siglo XXI. Tanto en los Estados Unidos de Norteamérica (EE.UU.) como en la Unión Europea (UE), el tema está siendo abordado gubernamentalmente y ya está en la agenda política internacional, lo cual constituye un primer paso. El segundo paso, y que es precisamente el que estamos dando mediante la presente contribución científica jurídica, es el de la creación de un Derecho de la Inteligencia Artificial (DIA).

En efecto, frente a los desafíos que plantea el desarrollo de la IA se han observado en general, respuestas regulativas puntuales de los Estados nacionales de la sociedad mundial, pero no, la configuración expresa de un DIA, cuyo objetivo sea, precisamente, orientar el desarrollo de la IA en función de objetivos de sustentabilidad y continuidad de la humanidad; de protegerla de una eventual orientación transhumanista, es decir, de una orientación regulativa inductora hacia la posthumanidad.

¿Cómo podrían los seres humanos de esta generación proteger a la humanidad de la puesta en marcha de proyectos de IA que tengan la potencialidad de contribuir a la superación del hombre, por ejemplo, mediante su paulatina ciborgización?<sup>(18)</sup> ¿Cómo hacerlo si no existe un DIA expresamente formulado?

La hipótesis que se plantea en este artículo sugiere que los riesgos existenciales que puede generar el desarrollo de la IA, de seguir un curso de desarrollo sin controles éticos y jurídicos, y en función, además, de objetivos transhumanistas, pueden ser enfrentados recurriendo a la doctrina universalmente aceptada de los derechos humanos y del derecho internacional de los derechos humanos (DIDH).

La creación de una superinteligencia artificial en cualquier jurisdicción o lugar del planeta Tierra, amenazará a la humanidad entera, pues tendrá el poder de expandirse y evolucionar autónomamente, sin tener un adversario humano de su mismo nivel, salvo el de otras superinteligencias artificiales con las cuales pudiera entrar en una espiral de competencia.

---

(18) Nota del director: sobre los *cyborgs* que ya existen actualmente ver Gil Domínguez, Andrés, "Transhumanismo y derechos ¿humanos?", punto 5, tomo I del presente *Tratado*.

La prevención y protección de la humanidad de los riesgos existenciales provocados por el desarrollo de la IA, entonces, debe ser universal, pues si solo unos pocos Estados regulan y otros no, de manera casi inevitable se generarán incentivos para que la industria de la IA (sin regulaciones éticas y jurídicas) se mueva —igual que el capital financiero— hacia esos paraísos desregulados (o de regulación permisiva). Por consiguiente, la regulación debe tender hacia a la internacionalización, lo mismo que acontece con la regulación de la energía atómica y los intentos por preservar el medio ambiente.

Por otra parte, frente a la ausencia de regulaciones legales específicas sobre desarrollo de la IA a nivel nacional, como acontece actualmente en la mayoría de los Estados miembros de las Naciones Unidas, el DIDH es, en última instancia, el instrumento disponible para hacer frente a los riesgos existenciales que puedan provenir del desarrollo de la IA. Vale decir, aun cuando no existan regulaciones estatales específicas en la materia, frente a los desarrollos de IA claramente atentatorios contra los derechos humanos (vida, libertad, igualdad, no discriminación arbitraria, privacidad, debido proceso, etc.), las personas tienen derecho a recurrir a los sistemas de protección internacional de los derechos humanos, si acaso sus Estados no los amparan en sus reclamos.

Considerando lo anteriormente expuesto, hago presente que en este artículo se indagará básicamente en el sistema universal de protección de los derechos humanos, pero evidentemente no se descarta la factibilidad de recurrir a los sistemas europeo, interamericano y africano de protección. De hecho, normativamente esa posibilidad existe, ya sea por los tratados que los configuran, así como por su conexión sustantiva con el sistema universal. Sin embargo, no se explorará esa alternativa por exceder los límites de este trabajo.

En síntesis, el objetivo jurídico es este artículo es contribuir a sentar las bases del DIA<sup>(19)</sup>. Para intentar lograrlo será necesario indagar en las fuentes materiales del proceso histórico y cultural que hace posible el surgimiento de esta nueva rama del derecho de proyecciones universales, de modo de así permitir su visualización y configuración en la audiencia.

En concreto el trazado argumentativo de esta contribución es el siguiente:

(a) La investigación que presento supone como punto de partida (o premisa) que el desarrollo de la IA sin controles o regulaciones éticas y

---

(19) He desarrollado latamente esta fundamentación en: López Oneto, Marcos, *Fundamentos para un derecho de la inteligencia artificial: ¿queremos seguir siendo humanos?*, Valencia, Tirant lo Blanch, 2019.

jurídicas es un peligro para la continuidad de la especie humana y, por consiguiente, para la humanidad. La IA tiene el potencial de afectar gravemente los derechos humanos, pues incluso puede afectar la continuidad de la especie humana. Este peligro aumenta cuando el desarrollo de la IA es inducido por la ideología transhumanista cuyo objetivo estratégico es conducir a la humanidad, mediante el uso de la tecnología, hacia la posthumanidad. El transhumanismo es una ideología que se encuentra a la vanguardia de las fuerzas que están guiando el desarrollo espiritual y tecnológico del siglo XXI hacia la utopía posthumana, para la cual el humanismo y sus principales conquistas son solo una etapa de la evolución. Así las cosas, la discusión crítica con el transhumanismo, a efectos de mostrar los riesgos existenciales que potencialmente implica, es un capítulo imprescindible de la fundamentación del DIA.

(b) La hipótesis que se discute en este artículo dice, asimismo, que, si bien expresamente no se ha formulado el DIA, sí existen implícitamente sus principios (o leyes más generales), los cuales son posibles de inferir de la doctrina universal de los derechos humanos y del DIDH en tanto principios implícitos de aquel, pues es dicho ordenamiento jurídico internacional el que tiene por misión proteger universalmente y en última instancia, al hombre y a la humanidad de los atentados a su existencia.

Desarrollaré la argumentación en tres secciones muy acotadas en su extensión.

En la primera (II) se pondrá de manifiesto las fuentes históricas y materiales del DIA mediante una breve descripción de la cuarta revolución industrial y la sociedad de la información, focalizando la presentación en IA. En esta sección se pondrán de relieve argumentos antropológicos, culturales, históricos y sociológicos sobre la evolución disruptiva de la IA.

En la sección segunda (III), se presentará la fundamentación ética del DIA en función del principio de responsabilidad conformado en sus lineamientos generales por Hans Jonas. La fundamentación ética del DIA, desde el principio de responsabilidad, no se efectuará *in abstracto*, sino que, discutiendo con el transhumanismo, la corriente de pensamiento identificada en esta investigación como la gran impulsora filosófica del fin del hombre y de la humanidad en aras de un futuro paraíso posthumano. En esta sección se indagará la evolución disruptiva de la IA, básicamente, a partir de las reflexiones éticas y filosóficas más relevantes que han surgido.

En la sección tercera (IV) se analizarán los desafíos normativos jurídicos que plantea la inteligencia artificial, proponiendo la configuración del DIA a partir de un proceso de inducción jurídica desde el DIDH. El proceso de configuración de los fundamentos del DIA, cier-

tamente, impondrá el deber metodológico de diferenciarlo o delimitarlo respecto de otras áreas jurídicas relacionadas con él, como son la Informática Jurídica (IF), el Derecho Informático (DIF) y el campo de la *Artificial Intelligence and Law*.

En la sección final (V), se presentarán algunas conclusiones.

## II. LA CUARTA REVOLUCIÓN INDUSTRIAL Y LA ERA DE LA INFORMACIÓN COMO FUENTES MATERIALES DEL DERECHO DE LA INTELIGENCIA ARTIFICIAL (DIA)

La fundamentación del Derecho de la Inteligencia Artificial (DIA) requiere como paso previo que se describa y analice el fenómeno que está orientando el cambio en la base material y espiritual de la sociedad mundial, a saber: la denominada cuarta revolución industrial (4RI) que opera como su fuente material<sup>(20)</sup>. Se trata de un cambio profundo, que genera expectativas de diversa índole en los sujetos y en la sociedad, complejizando el entorno ético-jurídico, por la proyección del impacto que, para la historia de la humanidad y de su curso evolutivo, podría llegar a provocar. El DIA no surge de la nada, sino que, como todo derecho o sistema normativo, se autoconstituye a partir de un contexto social que, para su reproducción y diferenciación, debe reducir la complejidad del entorno creado por las expectativas normativas que el mismo proceso genera en los sujetos<sup>(21)</sup>.

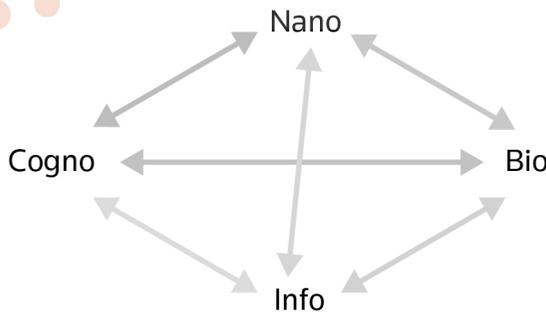
---

(20) En general y desde una perspectiva normativista, en este trabajo se entiende por fuentes materiales del derecho, los diversos factores (económicos, sociales, morales, religiosos, científicos, artísticos, técnico; la cultura, en general, etc.) que influyen en la producción, modificación y derogación de las normas jurídicas. Por fuentes formales, a su vez, entendemos “los distintos procedimientos de creación de normas jurídicas, así como los modos de exteriorización de estas y los continentes donde es posible hallarlas, tras los cuales procedimientos es posible identificar un órgano, una autoridad, una fuerza social o sujetos de derecho que tienen competencia para producir normas jurídicas, competencia para producir normas jurídicas, competencia que les es otorgada por el mismo ordenamiento jurídico al que pasan a incorporarse las nuevas normas por ellos creados” (Squella, Agustín, *Introducción al derecho*, 2a actualizada y ampliada, Editorial Jurídica de Chile, Santiago, 2011, p. 294). Así como la cultura, en un sentido muy amplio, es la fuente material por excelencia del derecho, la ley, a su vez, y también afirmado esto en un sentido muy general, es la fuente formal más reconocida del derecho.

(21) La concepción del derecho como sistema autopoiético (autoconstituido) por los procesos de diferenciación de la sociedad moderna mediante la reducción de las expectativas normativas de los sujetos es uno de los aportes de la omnicompreensiva sociología de Luhman, Niklas (*La sociedad de la sociedad*, trad. Javier Torres Nafarrate, Herder, México, 2007; *El derecho de la sociedad*, trad. Javier Torres Nafarrate, Herder, México, 2005).

La 4RI se caracteriza por el fenómeno de la convergencia de las tecnologías (CT). La IA no es un fenómeno tecnológico cuyo desarrollo esté teniendo lugar de forma aislada. La IA es una pieza constitutiva de la CT, descrita esta como “la interacción de diferentes disciplinas científicas-tecnológicas para abordar e intentar solucionar problemas comunes a todas ellas mediante la cooperación trans, inter y multidisciplinar”<sup>(22)</sup>. La CT se traduce en relaciones interactivas, sinergias o fusiones de diferentes campos de investigación y desarrollo tales como: nanociencia y nanotecnología; biotecnología, ciencias de la vida, biomedicina, ingeniería genética; tecnologías de la información y de la comunicación; robótica<sup>(23)</sup> e inteligencia artificial, así como las ciencias cognitivas, las neurociencias y las neurotecnologías.

No debe perderse de vista que la CT suele identificarse con el acrónimo NBIC (nano-bio-info-cogno), donde la letra N hace referencia a la nanotecnología y las nanociencias; la B, a la biotecnología, biomedicina, incluyendo a la ingeniería genética; la I, a las tecnologías de la información, computación y comunicaciones y la C, a las ciencias cognitivas, que incluye, por cierto, a las neurociencias. La expresión CT, por tanto, conceptualmente alude a la combinación sinérgica de las NBIC, en tanto fenómeno tecnológico en pleno desarrollo y auge, que está guiando la evolución y el progreso de la sociedad humana. Esta combinación sinérgica es representada, generalmente, a través de la figura del Tetraedro NBCI<sup>(24)</sup>:

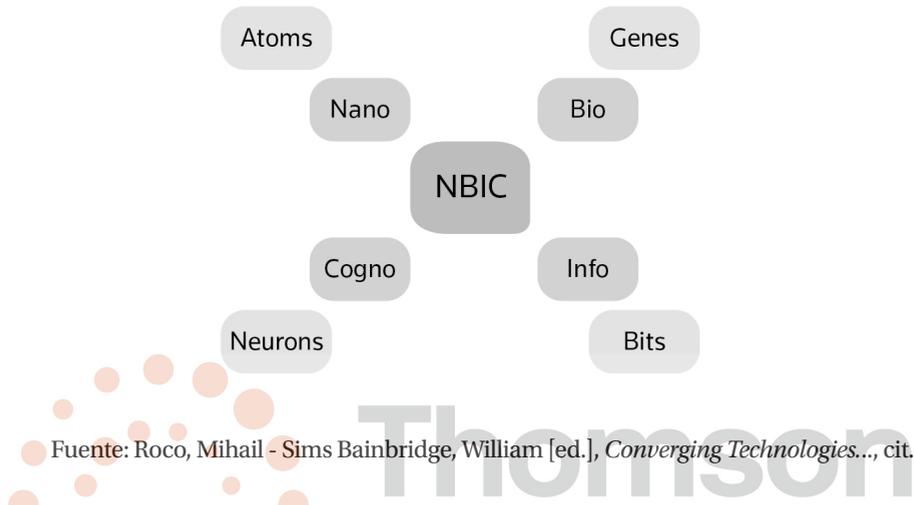


(22) Ursua, Nicanor, “La ‘convergencia de tecnologías’ (CT) y la ‘mejora técnica del ser humano’: una visión tecno-futurista”, *Thémata Revista de Filosofía*, nro. 46, p. 67.

(23) Nota del director: sobre el concepto de robótica y su relación con el campo de la inteligencia artificial ver Corvalán, Juan G. - Danesi, Cecilia - Carro, María Victoria “Responsabilidad civil de la inteligencia artificial”, tomo II del presente *Tratado*.

(24) Roco, Mihail - Sims Bainbridge, William (ed.), *Converging Technologies for Improving Human Performance. Nanotechnology, Biotechnology, Information Technology and Science*, National Science Foundation, Washington DC, 2003, p. 2.

El fenómeno NBIC, del cual la IA es parte fundamental, se basa en la unidad de la ciencia a nano escala ( $10^{-9}$ ) y en las tecnologías de integración nanométrica que la hacen posible. Sus unidades básicas de construcción sistémica son átomos, neuronas, genes y bits:



Fuente: Roco, Mihail - Sims Bainbridge, William [ed.], *Converging Technologies...*, cit.

La perspectiva y visión NBIC es una propuesta de desarrollo integral de las ciencias y tecnologías del siglo XXI, asumida oficialmente por los Estados Unidos de Norteamérica (EE.UU.) y Europa. Es el motor y modo actual del trabajo científico y tecnológico, y quienes no lo asuman se quedarán en el pasado.

El comienzo oficial de la estrategia NBIC se sitúa en el reporte de la National Science Foundation de USA (NSF), titulado “Converging Technologies for Improving Human Performance. Nanotechnology, Biotechnology, Information Technology and Cognitive Science” (2003)<sup>(25)</sup>. Las ciencias NBIC se despliegan en varios campos y, en lo que nos interesa, estas ciencias se están focalizando en la expansión de la cognición y comunicación, así como en la mejora de la salud y capacidades fenotípicas humanas mediante interfaces de cerebro a cerebro y de cerebro a máquina.

En efecto, el citado reporte enfatiza que los rápidos avances en tecnologías convergentes (NBIC) tienen el potencial de mejorar el desempeño humano y la productividad de la Nación. Los ejemplos de beneficios incluirían la mejora de la eficiencia y el aprendizaje en el trabajo; la mejora de las capacidades sensoriales y cognitivas individuales; la creación de nuevos procesos de fabricación y la aparición de productos mejorados;

(25) Ídem.

cambios revolucionarios en la atención médica; la mejora de la eficiencia individual y grupal así como el desarrollo de técnicas de comunicación efectiva que incluyen interacción cerebro a cerebro, de máquinas y cerebros gracias a la ingeniería neuromórfica para uso industrial y personal, mejorando las capacidades del capital humano para propósitos de producción y defensa, alcanzando un desarrollo sostenible y mejorando el deterioro físico y cognitivo que es común en el envejecimiento de la mente.

El enfoque de las ciencias NBIC de los EE.UU. se centra en la mejora técnica del ser humano (*human enhancement*), defendiendo una visión nanométrica del desarrollo científico para conseguir ese objetivo. Este enfoque puede resumirse en la siguiente consigna que W. A. Wallace, uno de los participantes del *workshop* que dio origen al reporte, pronunció:

“In the Cognitive Scientist can think it

”The Nano people can build it

”The Bio people can implement it, and

”The IT people can monitor and control it”<sup>(26)</sup>.

Esto es: si el científico cognitivo puede concebirlo, el nano puede construirlo, el bio puede aplicarlo y el info puede monitorizarlo y controlarlo.

Debe también considerarse que la Unión Europea (a través de la Comisión Europea), patrocinó el proyecto CONTECS que dio como fruto al reporte “Converging Technologies and their impact on the Social Sciences and Humanities (CONTECS). An analysis of critical issues and suggestion for a future research agenda, Final Report, May 2008”<sup>(27)</sup>. En dicho informe, al igual que en el de la NSF, juntamente con hacer una disensión técnica a las NBIC, se alzan algunas las voces en torno a los problemas éticos que el prometido mundo futuro de la IA y las NBIC, están generando. El proyecto CONTECS señala que si bien las actividades de política de investigación sobre convergencia, que hasta hace poco se habían limitado en gran medida a proyectos de prospectiva, reflejan la diversidad internacional de agendas, el debate sobre CT, incluso en la

---

(26) *Ibíd.*, p. 13.

(27) Andler, Daniel *et al.*, *Converging Technologies and their impact on the Social Sciences and Humanities (CONTECS). An analysis of critical issues and a suggestion for a future research agenda. Final report*, Fraunhofer Institute for Systems and Innovation, Germany, 2008.

Unión Europea, exhibe un enfoque claro en cuanto a que las tecnologías pueden ser utilizadas para la mejora humana; para una modificación masiva de cuerpos humanos en términos de una posible reconstrucción del hombre, o incluso para la creación de seres posthumanos. Estas visiones de gran alcance que no solo han acompañado, sino que, en cierto modo, han determinado el discurso sobre la CT, ya han provocado debates en campos de investigación como la ética de la tecnología, la evaluación de la tecnología, los estudios utópicos, la teología y en diversos subcampos de la ciencia y estudios tecnológicos, controvirtiendo el argumento naturalista de que los seres humanos son fisiológicamente lo que la evolución ha hecho de ellos<sup>(28)</sup>.

## 2.1. CARACTERIZACIÓN DE LA 4RI

La 4RI<sup>(29)</sup> es el contexto desde donde emerge y se está autoconstituyendo el DIA. La caracterizaremos siguiendo la perspectiva que sobre estas materias tiene el economista Klaus Schwab, quien introduce sus ideas señalando que el primer cambio revolucionario que, en sus estructuras socioeconómicas experimentó la humanidad, fue la transición de una economía recolectora a otra centrada en la agricultura, haciendo posible una forma de vida campesina.

Para Schwab la humanidad ha experimentado cuatro revoluciones industriales (esta última, la cuarta, todavía en sus albores, en su tema central). La periodización que propone señala que la primera revolución industrial aconteció aproximadamente entre 1760 hasta 1840, y fue desencadenada por la construcción del ferrocarril y la invención del motor de vapor, inaugurando los sistemas de producción mecánica. La segunda revolución industrial, tuvo lugar entre finales del siglo XIX y principios del XX, y se caracterizó por la producción en masa, gracias a la incorporación de la electricidad y la cadena de montaje. La tercera revolución industrial, comienza en 1960. Se trata de la revolución digital “catalizada por el desarrollo de los semiconductores, la computación mediante

(28) Ursua, Nicanor, “La ‘convergencia de tecnologías’ (CT) y la ‘mejora técnica del ser humano’: una visión tecno-futurista”, *Thémata Revista de Filosofía*, nro. 46, 2012, p. 70.

(29) Solo mencionaré, ya que los límites de extensión de este trabajo impiden un mayor desarrollo, que Jeremy Rifkin sostiene que estamos viviendo los tiempos de la 3RI, ofreciendo además una perspectiva conceptual distinta, véase: Rifkin, Jeremy, *La tercera revolución industrial. Cómo el poder lateral está transformando la energía, la economía y el mundo*, 1a ed., 5a reimp., trad. Albino Santos Mosquera, Paidós, Barcelona, 2016.

servidores tipo *mainframe* (en los años sesenta), la informática personal (décadas de 1970 y 1980) e internet (década de 1990)”<sup>(30)</sup>.

Pues bien, plenamente consciente de que algunos académicos y profesionales consideran que la evolución que analiza en su libro es simplemente parte de la 3RI, Schwab propone también la existencia de una 4RI que estaría en pleno desarrollo, que se desmarcaría del modelo de la 3RI, y “que no se parece a nada que la humanidad haya experimentado antes”. Hace esta aseveración, fundado en tres observaciones empíricas: (1) la velocidad de esta revolución que avanza a ritmo exponencial; (2) su amplitud y profundidad signada por la convergencia de las tecnologías que están cambiando sin precedentes la sociedad, la economía, los negocios, las personas, lo cual le lleva a sostener que no solo está cambiando el “qué” y el “cómo” hacer las cosas, sino el “quién somos”, y, (3) el impacto en los sistemas, es decir, en el funcionamiento estructural de la sociedad<sup>(31)</sup>.

Esta 4RI que, por cierto, tiene como base los avances tecnológicos que se alcanzaron en la tercera, comenzó a principios del siglo XXI y se caracteriza por un internet más ubicuo y móvil, por sensores más pequeños y potentes que son cada vez más baratos, y por la inteligencia artificial y el aprendizaje de la máquina. Sin embargo, la 4RI, argumenta Schwab, “no solo consiste en máquinas y sistemas inteligentes. Su alcance es más amplio”<sup>(32)</sup>.

Explica también Schwab que lo nuevo o revolucionario de estas transformaciones no se constituye por la existencia de las tecnologías digitales (computación y redes), sino porque estas son cada vez son más sofisticadas, complejas e integradas, cuestión que está transformando las sociedades y la economía mundial, en términos tales que todo esto se manifestará a través de la automatización y creación de cosas sin precedentes en la historia de la humanidad. La industria de la 4RI es el modelo “industria 4.0”, que dibuja un paisaje tecnológico poblado de fábricas inteligentes, sistemas de fabricaciones virtuales y físicos interconectados globalmente<sup>(33)</sup>.

---

(30) Schwab, Klaus, *La cuarta revolución industrial*, 3a reimp. de la 1a ed. 2016, Random House Grupo Editorial, Barcelona, 2017, p. 20.

(31) *Ibíd.*, p. 21.

(32) *Ibíd.*, p. 20.

(33) El término Industria 4.0 fue introducido por primera vez en la Feria de Hannover en el año 2011. Se trata de una investigación profunda que postula que estamos en una cuarta revolución industrial por obra de la introducción de la internet de las cosas y de los servicios, véase: Kagermann, Henning - Wahlster, Wolfgang - Johannes Helbig, “Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0

La 4RI es un fenómeno complejo que, por su profundidad, se extiende desde la secuenciación genética hasta la nanotecnología, y de las energías renovables a la computación cuántica. Así es como por todo lo anteriormente expuesto, Schwab concluye que es la fusión de estas tecnologías y su interacción a través de los dominios físicos, digitales y biológicos lo que hace que la cuarta revolución industrial sea fundamentalmente diferente a las anteriores.

Frente a estos vertiginosos cambios, Schwab alza su voz de preocupación porque, a su juicio, existen al menos dos factores que podrían limitar el potencial de la 4RI, a saber: (1) la falta de adecuación del marco institucional para orientar “la difusión de la innovación y mitigar la disrupción”, y, (2) la inexistencia de “una narrativa consistente, positiva y común que describa las oportunidades y los desafíos de la cuarta revolución industrial, una narrativa que es esencial si queremos empoderar a un conjunto diverso de individuos y comunidades, y evitar una violenta reacción popular contra los cambios fundamentales en curso”<sup>(34)</sup>.

Analíticamente Schwab distingue megatendencias impulsoras y puntos de inflexión de la 4RI<sup>(35)</sup>.

Las megatendencias impulsoras, que aprovechan el poder de penetración de la digitalización y las tecnologías de la información, se manifiestan en tres grupos: físicos, digitales y biológicos.

- (1) Las manifestaciones físicas escogidas son: vehículos autónomos, impresión 3D, robótica avanzada y nuevos materiales.
- (2) Las manifestaciones digitales que destaca Schwab son: internet de las cosas y/o internet de todas las cosas, y ciertas formas de colaboración virtual, como el *blockchain* y los negocios de plataforma que hacen posible lo que se denomina “economía bajo demanda o consumo colaborativo”.
- (3) Y la secuenciación e ingeniería genética sobre todo retroalimentada por el enorme poder de cómputo actual, junto con la biología sintética, son las megatendencias impulsoras biológicas.

- Securing the future of German manufacturing industry. Recommendations for implementing the strategic initiative INDUSTRIE 4.0. Final report of the Industrie 4.0 Working Group”, Frankfurt/Main, National Academy of Science and Engineering, Federal Ministry of Education and Research, 2013, p. 5.

(34) Schwab, Klaus, *La cuarta revolución industrial*, 3a reimp. de la 1a ed. 2016, Random House Grupo Editorial, Barcelona, 2017, p. 22.

(35) *Ibíd.*, ps. 29-44.

Los puntos de inflexión son un conjunto de predicciones realizado en el año 2015 por el Foro Económico Mundial, que presenta una serie cambios tecnológicos específicos que darán forma al futuro mundo digital hiperconectado. Se trató de una encuesta realizada entre 800 ejecutivos y expertos del sector tecnologías de la información y de las comunicaciones. Se les preguntó qué cambios esperarían que ocurrieran antes del año 2025. Se levantaron 21 cambios que se esperan que ocurran. No los mencionaremos todos. Aquí presentamos algunos ejemplos: 91,2% de los encuestados cree que las personas humanas usarán ropa conectada a internet; 85,5% estimó que el 10% de las gafas estarán conectadas a internet; el 84,1% de los encuestados señaló que antes del 2025 el primer automóvil impreso en 3D estará en producción; el 78,8% estimó que el 90% de la población mundial tendrá acceso a internet; el 76,4% de los consultados espera que el primer trasplante de un hígado impreso en 3D se efectúe; el 78,2% pronostica que el 10% de los automóviles que transiten por las carreteras de USA, serán sin conductor; el 69,9% consideró que más del 50% del tráfico de internet en los hogares será para electrodomésticos y dispositivos.

#### 2.1.1. LA CUARTA REVOLUCIÓN INDUSTRIAL EN LA ERA DE LA INFORMACIÓN

Una mirada aún más general de la 4RI indica que esta está teniendo lugar en una etapa especial del desarrollo de la humanidad. La interpretación sociológica e histórica que propongo señala que la 4RI podría ser contextualizada dentro de la Era de Información, así como es descrita y teorizada por Castells.

La era de la información se trata, según Castells:

“(…) de un acontecimiento histórico al menos tan importante como lo fue la revolución industrial del siglo XVIII, inductor de discontinuidad en la base material de la economía, la sociedad y la cultura”. Y, “a diferencia de cualquier otra revolución, el *núcleo* de la transformación que estamos experimentando en la revolución en curso remite a *las tecnologías del procesamiento de la información y de la comunicación*. La tecnología de la información es a esta revolución lo que las nuevas fuentes de energía fueron a las sucesivas revoluciones industriales del motor de vapor a los combustibles fósiles e incluso a la energía nuclear, ya que la generación y distribución de energía fue el elemento clave subyacente en la sociedad industrial”<sup>(36)</sup>.

---

(36) Castells, Manuel, *La era de la información. Economía, sociedad y cultura*, Siglo XXI, México, 2008, t. I, p. 57.

Para no introducir equívocos, Castells aclara:

“(…) lo que caracteriza a la revolución tecnológica actual no es el carácter central del conocimiento y la información, sino la aplicación de ese conocimiento e información a aparatos de generación de conocimiento y procesos de la información, en un círculo de retroalimentación entre la innovación y sus usos (...) Como resultado, la difusión de la tecnología amplifica infinitamente su poder al apropiársela y redefinirla sus usuarios (...) Por primera vez en la historia, la mente humana es una fuerza productiva directa, no sólo un elemento decisivo del sistema de producción”<sup>(37)</sup>.

Otra importante nota distintiva de la actual revolución tecnológica es el carácter global de esta, la cual se ha extendido desigualmente y de forma selectiva, tanto social como funcionalmente, pero bajo una lógica de aplicación inmediata enlazando el mundo mediante la tecnología. Este paradigma de la Tecnología de la Información se caracteriza por una serie de rasgos que constituyen la base de la sociedad de la información, a saber: (1) la información es su materia prima; (2) una capacidad de penetración en todos los procesos de la vida individual y colectiva por los nuevos medios tecnológicos; (3) una lógica interconexión de todo sistema o conjunto de relaciones que utilizan estas nuevas tecnologías de la información; (4) el paradigma de la Información se basa también en la flexibilidad (reversibilidad, adaptabilidad); (5) una quinta característica de esta revolución tecnológica es la convergencia creciente de tecnologías específicas en un sistema altamente integrado (convergencia tecnológica).

Como puede observarse, Castells incorpora en las características 3) y 5) de la era de la información, el fenómeno de la CT y anticipa, de algún modo, la lógica de funcionamiento sinérgico de las tecnologías NBIC. Castells también destaca que el surgimiento de esta nueva estructura social está asociado con la emergencia de un nuevo modo de desarrollo. Los modos de desarrollo son los dispositivos tecnológicos mediante los cuales el trabajo actúa sobre la materia para generar el producto, determinando en definitiva la cuantía y calidad del excedente.

El informacionalismo es el modo de desarrollo actual. Y se define históricamente por la reestructuración del modo capitalista de producción hacia finales del siglo XX. En este nuevo modo de desarrollo, la fuente de la productividad estriba en la tecnología de la generación de conocimiento, el procesamiento de la información y la comunicación de símbolos, en tanto se trata de la acción del conocimiento sobre sí mismo como principal fuente de productividad, hecho distintivo y circunstancia estructural que no se da los otros modos de desarrollo (industrialismo, por ejemplo).

(37) *Ibíd.*, p. 58.

La hipótesis de Castells dice que la sociedad informacional está cambiando las formas de comunicación de la especie humana y con ello, el mundo, la sociedad, la humanidad. La especie humana se define por su capacidad de comunicación significativa sustentada en el lenguaje reflexivo. El lenguaje humano posee aquello que los antropólogos denominan universalidad semántica, es decir, capacidad para comunicar “información acerca de aspectos, dominios, propiedades, lugares o acontecimientos del pasado, del presente o del futuro, tanto reales como posibles, verdaderos o imaginarios, cercanos o lejanos”<sup>(38)</sup>. El operar o despliegue del lenguaje y la comunicación construyen el mundo que habitamos. Por tanto, si los ritmos y procesos de la comunicación cambian de modo sustancial, es razonable esperar una transformación en el mundo.

El concepto era de la información captura la esencia del tiempo histórico industrial de fines de siglo XX y principios del XXI, signado por la emergencia de un paradigma tecnológico organizado en torno a las tecnologías de la información. Este paradigma —epistemológicamente— implica también una comprensión del mundo a partir de la información como unidad analítica básica de ella. El hombre, la sociedad, sus instituciones y sujetos son vistos como bits de información; átomos, neuronas, genes; como sistemas que pueden ser descodificados, así como el ADN. Lo central en este paradigma, como destaca Castells, es la aplicación de la información y el conocimiento a aparatos de generación de conocimiento y procesamiento de la información/comunicación (algoritmos evolutivos, sistemas expertos), en un círculo de retroalimentación acumulativo entre la innovación y los usos.

## 2.2. LAS TECNOLOGÍAS DE LA CUARTA REVOLUCIÓN Y EN PARTICULAR DE LA INTELIGENCIA ARTIFICIAL

Refiriéndose a la primera revolución industrial, decía Hobsbawm, que hablar de revolución industrial es hablar de algodón<sup>(39)</sup>. Siguiendo esa analogía, conviene preguntarse: ¿de qué se habla cuando se habla de 4RI? La respuesta a esta altura de la investigación es evidente. En términos generales, se habla de tecnologías de la información. Y más específicamente, no se puede hablar de 4RI, sin hablar de tecnologías de la información operando e interactuando fusionadamente en los dominios físicos, digitales y biológicos: IA, robótica, Internet de las cosas y de los

(38) Harris, Marvin, *Introducción a la antropología general*, 3a reimp. de la 1a ed., trad. de Francisco Muñoz de Bustillo [original en inglés: *Culture, People, Nature: An introduction to General Anthropology*], Alianza Editorial, España, 2011, p. 164.

(39) Hobsbawm, Eric J., *Industria e Imperio. Una historia económica de Gran Bretaña desde 1750*, 2a ed., Ariel, Barcelona, 1982, p. 55.

servicios, Nanotecnología, Biotecnología (biología sintética y genética), neurociencias, en otras palabras: tecnologías NBIC.

La 4RI representa un salto cualitativo en relación con la tercera, pues en este estadio de desarrollo los productos de las tecnologías de la información (internet, máquinas, robots, secuenciación genética, etc.) operan conectados y de manera inteligente, es decir, retroalimentándose y aprendiendo con su propio operar. Su avance es exponencial y el poder penetración irradia prácticamente todas las áreas del conocimiento y del quehacer humano.

En el contexto de la era de la información y de la 4RI, las tensiones éticas y jurídicas más evidentes de detectar provienen del campo de la IA y de las tecnologías de la vida que buscan el mejoramiento de los seres humanos (*human enhancement*). Ambas tecnologías, así como en general las NBIC, si se orientan sin control normativo hacia el mejoramiento del ser humano, pueden terminar por cambiar por completo su naturaleza. Pues bien, en la subsección que sigue, presento una breve descripción de la IA para crear la imagen de cómo esta tecnología podría modificar al ser humano y a la humanidad.

### 2.2.1. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

Comienzo proponiendo un concepto operacional de inteligencia, para luego aproximarnos a la definición de la IA. La inteligencia es un subconjunto estricto de la *computación*, es decir, de transformación de la información. Hay que tener en cuenta que la computación es un proceso físico, no matemático: requiere tiempo, espacio y energía. La inteligencia, así, puede ser vista como un subconjunto de la computación que transforma un contexto en acción<sup>(40)</sup>.

No existe una definición única de IA que sea universalmente aceptada por los profesionales del campo<sup>(41)</sup>. Algunos definen la IA como un sistema computarizado que exhibe un comportamiento que comúnmente se considera que requiere inteligencia. Otros definen la IA como un sistema capaz de resolver racionalmente problemas complejos o tomar acciones

(40) Bryson J., Joanna, "La última década y el futuro del impacto de la IA en la sociedad", en *¿Hacia una nueva ilustración? Una década trascendente*, BBVA/OpenMind, Madrid, 2018, p. 131.

(41) Nota del director: los organismos internacionales no proponen una definición uniforme de inteligencia artificial. Sobre cada una de ellas ver Corvalán, Juan G. - Díaz Dávila, Laura - Simari, Gerardo, "Inteligencia artificial: bases conceptuales para una aproximación interdisciplinar", apart. "Un menú de definiciones de 'inteligencia artificial'", tomo I del presente *Tratado*.

apropiadas para lograr sus objetivos en cualquier circunstancia del mundo real que se encuentre<sup>(42)</sup>.

Por la variedad de fuentes de las cuales se nutre, no es fácil definirla, ya que depende de la perspectiva o enfoque desde el cual se pretenda asir su esencia. Y es que la IA es una ciencia compleja construida con fundamentos extraídos principalmente de la filosofía, las matemáticas, la economía, la teoría de la información, las neurociencias, la psicología, la ingeniería computacional (*hardware y software*), la teoría del control, la cibernética y la lingüística, entre otras disciplinas.

Sin perjuicio de las dificultades definitorias aludidas, en términos generales la IA podría ser comprendida como una disciplina científica que busca desarrollar métodos y algoritmos soportados en sustratos artificiales que permitan generar comportamientos inteligentes. La IA sería aquella que exhiben ciertos sistemas o artefactos construidos por el hombre. Así es como se ha conjeturado que un sistema artificial poseería inteligencia cuando es capaz de llevar a cabo tareas que, si fuesen realizadas por un humano, se diría de este que es inteligente.

Uno de los libros más autorizados y reconocidos del campo de la IA<sup>(43)</sup>, luego de analizar las definiciones de IA en ocho libros de texto, señala que es posible distinguir los siguientes enfoques en la materia:

- (1) Comportamiento humano: el enfoque del test de Turing<sup>(44)</sup>: un sistema de IA es inteligente en la medida en que para un observador realiza conductas humanas.

---

(42) National Science and Technology Council, *Preparing for the future of artificial intelligence*, Executive Office of the President, National Science and Technology Council, Committee on Technology, Washington DC, 2016, p. 6.

(43) Russell, Stuart - Norvig, Peter, *Inteligencia artificial. Un enfoque moderno*, Pearson/Prentice Hall, Madrid, 2008, trad. Juan Manuel Corchado Rodríguez *et al.*

(44) El test de Turing tiene por objeto determinar si un ordenador puede convencer que es humano a un observador que lo interroga, durante el curso de una conversación experimental. El test debe su nombre a su creador, Alan Mathison Turing (1912-1954), matemático y lógico, considerado uno de los padres fundadores de las ciencias de la computación y la inteligencia. Cabe indicar que Turing planteó diversas versiones de su test. El test de Turing es una versión del denominado juego de imitación que, en su versión clásica, como el mismo Turing explica, “se juega con tres personas, un hombre (A), una mujer (B) y un interrogador (C) que puede ser de cualquier sexo. El interrogador se queda en una habitación separada frente a las otras dos. El objetivo del juego para el interrogador es determinar cuál de los otros dos es el hombre y cuál es la mujer. Él los conoce con las etiquetas X e Y, y al final del juego dice ‘X es A e Y es B’ o ‘X es B e Y es A’” (Turing, Alan, “Computing Machinery and Intelligence”, *Mind*, 1950, 59 [236]: 433-460). El interrogador formula preguntas y los jugadores responden por escrito, para que el timbre de la voz no le de pistas de identificación. Pues bien, Turing, modifica el juego de imitación, reemplazando a uno de

- (2) Pensar como humano: el enfoque del modelo cognitivo: un sistema de IA es inteligente en la medida que piense como un humano. Para esto se necesita tener una teoría del conocimiento humano y expresarla, además, algorítmicamente y llevarla al *software*. Esta tendencia en IA está muy unida a las ciencias cognitivas. Ambos campos científicos se retroalimentan, sobre todo en las áreas de visión de colores y lenguaje natural.
- (3) Pensamiento racional: el enfoque de las “leyes del pensamiento”. Para esta concepción de la IA, un sistema es inteligente en la medida que resuelve problemas lógicos formales. La llamada tradición logicista dentro del campo intenta construir programas que puedan resolver problemas descritos en notación lógica. La gran dificultad de este enfoque deriva del hecho que no es simple traducir el conocimiento informal a un sistema de notación lógica.
- (4) Actuar de forma racional: sobre este enfoque, explican Russel y Norvig que un agente es algo que razona. Pero explican que de los agentes informáticos se espera que tengan otros atributos que los distinguen de los “programas” convencionales, como que estén dotados de controles autónomos, que perciban su entorno, que persistan durante un período de tiempo prolongado, que se adapten a los cambios, y que sean capaces de alcanzar objetivos diferentes. Un agente racional es aquel que actúa con la intención de alcanzar el mejor resultado o, cuando hay incertidumbre, el mejor resultado esperado<sup>(45)</sup>. Las ventajas de este enfoque es que no descarta el razonamiento lógico formal, pero entiende que hay muchas conductas que no pueden inferirse lógicamente y que, sin embargo, pueden ser consideradas correctas, así como también hay situaciones donde no existe una salida correcta.

los jugadores por un ordenador digital. Si al final del juego el interrogador no identifica correctamente quién es el humano o quién es la máquina, entonces la conclusión es que el ordenador desarrolló un conjunto de respuestas que aparentan haber sido realizados por un proceso mental humano, indistinguible de la máquina computacional. Turing no pretendió definir con el test un concepto de pensamiento en términos de aplicarlo, digamos, ontológicamente, a la distinción pensamiento humano/pensamiento computacional. Sabía que ese tema era como un zapato chino y prefirió mostrar cuán complejo puede resultar determinar el sentido de lo que entendemos por pensamiento y resaltar con eso la existencia de un límite entre las capacidades físicas y las intelectuales humanas. Sin embargo, con el correr de los años, el test de Turing ha sido convertido en una especie test para saber cuándo los ordenadores digitales podrán igualar o superar el pensamiento humano.

(45) Russell, Stuart - Norvig, Peter, *Inteligencia artificial. Un enfoque moderno*, trad. Juan Manuel Corchado Rodríguez *et al.*, Pearson/Prentice Hall, Madrid, 2008, ps. 2-5.

Otra forma de acercarse a la comprensión conceptual de la IA es a través de la distinción IA General e IA Especial o Experta.

La primera es aquella que puede realizar cualquier tarea cognoscitiva posible de observar en un ser humano. Es la gran meta de la IA.

La IA especial es aquella se resuelve problemas y entrega resultados en áreas particulares, emulando y muchas veces superando las capacidades humanas, sobre todo por las deducciones que se obtienen mediante el análisis de volúmenes gigantes de información.

La IA especial o experta solo puede abocarse a la realización de las tareas que el programa le permite. Así, por ejemplo, los programas que juegan ajedrez a nivel de Gran Maestro son incapaces de jugar a las damas a pesar de ser un juego mucho más sencillo. Se requiere diseñar y ejecutar un programa distinto e independiente del que le permite jugar al ajedrez para que el mismo ordenador juegue también a las damas. En el caso de los seres humanos no es así, ya que cualquier jugador de ajedrez puede aprovechar sus conocimientos sobre este juego para, en cuestión de segundos, jugar a las damas<sup>(46)</sup>. Cabe tener presente que en la actualidad la IA ha logrado éxitos espectaculares en sistemas expertos.

Otra mirada al tema distingue entre IA Fuerte e IA Débil<sup>(47)</sup>. Esta distinción fue introducida críticamente por John Searle. La IA fuerte sería aquella que no simula una mente, sino que es una mente y, por tanto, debería pensar igual que un humano. Searle construyó estas distinciones, en conjunto con el experimento de la Habitación China, para demostrar que la IA fuerte es imposible.

La IA Débil correspondería a la inteligencia que podemos observar en los sistemas expertos o especializados. La IA Fuerte siempre será necesariamente general, pero es posible distinguir IA generales que no son fuertes, es decir, “que simulen la capacidad de exhibir inteligencia general de la mente, pero sin ser mentes”<sup>(48)</sup>. Hasta el momento, como sabemos, no se han construido IA Fuertes y tampoco IA generales.

---

(46) López de Mántaras Badia, Ramón - Meseguer González, Pedro, *Inteligencia artificial*, CSIC/Los Libros de la Catarata, Madrid, 2017, e-book, loc. 67-74.

(47) *Ibíd.*, loc. 81.

Nota del director: sobre esta distinción también Corvalán, Juan G. - Díaz Dávila, Laura - Simari, Gerardo, “Inteligencia artificial: bases conceptuales para una aproximación interdisciplinar”, tomo I del presente *Tratado*.

(48) López de Mántaras Badia, Ramón - Meseguer González, Pedro, *Inteligencia artificial*, CSIC/Los Libros de la Catarata, Madrid, 2017, e-book, loc. 81.

Esta distinción entre IA Débil y Fuerte permite proponer, respectivamente, dos definiciones de IA.

IA Débil sería la ciencia e ingeniería que permite diseñar y programar ordenadores de forma que realicen tareas que requieran inteligencia (resolver problemas matemáticos, estadísticos, jugar ajedrez, dirigir un centro logístico, inferir tendencias —*big data*—, dirigir un misil, hacer radiología, etc.). Por su parte, IA fuerte, sería la ciencia e ingeniería que permitiría replicar la inteligencia humana mediante máquinas.

Si se busca una definición operacional de IA, la que propone el Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial, creado por la Comisión Europea en junio de 2018, resulta una alternativa muy autorizada, a saber: sistemas que manifiestan un comportamiento inteligente, pues son capaces de analizar su entorno y pasar a la acción —con cierto grado de autonomía— con el fin de alcanzar objetivos específicos, y que pueden consistir simplemente en un programa informático (p. ej., asistentes de voz, programas de análisis de imágenes, motores de búsqueda, sistemas de reconocimiento facial y de voz) o estar incorporada en dispositivos de *hardware* (p. ej., robots avanzados, automóviles autónomos<sup>(49)</sup>, drones<sup>(50)</sup> o aplicaciones del internet de las cosas)<sup>(51)</sup>.

En suma, convengamos en que, si pudiera definirse en función de los objetivos la IA, se podría señalar que estos son construir sistemas (físicos o virtuales) capaces de deducir, razonar, resolver problemas, planificar, aprender; procesar lenguajes naturales, mostrar creatividad, inteligencia social y general, así como tener también capacidad de movimiento y percepción, en el caso de lo que comúnmente conocemos como robots físicos.

El objetivo final de la IA se resumiría en construir una máquina que tenga una inteligencia de tipo general similar a la humana (portadora de sentido común). Aquel objetivo supone crear los programas que puedan

(49) Nota del director: sobre automóviles autónomos ver Corvalán, Juan G. - Danesi, Cecilia, Carro M. Victoria, “Responsabilidad civil de la inteligencia artificial”, tomo II del presente *Tratado*.

(50) Nota del director: sobre drones ver Dupuy, Daniela, “Una mirada integral del fenómeno de la ciberdelincuencia y el impacto de la inteligencia artificial en las investigaciones digitales”, tomo IV del presente *Tratado*. Asimismo, ver Martino, Antonio A., “Ética y sistemas inteligentes”, apart. “La cosa se pone más peligrosa”, tomo I del presente *Tratado*.

(51) European Commission, *A definition of IA: main capabilities and disciplines*, Brussels, High-Level Expert Group on Artificial Intelligence, 2019.

hacerlo (descubrir los algoritmos) y la construcción de los soportes físicos para el operar de esos programas (las máquinas, el computador, los robots). Ese objetivo no se ha cumplido todavía. Pero existen autorizadas predicciones que informan que, gracias al operar de las tecnologías NBIC, la humanidad se está acercando a estadios de desarrollo que permitirían mantener el objetivo como una meta razonable.

En la actualidad, gran parte del optimismo deriva de los sorprendentes avances en sistemas expertos y *machine learning*; nanotecnología, computación cuántica y del desarrollo exponencial que el poder de cómputo puede llegar a experimentar, si se mantiene el ritmo de crecimiento observado desde su nacimiento. No hay que perder de vista también, que hay vertientes dentro del tecno-optimismo, que predicen una era en que el desarrollo exponencial del poder de cómputo nos transportará a la Singularidad (en el 2045), definida esta como un estadio de desarrollo donde se fusionará la inteligencia biológica con la no-biológica, creando una superinteligencia que se expandirá por todo el universo<sup>(52)</sup>.

### 2.2.2. IMPACTO DE LA IA EN LA SOCIEDAD

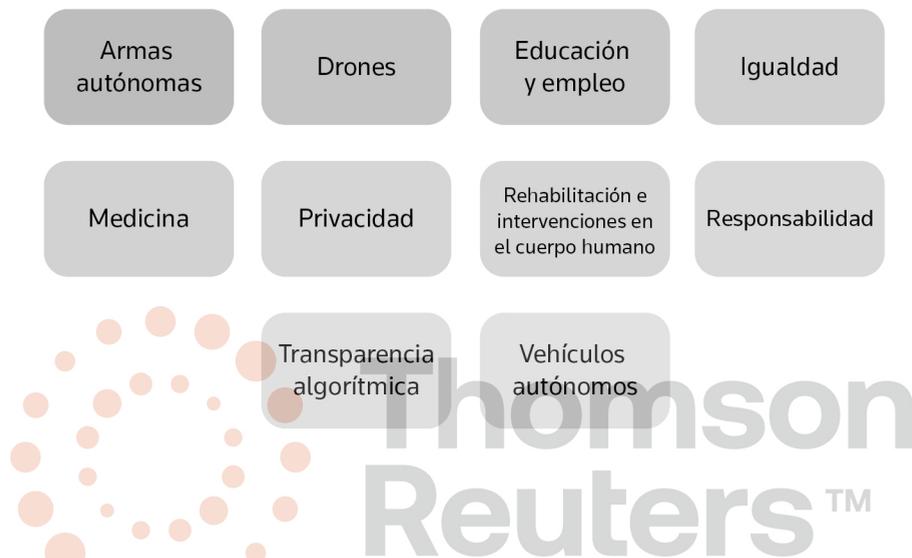
Dado que el objetivo de este estudio es fundamentar el DIA, a continuación, se presentará el impacto de la IA donde ya está jugando un rol transformador y en donde, potencialmente, ya se avizora que lo hará. Estas transformaciones reales y potenciales están generando tensiones éticas jurídicas en las sociedades del siglo XXI. Estas tensiones se originan porque las tecnologías de la 4RI, en especial la IA, producen expectativas normativas en uno o varios sentidos sin que, hasta la fecha, se hayan podido procesar satisfactoriamente en un orden jurídico estable. Ese conjunto desordenado de expectativas jurídicas, es decir, ese conjunto de expectativas no ordenadas en función de algún sentido normativo general, puede ser representado como un estado complejo; como un estado donde resulta extremadamente difícil poner en relación cada elemento en un determinado estado de cosas. La interpretación sociológica jurídica que sigo en esta investigación, dice que el torbellino de expectativas normativas que la 4RI genera en la sociedad, y que por lo mismo posibilita su ordenación y autoconstitución sistémica, es el motor sistémico diferenciador de configuración del DIA.

---

(52) Kurzweil, Raymond, *La singularidad está cerca. Cuando los humanos trascendamos la biología*, trad. Carlos García Hernández, Lola Books GbR, Berlín, 2012, e-book, loc. 352.

Entre otras, las áreas de impacto particulares más visibles de la IA, que están produciendo un aumento de las expectativas normativas de los sujetos, son las siguientes<sup>(53)</sup>:

### Algunas de las principales áreas actuales de riesgo e impacto de IA



En este contexto de aumento de las expectativas normativas, surgen propuestas de nuevos derechos. Pérez Luño, por ejemplo, señala que estaría emergencia una tercera generación de derechos humanos: los derechos de la era tecnológica<sup>(54)</sup>:

(53) Establecidas a partir de la bibliografía citada, principalmente las provenientes de fuentes gubernamentales, organismos internacionales e instituciones académicas privadas y públicas.

(54) Pérez Luño, Antonio, *La tercera generación de derechos humanos*, Aranzadi, Navarra, 2006. La terminología “derechos humanos de la tercera generación” fue acuñada por Karel Vasak, director del Departamento Jurídico de la UNESCO. Vasak pronunció en 1979, la Lección Inaugural de la Décima Sesión del Instituto Internacional de Derechos Humanos de Estrasburgo que llevaba por título: *Pour les droits de l’homme de la troisième génération*. Es una categoría muy controversial. Algunos autores proclaman el fin de la teoría de las generaciones de derechos humanos, véase: Fajardo, Luis A., “Globalización de los derechos humanos. La ‘teoría de las generaciones’ ha muerto. ¡Viva la teoría de la generación de Vierna!”, *Revista IUSTA*, 2 (31), 2016, disponible en <https://revistas.usantotomas.edu.co/index.php/iusta/article/view/3067> (consultado el 08/03/2021).

Los derechos relativos al medio ambiente, la calidad de la vida y la paz

Los derechos en el ámbito de las TIC's

Los derechos en la esfera de la bioética y de las biotecnologías

Derechos afectados: intimidad, identidad, libertad, autonomía, igualdad

Riofrío Martínez, por su parte, postula una cuarta generación de derechos humanos, integrada por los siguientes<sup>(55)</sup>:

a) El derecho a existir digitalmente

b) El derecho a la reputación digital

c) La estima digital

d) La libertad y responsabilidad digital

e) La privacidad virtual, el derecho al olvido, el derecho al anonimato

f) El derecho al *big-reply*

g) El derecho al domicilio digital

h) El derecho a la técnica, al *update*, al parche

i) El derecho a la paz cibernética y a la seguridad informática

j) El derecho al testamento digital

(55) Riofrío Martínez-Villalba, Juan Carlos, "La cuarta ola de derechos humanos: los derechos digitales", *Revista Latinoamericana de Derechos Humanos*, vol. 25 (1), 2014, ps. 15-45.

Desde el punto de vista de la protección de datos en la era de la IA, a su turno, Corvalán postula los siguientes derechos específicos que surgen a partir de la generación de perfiles, *rankings* y subastas: (1) derecho a no ser perfilado, (2) derecho a que no se tomen decisiones basadas únicamente en perfiles realizados a partir de IA, (3) derecho a que el perfilamiento no sea realizado sobre datos sensibles, (4) derecho a conocer en qué consiste un perfilado o *scoring*, (5) derecho a conocer las condiciones y criterios de perfilamiento y (6) derecho a conocer qué se entiende por subasta y si las empresas realmente las realizan<sup>(56)</sup>.

Como se aprecia, el impacto de la IA en diferentes campos y áreas de las relaciones sociales, genera un cúmulo de expectativas normativas. A continuación, me enfocaré en dos áreas de impacto que, por su peso en la estructura social, considero de vital importancia profundizar.

#### 2.2.2.1. EL IMPACTO DE LA IA EN EL TRABAJO Y EN EL EMPLEO

Los efectos del incremento (cualitativo y cuantitativo) de las TIC en el trabajo han sido objeto de múltiples investigaciones, sobre todo en el contexto más general del estudio de las revoluciones industriales<sup>(57)</sup>. En

(56) Corvalán, Juan G., *Perfiles digitales humanos. Proteger datos en la era de la inteligencia artificial. Retos y desafíos del tratamiento automatizado*, Thomson Reuters-La Ley, Buenos Aires, 2020, ps. 133-138.

Nota del director: asimismo, sobre principios y derechos de protección de datos personales ver Corvalán, Juan G. - Papini, Carina M., “Explorando los perfiles digitales humanos: cómo los algoritmos nos conocen más que nosotros mismos. Incluye el análisis del nuevo proyecto de Ley de protección de datos personales de Argentina”, tomo II del presente *Tratado*.

(57) Nota del director: muchos autores en esta obra han abordado los efectos de las nuevas tecnologías y la automatización en el mercado laboral. En primer lugar, estos se analizan en Etcheverry, Milagros, “El impacto de la inteligencia artificial en el mundo del trabajo”, tomo III del presente *Tratado*. Desde una óptica pesimista, se ha hecho hincapié en que los impactos de las innovaciones en el mercado laboral pueden resultar en una disminución de hasta un 47% de puestos de trabajo en las próximas décadas, examinando así las transformaciones que se producen en la estructura empresarial y en la organización del trabajo como la precarización y reducción de complejidad de las tareas, ver Soares Valentini, Rómulo, “Industria 4.0 y gestión laboral por inteligencia artificial. Las plataformas de trabajo y los nuevos desafíos del derecho laboral”, tomo III del presente *Tratado*. No obstante, se han manifestado opiniones positivas, postulando que el temor por la pérdida de empleos en manos de la inteligencia artificial resulta infundado. Particularmente se postula que la profesión de la abogacía “sobrevivirá” y seguirá siendo útil, sin perjuicio de que los abogados deban formarse y adaptarse al nuevo paradigma, ver Braun Pellegrini, Ezequiel, “Breve. Más tiempo para los abogados del futuro”, tomo III del presente *Tratado*. En el mismo sentido, se ha expresado que, aunque disminuya la demanda o relevancia del abogado en su ejercicio tradicional, no significa que haya menos puestos de trabajo para abogados en el futuro, ver Marsiglia, Gabriella - Renella, María Paula, “Inteligencia profesional: un diferencial frente a la IA”, tomo III del presente *Tratado*. Desde una perspectiva ética se ha postulado que se debe garantizar a los trabajadores que per-

lo que respecta a la 4RI, las indagaciones han estado encaminadas en discernir el grado de influencia de las tecnologías de la información, en especial, las derivadas de la IA (robótica, automatización)<sup>(58)</sup>. Y es que una de las zonas de impacto de la IA que despierta mayor preocupación a nivel de organismos internacionales y centros de estudios es la constituida por el mundo del trabajo.

La robotización y el aprendizaje automático (*machine learning*) en la industria muestran que las máquinas en muchas tareas o tipos de trabajo son considerablemente más eficientes y productivas que los seres humanos y amenazan con desplazarlos de sus puestos de empleo. La Organización Internacional del Trabajo (OIT) lanzó oficialmente el año 2015 la “Iniciativa del Centenario relativa al futuro del trabajo”<sup>(59)</sup>. En el “Informe inicial para la Comisión Mundial sobre el Futuro del Trabajo” se identifican cuatro grandes tendencias e implicaciones para el futuro del trabajo, a saber: tecnología, demografía, globalización y cambio climático. A propósito de la influencia de la tecnología, enfatiza el impacto transformador que esta proyecta sobre la naturaleza del trabajo, en particular por el aumento de los robots en la industria. Advierte el Informe que, sin bien es cierto que todos los cambios tecnológicos implican destrucción de empleos a corto plazo, la tendencia observada en el pasado indica que estos se recuperan y luego surgen nuevos y mejores empleos, cuestión que no se puede asegurar que acontezca en este nuevo ciclo de cambios, dada

dieron empleos ayudas no económicas como formación y acceso a derechos básicos, ver Martino, Antonio, “Ética y sistemas inteligentes”, tomo I del presente *Tratado*. Al examinar el rol de la IA en el ámbito de educación, se ha dicho que, lejos de promover el fin del trabajo humano, la IA complementará el trabajo del profesor en el aula, ver Schapira, Débora, “Inteligencia artificial: un nuevo desafío para la educación”, tomo III del presente *Tratado*. Finalmente, se ha evaluado el impacto en el mercado de trabajo no solo del cambio tecnológico, sino también del envejecimiento poblacional y la pandemia del COVID-19 desde una perspectiva del derecho tributario en Iriarte Yanicelli, Adolfo A. “Inteligencia artificial y derecho tributario”, tomo IV del presente *Tratado*.

(58) En los últimos años se han publicado varios estudios. Algunos de ellos: Acemoglu, Daron - Pascual Restrepo, *Robots and Jobs: evidence from US labor Market*, National Bureau of Economic Research, *working paper* 23285, Cambridge, MA, 2017; McKinsey Global Institute, *A future that Works: automation, employment, and productivity* y *Jobs lost, Jobs Gained: workforce transitions in a Time of Automation*, 2017; World Bank Group, *World Development Report 2016. Digital Dividends*, World Bank Group, Washington DC, 2016; World Economic Forum, *The Future of Jobs. Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution*, World Economic Forum, Geneva, 2016; Brynjolfsson, Erik - McAfee, Andrew, *The Second Machine Age*, W. W. Norton & Company, Inc., New York/London, 2014.

(59) Todos los documentos, informes y datos empíricos sobre el tema han sido obtenidos en el web de la OIT sobre la “Iniciativa sobre el Futuro del Trabajo”: <https://www.ilo.org/global/topics/future-of-work/lang-es/index.htm> (consultado el 14/09/2018).

la velocidad y avance de los cambios tecnológicos actuales, que no reconocen precedente en la historia de la humanidad<sup>(60)</sup>. Considerando, sin embargo, exageradas las hipótesis pesimistas, a juzgar por los índices de empleo oficiales, las disrupciones originadas por los cambios tecnológicos en el trabajo que le preocupan a la OIT son los denominados dividendos digitales o tecnológicos, es decir, como se distribuyen y reparten los beneficios del cambio que está promoviendo la irrupción de la tecnología (crecimiento, empleo, servicios).

Según la OIT y el Banco Mundial, la revolución tecnológica digital ha afectado negativamente el empleo, en el sentido que los mercados laborales se han polarizado y la desigualdad ha aumentado. Además, se suele mencionar en esta clase de análisis, que los efectos de la revolución tecnológica mundial todavía no se advierten en un aumento de la productividad; reactualizando con aquella constatación, el debate sobre la famosa paradoja de la productividad ya sea que se mida la productividad como productividad laboral o como productividad total de los factores<sup>(61)</sup>.

Sin perjuicio de que la preocupación de la coyuntura económica-social está signada por el tema de la distribución de los dividendos digitales, los estudios más autorizados en la materia coinciden en predecir que la 4RI tiene el potencial de producir la desaparición o el cambio radical de más de la mitad de los empleos actuales producto de la irrupción de los robots en el mundo del trabajo. Por consiguiente, si bien el problema de los dividendos digitales es la preocupación de la hora presente, no deben perderse de vista los inminentes riesgos y peligros que, para la existencia del trabajo, se instalan en el horizonte de la humanidad: un mundo donde los robots le quitan el puesto de trabajo a los humanos. Así, de la lucha de clases entre proletarios y capitalistas, pasaríamos a la lucha entre la clase de los trabajadores humanos y los robots que amenazan con hacerse cada vez más inteligentes.

En efecto, estamos viviendo en un mundo donde el empleo es un bien escaso y en donde, conseguir un *McDonalds's Job*, estadísticamente en los EE.UU., es más difícil que ser aceptado en Harvard<sup>(62)</sup>. Es el mundo de Google, gigante de la IA, que en 2012 generó US\$ 14.000 millones de utilidades con alrededor de 38.000 trabajadores versus General Mo-

(60) OIT, *Informe inicial para la Comisión Mundial sobre el Futuro del Trabajo*, OIT, Ginebra, 2017, ps. 10-11.

(61) Schwab, Klaus, *La cuarta revolución industrial*, 3a reimp. de la 1a ed. 2016, Random House Grupo Editorial, Barcelona, 2017, ps. 49-53; World Bank Group, *World Development Report 2016. Digital Dividends*, World Bank Group, Washington DC, 2016, p. 2.

(62) Ford, Martin, *El auge de los robots. La tecnología y la amenaza de un futuro sin empleo*, trad. Andrea Galvéz de Aguinaga y Víctor Cuchí Espada, Paidós, Barcelona, 2017, p. 29.

tors que, en su momento de mayor apogeo 1979, registró beneficios por US\$ 11.000 millones, ocupando 840.000 trabajadores (actualizado según inflación)<sup>(63)</sup>. Es también el mundo Amazon, que en marzo de 2012 compró por la suma de US\$ 775 millones, Kiva Systems, empresa de robótica gracias a la cual reforzó la logística robotizada de sus almacenes de distribución. La empresa absorbida ahora se llama Amazon Robotica<sup>(64)</sup> y gracias a ella, ahora Amazon cuenta con un disciplinado ejército de decenas miles de robots logísticos repartidos alrededor del mundo (que no están sindicalizados, no hacen huelgas, no se quejan, no demandan judicialmente la empresa, no cobran salario, no discuten con sus colegas o sus jefes, trabajan en silencio 24/7 y no sufren de estrés laboral, por cierto).

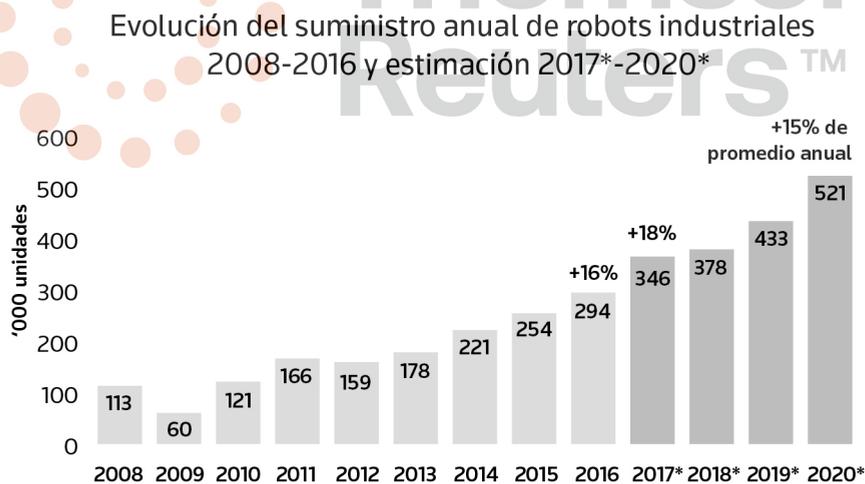
Organización	Estimaciones
Universidad de Oxford	El 47 por ciento de los trabajadores de los Estados Unidos de América corren el riesgo de que sus puestos de trabajo se automaticen.
PricewaterhouseCoopers	El 38 por ciento de los puestos de trabajo en los Estados Unidos de América, el 30 por ciento en el Reino Unido, el 21 por ciento en el Japón y el 35 por ciento en Alemania están amenazados por la automatización.
OIT (Chang y Huynh)	ASEAN-5: en los próximos veinte años, el 56 por ciento de los puestos de trabajo corren el riesgo de automatización.
McKinsey	El 60 por ciento de todas las ocupaciones comprenden al menos un 30 por ciento de actividades que técnicamente pueden automatizarse.
OCDE	Promedio de la OCDE: el 9 por ciento de los puestos de trabajo corren un alto riesgo. Hay un riesgo reducido de automatización completa, pero una gran proporción (entre el 50 por ciento y el 70 por ciento) de tareas con riesgo de automatizarse.
Roland Berger	Europa occidental: de aquí a 2035 se habrán perdido 8,3 millones de empleos en el sector industrial frente a 10 millones de puestos creados en el sector servicios.
Banco Mundial	Dos tercios de todos los puestos de trabajo de los países desarrollados pueden automatizarse.

Fuente: OIT (2017), *Informe inicial para la Comisión Mundial sobre el Futuro del Trabajo*, p. 29.

(63) *Ibidem*, ps. 79-80.

(64) Una buena forma de dimensionar cómo operan los robots en los centros de distribución de Amazon puede ser apreciada en <https://www.amazonrobotics.com/#/> (consultado el 02/12/2018).

En parte, las predicciones de reemplazado total o parcial de puestos de trabajo se debe al aumento sostenido de la venta de robots y sistemas de automatización. Se trata de una de las industrias más pujantes y dinámicas de la sociedad global. La Federación Internacional de la Robótica, en sus informes más recientes sobre el desarrollo de la industria, calcula que para el año 2020 el stock mundial de robots industriales operativos aumentará de las 1.828.000 unidades estimadas a finales de 2016 a las 3.053.000 unidades, lo que supone una tasa de crecimiento anual promedio del 14% entre 2018 y 2020<sup>(65)</sup>. Un incremento aún mayor, del orden de 20% al 25% en promedio por año, entre 2018-2020 (sobre las 400.000 mil unidades), se proyecta para los robots de servicios profesionales<sup>(66)</sup>. A este volumen de máquinas, debe sumarse el aumento exponencial de toda clase dispositivos de inteligencia artificial móviles que se usan cotidianamente tales como *smartphone* y derivados; videos juegos, *ebooks*, internet de las cosas, *eCommerce*, e-Educación, Bots, Robótica de la Nube, etc. Todas las proyecciones predicen un mundo futuro poblado de toda clase de sistemas derivados de las tecnologías de la información.



Fuente: International Federation of Robotics (IFR), *The Impact of Robots...*, cit.

La invasión de la robótica a la nación-trabajo genera ansiedad por el efecto de corto o mediano plazo conocido como desempleo tecnológico,

(65) International Federation of Robotics (IFR), *The Impact of Robots on Productivity, Employment and Jobs*, IFR, Frankfurt, 2017.

(66) Ídem.

en el sentido que Keynes daba a esa expresión, es decir, “el desempleo causado por los medios descubiertos para economizar mano de obra que ocurre mucho antes de que podamos encontrar nuevos usos para la misma”<sup>(67)</sup>.

El desempleo tecnológico que augura la invasión robótica no solo podrá afectar a las tareas rutinarias y que “involucran actividades físicas en entornos altamente estructurados y predecibles”, incluidas las vinculadas con la recopilación y procesamiento de datos, y que en los EE.UU. constituye el 51% de las actividades económicas, tales como manufactura, alojamiento, industria alimentaria, comercio a menudeo y empleos medianamente calificados, sino que además la oleada de automatización se extenderá hacia los puestos de trabajo “de cuello blanco”<sup>(68)</sup>.

Como se ha mostrado en las secciones precedentes, la IA de los sistemas expertos gana terreno todos días en los más diversos campos de la economía existentes, a la par que crea nuevas industrias, a saber, financiera, médica, comercio, transporte, comunicaciones, entretenimiento, por nombrar las más visibles al sencillo ciudadano de a pie. La IA cada día que pasa se hace más inteligente en términos tales que, incluso, “acabará produciendo el desmantelamiento de las profesiones tradicionales” como ha sido analizado por los investigadores del campo<sup>(69)</sup>.

Sin embargo, de acuerdo con los datos estadísticos comúnmente aceptados, no hay razones para ser tecno-pesimista, por cuanto el desempleo tecnológico suele ser de corto o mediano plazo. Pero los cambios en las profesiones y en el trabajo que la 4RI está provocando son tan profundos que, incluso la propia OIT, a pesar de no tener una visión tecno-pesimista, hace un llamado de alerta en cuanto a que, esta vez, las cosas pueden ser diferentes, sobre todo por la polarización del trabajo y la distribución desigual de los dividendos digitales que empieza a observarse<sup>(70)</sup>.

---

(67) Keynes, John Maynard, “Economic possibilities for our Grandchildren”, en *Essays in Persuasion*, Harcourt Brace, New York, 1930, ps. 358-373. [https://assets.aspeninstitute.org/content/uploads/files/content/upload/Intro\\_and\\_Section\\_I.pdf](https://assets.aspeninstitute.org/content/uploads/files/content/upload/Intro_and_Section_I.pdf) (consultado el 16/09/2018).

(68) McKinsey Global Institute, *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation*, McKinsey Global Institute, December 2017.

(69) Susskind, Richard - Susskind, Daniel, *El futuro de las profesiones*, trad. Juan Carlos Ruiz Franco, Teell, España, 2016, p. 2.

(70) Se muestra en este estudio que “la desigualdad del ingreso está aumentando al mismo tiempo que van disminuyendo los costos del almacenamiento de los macrodatos” y se proyecta un mayor riesgo de polarización del empleo en los próximos años. Y añade: “A medida que se destruyen puestos de trabajo en el sector manufacturero y parte del sector de los servicios, ha aumentado el empleo en las ocupaciones

Se constata también una atmósfera de pérdida del sentido del trabajo, pues la retribución no resulta en ningún caso satisfactoria. Y dado que el trabajo es un bien escaso, las personas del extremo pobre del polo laboral (que son la mayoría), se autoperciben como prisioneras de un sistema injusto que no les proporciona bienestar. Así las cosas, la OIT hace un llamado a brindar una protección social efectiva para el siglo XXI, haciéndose parte del debate sobre la conveniencia de estatuir una renta básica universal<sup>(71)</sup>.

Quizás todo lo anteriormente expuesto sobre el impacto de la IA en el mundo del trabajo pueda resumirse en un muy citado y apócrifo intercambio verbal, supuestamente sostenido entre Walther Reuther, el legendario líder de la *United Automobile Workers* y Henry Ford II, cuando visitaban una planta automatizada. En eso estaban cuando de pronto, Ford, distendidamente le dice a Reuther: “Walter, ¿cómo harás para que los robots paguen las cuotas sindicales?, a lo que Reuther habría contestado: Henry, ¿cómo vas a hacer para que ellos compren tus autos?”<sup>(72)</sup>.

de baja y alta calificación. Los estudios sobre la robotización muestran que el desplazamiento es alto en lo que atañe a las tareas rutinarias (es decir, tareas que pueden realizar fácilmente robots informatizados), incluso en muchos segmentos del sector de los servicios donde la digitalización y la inteligencia artificial han llegado a desempeñar un papel más importante. Al no haber políticas de transición efectivas que incluyan la oferta de oportunidades adecuadas para adquirir nuevas competencias laborales pertinentes (véase la nota informativa núm. 8), muchas de las personas que corren el riesgo de perder su empleo pueden verse obligadas a aceptar trabajos menos calificados y peor remunerados, intensificando así la presión que se ejerce sobre los salarios en el segmento peor remunerado (Dauth *et al.*, 2017). De hecho, la mayoría de los trabajos rutinarios de calificación media que estaban comprendidos en los contratos de trabajo estándar con horarios de trabajo regulares han sido reemplazados por formas de empleo atípicas, en ocupaciones no rutinarias tanto de carácter intelectual como manual (OCDE, 2015)”; OIT (2018). La influencia de la tecnología en la calidad y la cantidad de empleo. Nota informativa preparada para la segunda reunión de la Comisión Mundial sobre el Futuro del Trabajo 15-17 de febrero de 2018, Grupo 3: Tecnología al servicio del desarrollo social, ambiental y económico. Ginebra, Comisión Mundial sobre el Futuro del Trabajo, p. 4.

(71) OIT, *Informe inicial para la Comisión Mundial sobre el Futuro del Trabajo*, Ginebra, 2017, p. 24.

(72) El supuesto diálogo entre Reuther y Ford II es citado por diversas fuentes, entre otras: Ford, Martin, *El auge de los robots. La tecnología y la amenaza de un futuro sin empleo*, trad. Andrea Galvéz de Aguinaga y Víctor Cuchí Espada, Paidós, Barcelona, 2017, p. 185; *The Economist*, <https://www.economist.com/babbage/2011/11/04/difference-engine-luddite-legacy> (consultado el 02/10/2018); Huffpost, [https://www.huffingtonpost.com/george-zarkadakis/will-a-robot-replace-you\\_b\\_8506324.html](https://www.huffingtonpost.com/george-zarkadakis/will-a-robot-replace-you_b_8506324.html) (consultado el 02/10/2018).

### 2.2.2.2. EL IMPACTO EN EL CURSO EVOLUTIVO DEL *HOMO SAPIENS SAPIENS* Y LA HUMANIDAD

El trabajo, desde sus orígenes, es la base organizativa de todo orden social, de modo tal que, cualquier cambio significativo en el orden laboral, probablemente generará —y la historia creemos que así lo ha demostrado— una dinámica de cambios y modificaciones en la estructura global de la existencia humana<sup>(73)</sup>. Se supone que la tecnología produce, al menos en el mediano plazo, un aumento en la productividad de los trabajadores. Pero cuando la tecnología alcanza niveles de perfeccionamiento tan altos, como los que se consiguen con la IA y la robótica, el trabajo humano tiende a ser reemplazado, y se difumina la frontera entre capital y trabajo, ya que las máquinas mismas, se transforman en *trabajadores*, pues hacen el trabajo.

La observación del paisaje tecnológico actual arroja indicios fuertes de que la tendencia actual va en esa dirección, impulsada por el desarrollo exponencial de las TICs que permiten profundizar los sistemas de IA, ya que aumentan el poder de cómputo, los cuales, a su vez, permiten desarrollar aún más las TICs pues el mayor poder de cómputo de los sistemas de IA contribuye a resolver los problemas que el desarrollo de las TICs van planteando en su avance. Se trata de la inteligencia trabajando sobre la inteligencia en el contexto de una sociedad cuya economía, además, está estructurándose sobre la base de ese proceso de retroalimentación.

El trabajo, decía, ha cambiado al mundo porque el trabajo es la base organizativa de todo orden social, de modo tal que, cualquier cambio significativo en él, impone una dinámica de cambios y modificaciones en la estructura global de la existencia humana. La hipótesis que seguimos y hemos desarrollado a lo largo de estas páginas, dice que los cambios que el trabajo está experimentando y experimentará en el futuro, en virtud de la irrupción de las tecnologías NBIC, afectarán la sociedad completa e, incluso, hasta podrían generar un cambio evolutivo en el *homo sapiens*, pues las tecnologías NBIC, inauguran técnicamente esa posibilidad. La conciencia de estar en tránsito hacia un nuevo estadio de desarrollo civilizacional y evolutivo, asimismo, comienza a organizarse ideológicamen-

---

(73) Sobre la fundamentación de la relación entre trabajo y sociedad, y cómo el trabajo influye en la formación de la sociedad, en términos tales de ser estructurante, existe una extensa bibliografía que hemos recogido y analizado en dos investigaciones presentadas con anterioridad, y a las cuales ya nos hemos remitimos: López Oneto, Marcos, “Las metamorfosis del trabajo: del trabajo al empleo (notas de lectura para una reconstrucción histórica-conceptual del trabajo)”, *Revista Chilena de Derecho del Trabajo y de la Seguridad Social*, vol. 5, nro. 9, 2014, ps. 141-166; *El principio de protección a la fuente del empleo en Chile. Pasos hacia un derecho del trabajo bidireccional*, Thomson Reuters, Santiago, 2013.

te en función del fenómeno de la convergencia tecnológica expresado en las ciencias NBIC, en donde la IA (*the key element*) y la nanotecnología son los elementos catalizadores.

En efecto, tanto el gobierno de los Estados Unidos de Norteamérica como la Unión Europea, han advertido la importancia del tema, otorgándole legitimidad política social a las preocupaciones futuristas sobre el potencial transformador de la naturaleza humana que puede desencadenarse con el despliegue sin control de las potencialidades de las ciencias NBIC.

El comienzo oficial de este desarrollo político, como ya se ha señalado, suele ser focalizado en reporte de la *National Science Foundation (NSF)*, titulado “Converging Technologies for Improving Human Performance. Nanotechnology, Biotechnology, Information Technology and Cognitive Science” (2002)<sup>(74)</sup>, donde se analiza en detalle los avances tecnológicos que tienen el potencial de impactar en el curso evolutivo del *homo sapiens sapiens*, como son la *improving human performance* y los *human enhancement*, que consisten en el aumento tecnológico de las capacidades humanas y la modificación de la corporalidad y el intelecto humano. La Unión Europea (a través de la Comisión Europea), por su parte, patrocinó el proyecto CONTECS que dio como fruto el reporte “Converging Technologies and their impact on the Social Sciences and Humanities (CONTECS). An analysis of critical issues and suggestion for a future research agenda, Final Report, May 2008”<sup>(75)</sup>. En dicho reporte, al igual que en el de la NSF, juntamente con hacer una disensión técnica a las NBIC, se alzan las voces en torno a los problemas éticos que el prometido mundo futuro de la IA y las NBIC están generando.

La IA contribuye a que los proyectos *improving human performance* y los *human enhancement* avancen y profundicen sus posibilidades transformadoras de la naturaleza humana. La ingeniería genética no sería posible sin el poder de cómputo actual que permite la ejecución de complejos algoritmos de búsqueda, conexión, ordenación, inferencia y desciframiento (secuenciación genómica)<sup>(76)</sup>. El aumento de poder de la

(74) Roco, Mihail - Sims Bainbridge, William (ed.), *Converging Technologies for Improving Human Performance. Nanotechnology, Biotechnology, Information Technology and Science*, National Science Foundation, Washington DC, 2003.

(75) Andler, Daniel *et al.*, *Converging Technologies and their impact on the Social Sciences and Humanities (CONTECS). An analysis of critical issues and a suggestion for a future research agenda. Final report*, Fraunhofer Institute for Systems and Innovation, Germany, 2008.

(76) La edición genética ha dado un transcendental paso con la invención de la técnica CRISPR-Cas 9 que permite editar el genoma en la línea germinal, fue desarrollado en el año 2012 por Emmanuelle Charpentier y Jennifer Doudna (pero debe su

IA, por tanto, repercute en el desarrollo de la ingeniería genética (biología computacional). El perfeccionamiento de la ingeniería genética llevada a sus límites, como sería la edición del genoma en la línea germinal, crea la posibilidad no solo de la transgénesis en el *homo sapiens*, sino que directamente el cambio artificial de su curso evolutivo. Si a esta hipótesis de hibridación, le agregamos el componente *cyborgs*, esto es, de la simbiosis ser humano-máquina, por obra de la IA y la robótica<sup>(77)</sup>, lo que está en juego, entonces, ya no es la simple mejora del ser humano, sino definitivamente el fin de la especie humana.

El viejo ideal humanista de mejoramiento humano mediante la educación, así, comienza a quedar obsoleto con la mejora técnica del ser humano que incluiría la ampliación de los sentidos, la memoria, la inteligencia y el control del envejecimiento; cambiando nuestra conceptualización de la dignidad humana, fundamento clásico y último de todo derecho humano. En concreto, la posibilidad de cambio civilizacional que las tecnologías del mejoramiento humano prometen, a saber, un mundo posthumano y regido por una lógica reproductiva de hibridación ontológica permanente, ya no pertenece solo al terreno de la ciencia ficción, sino que es un asunto de Estado, que mantiene en alerta a los gobiernos del mundo, quienes en la materia comienzan a actuar conforme al principio de precaución.

El impacto de la IA en la sociedad genera ansiedad e inquietud. Al tener la humanidad hoy en sus manos, por primera vez en la historia, el poder de cambiar el curso evolutivo de la especie humana, las preguntas que antaño se formulaban sobre la naturaleza humana, adquieren

---

origen al microbiólogo español Francis Mojica), las cuales obtuvieron por tal aporte el premio Nobel de Química del año 2020. El método CRISPR-Cas 9 no es por cierto la única tecnología de edición genómica, pero sí, según los expertos, es la más sencilla, económica, eficiente y precisa, lo cual contribuye a una expansión del universo de aplicación y desarrollo de esta técnica. Estas últimas dos características, precisión y eficiencia, han llevado al convencimiento científico que el método se podría aplicar directamente “al embrión humano para modificar los genes en algunas o en todas las células del mismo; si se permite que el embrión se desarrolle, entonces el ser humano resultante tendrá esas modificaciones (y así su descendencia)” (Chan, Sarah, “Edición genética y desafíos en la regulación internacional: lecciones desde México”, en Capdeville, Pauline - Medina Arellano, María de Jesús [coords.], *Bioética Laica. Vida, muerte, género, reproducción y familia*, UNAM Instituto de Investigaciones Jurídicas, México, 2018, p. 65). La técnica CRISPR-Cas 9 inaugura pues la factibilidad técnica de la modificación genética de la especie humana.

(77) Nota del director: sobre el concepto de robótica y su relación con el campo de la inteligencia artificial ver Corvalán, Juan G. - Danesi, Cecilia, Carro María Victoria “Responsabilidad civil de la inteligencia artificial”, tomo II del presente *Tratado*. Sobre los *cyborgs* que ya existen actualmente ver Gil Domínguez, Andrés, “Transhumanismo y derechos ¿humanos?”, tomo I del presente *Tratado*.

otro significado, pues ahora existen las herramientas tecnológicas para pasar de la potencia al acto transformador: al acto fáustico (por sobre el prometeico)<sup>(78)</sup>. Acto fáustico que, como veremos más adelante, está siendo potenciado por el poderoso e influyente movimiento transhumanista.

### III. FUNDAMENTOS ANTROPOLÓGICOS, ÉTICOS Y FILOSÓFICOS PARA UN DERECHO DE LA INTELIGENCIA ARTIFICIAL (DIA)

La potencialidad de la IA de conducir a la humanidad hacia una era posthumana es sostenida antropológica, ética, ideológica y filosóficamente por el movimiento transhumanista. Intentar una conceptualización de aquel movimiento futurista, para luego confrontarlo críticamente, es la estrategia de argumentación de los fundamentos antropológicos, éticos y filosóficos del DIA por la que he optado, pues nos permite una argumentación no *in abstracto*, sino que conectada fuertemente con la realidad contemporánea de la humanidad (cultural e histórica).

#### 3.1. APROXIMACIÓN CONCEPTUAL AL TRANSHUMANISMO

En este artículo se conceptualiza al transhumanismo como un movimiento cultural que, comprendiendo al *homo sapiens* como una etapa más de la evolución y no la última, afirma la posibilidad, necesidad y conveniencia de tomar, mediante el uso de la tecnología (NBIC), el control de la evolución de la especie humana; para impulsar así su mejoramiento actual, a fin de eliminar el envejecimiento y mejorar en gran medida las capacidades humanas intelectuales, físicas y psicológicas, asumiendo explícitamente la posibilidad de producir, mediante el uso de la tecnología, una superación geno-fenotípica de aquel, representada icónicamente en la figura del posthumano y de la posthumanidad.

La conceptualización que los propios transhumanistas hacen de su movimiento sirve para justificar la propuesta:

“El transhumanismo es una forma de pensar sobre el futuro que se basa en la premisa de que la especie humana en su forma actual no representa el final de nuestro desarrollo, sino una fase comparativamente temprana (...)El transhumanismo es una clase de filosofías de la vida que

(78) La hipótesis del fin de la humanidad y el advenimiento de la posthumanidad que es expresamente defendida por el movimiento transhumanista y predicha para el año 2045 por uno de sus más prominentes intelectuales, Raymond Kurzweil, también ha sido tratada como tendencia histórica por Harari, véase. Harari, Yuval Noah, *De animales a dioses. Breve historia de la humanidad*, trad. de Joandomènec Ros, Debate, Buenos Aires, 2014; *Homo Deus. Breve historia del mañana*, trad. Joandomènec Ros, Debate, Barcelona, 2016.

busca la continuación y la aceleración de la evolución de la vida inteligente más allá de su forma humana actual y sus limitaciones humanas por medio de la ciencia y la tecnología, guiados por principios y valores que promueven la vida”<sup>(79)</sup>.

No en vano, uno de los fundadores de la filosofía transhumanista, Max More<sup>(80)</sup>, señala que el transhumanista aborda los problemas morales y éticos, sobre la base de una norma universal anclada no en la pertenencia a la especie humana, sino en las cualidades de cada Ser. Así es como postula que criaturas con niveles similares de sapiencia, inteligencia y experimentación del mundo, gozan de un *status* similar sin importar si son seres humanos, animales, ciborgs, máquinas inteligentes o extraterrestros de nuestro mundo. En esa misma línea, Martina Rothblatt, para quien la mente es el Ser, defiende la libertad de las formas para los seres, porque la tecnología de *software* del siglo XXI permitirá separar nuestras mentes de nuestros cuerpos, pasando de ser mentes basadas en carne (*flesh*) a mentes basadas en *software* (silicio). Cuando ello efectivamente acontezca y sea una práctica, será como cambiar de género de hombre a mujer o elegir ser hombre o mujer. Transexuales, así, escogeríamos una nueva forma en *software*, aunque seguiríamos siendo la misma persona. Advenirla una era de transgenerismo, ya que podríamos variar cuanto quisiéramos nuestras formas sin perder la identidad de nuestro Ser. Todos estos hechos proporcionarían a los sociobiólogos evidencias del surgimiento de una nueva especie:

“Como nuestra creatividad ha florecido, hemos madurado del *homo sapiens a persona creatus* (...) Un concepto transhumanista básico es que un humano no necesita tener un cuerpo de carne, así como una mujer no necesita tener una vagina real (...) La humanidad está en la mente, al igual que la identidad sexual. A medida que el *software* se vuelve cada vez más capaz de pensar, actuar y sentirse como un humano, debe ser tratado como un compañero humano, y bienvenido como miembro de la especie tecnológica *persona creatus*”<sup>(81)</sup>.

Como se advierte, el proyecto ideológico del transhumanismo radical se traduce en la consecución de un objetivo distópico, porque plantea el surgimiento de un mundo futuro donde el hombre y la humanidad se-

---

(79) Véase el sitio web de *Humanity+*, la organización rectora del transhumanismo a nivel planetario <https://humanityplus.org/philosophy/transhumanist-faq/> (consultado el 08/11/2018).

(80) More, Max, “The Philosophy of Transhumanism”, en More, Max - Vita More, Natasha (ed.), *The Transhumanist reader*, Wiley-Blackwell, UK, 2013, ps. 13-14.

(81) Rothblatt, Martine, “Mind is deeper than matter. Transgenderism, transhumanism, and the freedom of form”, en More, Max - Vita More, Natasha (ed.), *The Transhumanist reader*, Wiley-Blackwell, UK, 2013, p. 13.

rán el pasado de otras civilizaciones, por de pronto, las que emerjan de la especie *persona creatus* o de las otras que aparezcan en el Edén artificial transhumanista habitado por seres inmortales, mentes digitales y virtuales, que se reproducen a sí mismas: hermafroditas cibernéticos.

Pues bien, conceptualizado ya, asimismo, el transhumanismo podría, idealmente, caracterizarse del siguiente modo<sup>(82)</sup>:

- a) Como una eugenesia de nuevo cuño, con pretensiones éticas, que quiere pasar “del azar a la elección” (“From chance to choice”). De la lotería genética al supermercado genético y de la ortopedia y del *cyborg*<sup>(83)</sup>, al posthumano. No se trata de una eugenesia estatal democrática o bien de una totalitaria-exterminadora (nazi), sino que de una eugenesia privada; liberal diría Habermas<sup>(84)</sup>, cuyos pilares son la libertad personal, la igualdad (reparar las injusticias de la naturaleza), democrática (igualdad genética) y promocional, pues a diferencia de la exterminadora nazi, no buscaría la eliminación de los débiles sino repararlos y aumentar sus capacidades.
- b) Como desacralizador de la noción de naturaleza humana, en el sentido de que esta no es intocable, y que por tanto no existen impedimentos éticos para modificarla, mejorarla y aumentarla. Recordemos que para Harris<sup>(85)</sup> es una obligación moral mejorar a los hijos (principio de beneficencia procreativa). Si el humanismo es antropocéntrico, diríamos, que el transhumanismo es antropotécnico y antroprófugo, es decir, mediante la técnica intenta escapar de la condición humana en busca de la posthumanidad.
- c) Como un movimiento cultural que busca la inmortalidad representada en la figura del posthumano. Se trata de la profecía de la muerte de la muerte de la que hablan Cordeiro<sup>(86)</sup> y Laurent

(82) En una alguna medida hemos seguido la caracterización ideal que hace Ferry, Luc, *La revolución transhumanista. Cómo la tecnomedicina y la uberización del mundo van a transformar nuestras vidas*, trad. de Alicia Martorell Linares (original en francés: *La révolution transhumaniste*), Alianza Editorial, Madrid, 2017, ps. 50-74.

(83) Nota del director: sobre los *cyborgs* que ya existen actualmente ver Gil Domínguez, Andrés, “Transhumanismo y derechos ¿humanos?”, tomo I del presente *Tratado*.

(84) Habermas, Jürgen, *El futuro de la naturaleza humana. ¿Hacia una eugenesia liberal?*, trad. de R. S. Carbó, Paidós, Barcelona, 2017.

(85) Harris, John, “Los mejoramientos son una obligación moral”, en Bostrom, Nick - Savulescu, Julian, *Mejoramiento humano*, trad. por Inés Ramia y Alicia Jiménez, Tell Editorial, España, 2017; *Enhancing Evolution. The Ethical Case for Making Better People*, Princeton and Oxford, Princeton University Press, 2010, e-book.

(86) Cordeiro, José Lluís - Wood, David, *La muerte de la muerte. La posibilidad científica de la inmortalidad física y su defensa mortal*, Deusto, Barcelona, 2018.

Alexandre<sup>(87)</sup>, que amalgaman en función de los futuros avances en genómica, secuenciación del ADN, terapias genéticas, nanomedicina, investigación en células madres, bio-*big data* y la robótica.

- d) Como un movimiento cultural tecno-optimista y progresista, es decir, contrario a todas las formas de pesimismo que conducen al “bioconservadurismo”; contra las ideologías del declive y de la vuelta a la edad de oro, el transhumanismo reivindica una fe en el progreso que se puede comparar con la que animaba a los filósofos y científicos de la ilustración<sup>(88)</sup>. Como ha dicho Eric Schmidt, ex Presidente Ejecutivo de Google y uno de los ejecutivos más prestigiados del mundo de la informática y la IA: “Si lo hacemos bien creo que podremos reparar todos los problemas del mundo”<sup>(89)</sup>.
- e) Como un movimiento cultural antroprófugo, es decir, un movimiento cultural que propicia la huida del *homo sapiens* de su condición humana hacia una posthumana. Así, por ejemplo, la supresión de la frontera entre el ser humano y la máquina (entre lo real y lo virtual) es vista como una forma de liberación, lo mismo que el aumento de las capacidades humanas y la incorporación de otras, no propiamente genotípicas o fenotípicas. Para el transhumanismo la biología es, al mismo tiempo, condición de existencia y esclavitud de los seres humanos. La desaparición biológica de la especie humana no es vista con una pérdida, en la medida que las máquinas o las nuevas creaturas (“los hijos mentales” de que habla Movarec) preserven su cultura<sup>(90)</sup>. Por lo anteriormente expuesto y como lo ha justificado Diéguez, el transhumanismo propone un proyecto “de liquidación del ser humano (porque en esto consiste al final, por mucho que se lo intente pasar por una redención)”<sup>(91)</sup>. O, y como ha escrito Ferry refiriéndose al transhumanismo cibernético: “La posthumanidad no tendrá prácticamente nada de humano, pues ya no estará

---

(87) Alexandre, Laurent, *La mort de la mort. Comment la technomédecine va bouleverser l'humanite*, Jclattès, Paris, 2011.

(88) Ferry, Luc, *La revolución transhumanista. Cómo la tecnomedicina y la uberización del mundo van a transformar nuestras vidas*, trad. de Alicia Martorell Linares, Alianza Editorial, Madrid, 2017, p. 62.

(89) Ídem.

(90) Diéguez, Antonio, *Transhumanismo. La búsqueda tecnológica del mejoramiento humano*, Herder, Barcelona, 2017, p. 67.

(91) *Ibíd.*, p. 6.

arraigada en lo vivo, la lógica de las nuevas tecnologías es fundamentalmente la de la desmaterialización”<sup>(92)</sup>.

Más allá de la caracterización realizada, es importante consignar también que el estructuralismo francés en general y la obra de Michel Foucault, en particular, podrían ser considerados como un importante antecedente cultural del transhumanismo. Ya en su clásico “Las palabras y las cosas”, publicada en 1966, Foucault, crea las bases teóricas para la posterior crítica postmoderna del humanismo. Es ahí donde se defiende una concepción del hombre y de lo humano como algo contingente, no comprensible como categoría natural y universal; no esencialista, sino como fruto de una construcción cultural. Se trata de la negación del hombre en tanto portador de una esencia definitiva. También debe mencionarse que, de un modo directo, y siempre en el plano filosófico, sin duda que el pensamiento de Peter Sloterdijk, con su tesis del fracaso del humanismo, también ha contribuido a darle forma y peso al pensamiento transhumanista.

La conceptualización y caracterización del transhumanismo que se ha propuesto es bastante genérica y podría ser aceptada por un muy amplio y diverso auditorio de observadores, pues se ha construido teóricamente con elementos autodescriptivos de los propios Transhumanistas (Declaración y Transhumanist FAQ's), así como de miradas críticas o parciales, provenientes de diversas fuentes. A partir esa idea general que sobre el transhumanismo se ha construido, presento ahora una mirada más puntual, perfilando una concepción.

Es importante consignar que, el estructuralismo francés en general y la obra de Michel Foucault, en particular, podrían ser consideradas como un importante antecedente cultural del transhumanismo. Ya en su clásico “Las palabras y las cosas”, publicada en 1966, Foucault, crea las bases teóricas para la posterior crítica postmoderna del humanismo. Es ahí donde se defiende una concepción del hombre y de lo humano como algo contingente, no comprensible como categoría natural y universal; no esencialista, sino como fruto de una construcción cultural. Se trata de la negación del hombre en tanto portador de una esencia definitiva<sup>(93)</sup>.

---

(92) Ferry, Luc, *La revolución transhumanista. Cómo la tecnomedicina y la uberización del mundo van a transformar nuestras vidas*, trad. de Alicia Martorell Linares, Alianza Editorial, Madrid, 2017, p. 47.

(93) Foucault, Michel, *Las palabras y las cosas: una arqueología de las ciencias humanas*, trad. Elsa Frost, 2a ed., Siglo XXI, Buenos Aires, 2008.

Y de un modo directo, siempre en el plano filosófico, sin duda que el pensamiento de Peter Sloterdijk, con su tesis del fracaso del humanismo, también ha contribuido a darle forma y peso<sup>(94)</sup>.

### 3.1.1. EL TRANSHUMANISMO COMO UN NEO-GNOSTICISMO FUNDADO EN EL TECNOPTIMISMO

Judeo-cristiana y míticamente, sabemos que, por haber comido del fruto prohibido, la mortalidad fue el castigo que impuso Dios a Adán y Eva, y a toda su descendencia (la humanidad). La expulsión del Jardín del Edén marca el inicio de la humanidad cuya evolución será propulsada por la combustión que genera la contradicción vida/muerte. Quizás no exista un mito tan poderoso en el inconsciente colectivo de la civilización judeocristiana, que el mito de la pérdida del paraíso y el inicio de la humanidad como acto de desobediencia y, al mismo tiempo, de creación de la libertad. La ganancia de la libertad se obtuvo a costa de la pérdida del paraíso.

Pues bien, desde aquellos tiempos bíblicos hasta ahora, la recuperación del paraíso y de la inmortalidad ha sido la promesa que las religiones han ofrecido a sus creyentes. Salvación, vida eterna, reencarnación, nirvana, lo que fuere; las religiones han sabido canalizar el sentimiento de separatividad de la gente; de soledad esencial que la ilusión del amor por veces las hace olvidar al sentirse unificados e integrados al cosmos.

Pero la humanidad, a cada tanto, entra en crisis. Ahora, bajo el acecho del fin de los discursos de la modernidad, con la denominada postmodernidad y el fin de los metarrelatos, enfrentamos una crisis de sentido e identificación social<sup>(95)</sup>. Cualquiera que sea la construcción teórica que se intente hacer para interpretar el momento histórico que estamos viviendo, esta debe partir del hecho siguiente: el derrumbe de los sueños, de las utopías; el derrumbe de los muros de sentido que estructuraban la sociedad; su derrumbe por la fuerza irresistible del estrepitoso fracaso de los discursos de la modernidad, han permitido que la sociedad comience a buscar la salvación ya no exclusivamente en la religión, en las ideologías o utopías socio-políticas, sino que en otros discursos (ecologismo, feminismo, género, animalismo, *new age*, etc.).

---

(94) Presentada originalmente en su clásico *Normas para el parque humano. Una respuesta a la Carta sobre el humanismo de Heidegger*, trad. Teresa Rocha Barco, Siruela, Madrid.

(95) Para una revisión conceptual de la categoría “postmodernidad”, sin duda, Lyotard (2006) y para su crítica, Habermas, Jürgen, *El discurso filosófico de la modernidad*, Katz Editores, Buenos Aires, 2010.

La hipótesis que propongo dice que el transhumanismo es uno de esos nuevos discursos que emergen en esos momentos de transición y de cambio histórico. El transhumanismo es fruto del proceso de desencanto ideológico y religioso por el que atraviesa la humanidad, prometiendo la salvación no por la fe en el Dios judeocristiano o en cualquier otro Dios; tampoco por la promoción de una nueva utopía política social, sino que por la fe en el progresivo desarrollo tecnológico.

Frente a la desolación producida por el paraíso ya perdido completamente, y por la obsesiva y frustrante búsqueda de la inmortalidad que nunca llegó a pesar de los rezos, las penitencias, y también de los diezmos; el transhumanismo promete que la salvación laica advendrá por mor de la tecnología, del desarrollo exponencial o muy acelerado de las ciencias NBIC. El transhumanismo augura nuevos paraísos; paraísos artificiales habitados por “humanos” diseñados genéticamente con “cuerpos sin edad y almas llenas de alegría”<sup>(96)</sup>; por una pléyade de seres híbridos; seres digitales; todo en una “naturaleza” también rediseñada. Lo que Dios nos quitó, la inmortalidad, nos será devuelto por la *Gnosis*. Seremos capaces de escapar de la muerte; si queremos, incluso, podremos hasta prescindir del cuerpo; seremos capaces de evadir el castigo divino, por obra y gracia del cultivo de la *Gnosis*.

No es primera vez en la historia de la humanidad que se busca la salvación y liberación del alma, la inmortalidad, mediante la *Gnosis*. En efecto, fueron los gnósticos, esa heterodoxa corriente del cristianismo primitivo, declarada herética por Inereo de Lyon en el año 180 d. C., la que postulaba que el hombre no se salva por la fe en Cristo, sino que se salva mediante la *gnosis* o conocimiento introspectivo de lo divino, que es un conocimiento superior a la fe. La salvación se consigue mediante el conocimiento de uno mismo comprendido —el uno mismo— como chispa divina encerrada en el cuerpo, en la materia.

Para los gnósticos el ser humano es autónomo para salvarse a sí mismo<sup>(97)</sup>. Así, gnóstico es el hombre que se ha iniciado en el conocimiento de sí mismo como método de salvación, acaso como escape de este mundo fútil e incensario. Y es que los gnósticos, observa Pastorino, “creían que el mundo material en el que vivimos es una catástrofe cósmica y que, de alguna manera, chispas de la divinidad han caído, han quedado atrapa-

(96) “Ageless bodies and Happy Souls” como se expresa en The President’s Council on Bioethics, *Beyond Therapy. Biotechnology and the Pursuit of Happiness*, Washington, DC, 2003, disponible en <https://biotech.law.lsu.edu/research/psc/reports/beyondtherapy/> (consultado el 03/03/2021).

(97) Sobre el gnosticismo, hemos tenido a la vista la traducción César Vidal de los Evangelios Gnósticos, véase Vidal, César, *Los evangelios gnósticos*, Edaf, Madrid, 2008.

das en la materia, y necesitan escapar y volver a su origen. El escape de la materia lo logran cuando adquieren conciencia cabal de su situación y de su origen divino. Este conocimiento es la *Gnosis*"<sup>(98)</sup>.

Para el transhumanismo el hombre es una etapa, apenas el comienzo de la evolución de la inteligencia. Para ellos el ser humano es extremadamente débil y de vida muy corta, comparado con otras especies vivas. Su inteligencia es su don. Y ese don encierra el misterio de su permanente crecimiento, desarrollo y evolución mediante el uso del mismo. La inteligencia trabajando sobre la inteligencia aumenta la inteligencia y nos conducirá hacia la *Gnosis*, que en la lógica del transhumanismo sería algo así como el descubrimiento o creación del algoritmo de la vida, de la organización de la materia y del universo.

El transhumanismo propone pasar de la lotería genética al supermercado genético; de la inteligencia biológica a la inteligencia artificial fuerte; luego a la singularidad y de ahí, a la mismísima eternidad. Su programa consiste, pues, no solo en tomar el control de la evolución mediante la ciencia y la tecnología, sino que expandir la inteligencia más allá de las fronteras del planeta tierra.

Por todo lo anteriormente expuesto, en este estudio se conceptualiza al transhumanismo como una especie de *neo-gnosticismo*, por ubicar como motor de su proyecto de mejoramiento y superación del *homo sapiens*, a la ciencia: la inteligencia, la *Gnosis*.

Nos acompaña en esta tesis Reichmann, quien distingue los inicios de esta ciencia fáustica, en la *Nueva Atlántida* de Francis Bacon, donde se anticipa la biología sintética y la ingeniería genética; para luego distinguirla con nitidez en el siglo XX en los trabajos de J. D. Bernal y verla reaparecer, en gloria y majestad en el siglo XXI bajo el poder de las ciencias NBIC y las tesis de la singularidad, con un Raymond Kurzweil a la cabeza, preconizando que el aumento explosivo de la inteligencia permitirá a los seres humanos emanciparse del mundo material y dejar de ser organismos biológicos. Así es como Reichmann concluye:

“El transhumanismo es un nuevo gnosticismo —que se sitúa de lleno en la antiquísima tradición (órfica y gnóstica) del soma/sema, el cuerpo como tumba y cárcel—, y a mi entender se trata de una de las figuras más amenazantes del nihilismo para el siglo XXI”<sup>(99)</sup>.

---

(98) Pastorini, Miguel, “Los autores de los evangelios gnósticos ¿quiénes eran y qué creían?”, *Aleteia*, Aclaraciones Doctrinales de los Manuscritos de Nag Hammadi, 2017.

(99) Reichmann, Jorge, “¿Triunfara el nuevo gnosticismo? Notas sobre la biología sintética, nanotecnologías y manipulación genética en el Siglo de la Gran Prueba”, *ISEGORIA Revista de Filosofía Moral y Política*, (55), 2016, 409-441.

Para Reichmann, la pretensión de abolir la imperfección humana transforma inmediatamente el proyecto transhumanista en nihilista. Toma esta idea principalmente de Hans Jonas, quien sugería que el nihilismo moderno tendría su origen en la revolución científica, que al erradicar las causas finales contribuyó a despojar de cualquier sentido intrínseco al universo.

### 3.2. ÉTICA PARA UN DERECHO DE LA INTELIGENCIA ARTIFICIAL (DIA): RESPONSABILIDAD Y PRECAUCIÓN

Durante el curso del siglo XX, uno de los primeros filósofos que se detuvo a reflexionar particularmente sobre el impacto de la tecnología en función de la continuidad y futuro del ser humano y de la humanidad, y más en general del fenómeno de la vida en la Tierra, fue Hans Jonas (1903-1993), autor del influyente libro *El Principio de responsabilidad. Ensayo de una ética para la civilización tecnológica* (1995) (edición original, 1979: *Das Prinzip Verantwortung*)<sup>(100)</sup>.

Frente a la incertidumbre que generan los nuevos poderes tecnológicos que el hombre ha adquirido y seguirá adquiriendo, Jonas postula una “heurística del temor”: “sólo la previsible desfiguración del hombre nos ayudará a alcanzar aquel concepto de hombre que ha de ser preservado de tales peligros”<sup>(101)</sup>.

Ante el apocalipsis que la tecnología puede desencadenar, Jonas propondrá una fundamentación ontológica de la ética. Jonas razona desde la premisa que la naturaleza (digamos, el fenómeno de la vida) tiene un fin intrínseco e inherente, cual es la conservación de la vida. La finalidad de conservar la vida en la naturaleza es un valor en sí. Observa Jonas que la mera posibilidad de atribuir valor a lo que es, “independientemente de lo mucho o lo poco que se encuentre actualmente presente, determina la superioridad del ser sobre la nada”<sup>(102)</sup>. Pues bien, la finalidad de

---

(100) No olvido, por cierto, las reflexiones de Ortega y de Heidegger sobre la técnica. Pero se enfocan más en comprender el significado de la técnica en relación con la constitución ontológica del ser humano. Contenida en el caso de Ortega en “Meditación de la técnica”, desarrollada para un curso de la Universidad Internacional de Santander, 1933, y publicada junto a “Ensimismamiento y alteración”, 1939 (Ortega y Gasset, José, *Meditación de la técnica. Ensimismamiento y alteración*, Biblioteca Nueva, Madrid, 2015, e-book). Y respecto de Heidegger, en “La pregunta por la técnica” (conferencia 1953) y “Ciencia y meditación” (conferencia, 1953) (Heidegger, Martin, *Filosofía, Ciencia y Técnica*, 3a ed., Universitaria, Santiago, 1997).

(101) Jonas, Hans, *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*, Herder, Madrid, 1995, p. 16.

(102) *Ibidem*, p. 95.

la naturaleza, esa misteriosa pulsión hacia lo que denominados vida, es “una afirmación sustancial del ser que se pone en sentido absoluto como mejor respecto al no-ser”. Es mejor, tiene valor, “ser”, por sobre “no-ser”. El “ser” es preferible a la nada. El ser tiene superioridad por sobre el no-ser. Puesto que el “ser” tiende hacia su reproducción ontológica, lo cual implica que tiene valor en sí mismo, su conservación constituye un deber ser del hombre: aquel sujeto autoconsciente de su propia existencia, condición ontológica, que le impone la responsabilidad de cuidar la vida.

La autoafirmación de la vida misma, el ser siendo; quiero decir, el fin de continuar siendo por la finalidad intrínseca que constituye fenomenológicamente la vida, conduce el pensamiento de Jonas a concluir la posibilidad de superar la distinción entre ser y deber-ser. La pulsión de vida de la naturaleza, su irrupción frente a la ausencia de vida (no sobre la muerte), implica el deber-ser, pues implica un valor en sí, que la hace que “deba ser”. Si existe, ergo, debe ser. Por consiguiente, “ser” y “deber-ser”, ontológicamente vistos, son uno y se imbrican recíprocamente.

Pues bien, será a partir de la propuesta ontológica de superación de la distinción “ser/no ser”, que Jonas derivará un primer imperativo categórico fundamental: “que exista una humanidad”. Y es que Jonas sustituye el imperativo kantiano “actúa de tal modo que el principio de tu acción se transforme en una ley universal”, por el siguiente: “actúa de tal modo que los efectos de tu acción sean compatibles con la permanencia de una vida humana auténtica”. Y formulado negativamente: “no pongas en peligro la continuidad indefinida de la humanidad en la Tierra”<sup>(103)</sup>.

Pero, además, debe considerarse que la humanidad, en tanto fenómeno vital, tiene como condición necesaria de su existencia a la biosfera. Por consiguiente, el actuar del hombre de modo tal que su acción sea compatible con la permanencia de la vida humana auténtica, implica también la acción de conservación de la biosfera y de toda la naturaleza extrahumana, pues sin ella, el hombre y más en general, el fenómeno de la vida, al menos en la Tierra y hasta donde sabemos, no sería posible. Preservar la naturaleza significa, por tanto, también, preservar al ser humano y la humanidad. Hasta antes del advenimiento de la revolución tecnológica, la naturaleza se cuidaba a sí misma, no era objeto de responsabilidad humana. Ahora el poder de la tecnología hace posible su profunda alteración, sino su destrucción.

De una ética clásica antropocéntrica, la filosofía moral de Jonas nos permite pasar a una ética ecológica. Pensamos que hay argumentos para

---

(103) De Siqueira, José Eduardo, “El principio responsabilidad de Hans Jonas”, *Revista BioEthikos*, Centro Universitario Sao Camilo, 3 (2), 2009, p. 172.

admitir que el pensamiento de Jonas ha contribuido a la formación de la conciencia ecológica contemporánea<sup>(104)</sup>. En efecto, la autocomprensión del hombre, en tanto ser racional, no puede sino hacerlo responsable no solo del bien actuar frente a los hombres, sino que también frente a la naturaleza. Toda acción del hombre debe ser compatible con la permanencia de una vida humana auténtica, no poniendo en peligro la continuidad indefinida de la humanidad en la Tierra y de la naturaleza misma (la Tierra), pues ella es condición necesaria de su existencia (como especie, se entiende). Pesa sobre sus hombros, pues, un principio de responsabilidad del cual se infiere, asimismo, un principio de cautela en el actuar.

Como ha sido resaltado por De Siqueira, al formular su imperativo de responsabilidad, Jonas está pensando no tanto en el peligro de la pura y simple destrucción física de la humanidad, sino en su muerte esencial, aquella que adviene de la desconstrucción y la aleatoria reconstrucción tecnológica del hombre y del medio ambiente<sup>(105)</sup>.

Jonas configuró uno de los pilares filosóficos del discurso bioconservador de la humanidad que bien, podría sintetizarse en la siguiente frase: no es lícito apostar la existencia del hombre. La conservación de la vida exige una lógica precautoria, lo cual obliga a darle crédito a los pronósticos catastróficos más que a los favorables. El hombre, ser moral, “respecto a la vida de la humanidad es obligado —lo que no tiene por qué ser así en la vida de los pacientes individuales— preferir paliativos deficientes a una prometedora cura radical que comporta el riesgo de la muerte del paciente”<sup>(106)</sup>. Se produce así una inversión del principio cartesiano de la duda, pues tratándose de decisiones que afectan la vida, debemos tratar lo dudoso, como posible. El principio de responsabilidad implica, por consiguiente, un principio de precaución (PP)<sup>(107)</sup> que, de manera muy

(104) Sobre el particular, véase González Gómez, Germán, “El principio de responsabilidad de Hans Jonas a la luz de la conciencia ecológica”, *Nómadas. Revista Crítica de Ciencias Sociales y Jurídicas*, 46, 2015, 2.

(105) De Siqueira, José Eduardo, “El principio Responsabilidad de Hans Jonas”, *Revista BioEthikos*, Centro Universitario Sao Camilo, 3 (2), 2009, p. 172.

(106) Jonas, Hans, *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*, Herder, Madrid, 1995, p. 80.

(107) Existe una extensa bibliografía sobre el principio de precaución. Entre otros estudios, se ha tenido presente: O’Riordan, Timothy - Cameron, James (eds.), *Interpreting the precautionary principle*, Earthscan Publications, Londres, 1994; Artigas, Carmen, *El principio precautorio en el derecho y la política internacional*, Naciones Unidas, CEPAL/ECLAC, Santiago, 2001; Stevens, Mary, “The Precautionary Principle in the International Arena”, *Sustainable Development Law & Policy*, 2002, 2 (2): 13-22; Sunstein, Cass R., “Beyond the Precautionary Principle”, *The Law School. The University of Chicago, John M. Olin Law & Economics Working paper*, nro. 149 (2D series), disponible en: <https://dash.harvard.edu/handle/1/12876710> (consultado el

elemental, admitiría ser definido como una directriz normativa para tomar decisiones de curso de acción ante situaciones de incertidumbre científica en la evaluación y la gestión de riesgos. Ante la incertidumbre de los efectos que ciertas acciones pudieran tener en contra de la salud humana y el medio ambiente, el PP aconseja tomar todas las medidas eficaces para no incurrir en tales acciones. La finalidad del PP es, por consiguiente, “proteger a los seres humanos y al medio ambiente de los riesgos que entraña la actividad humana mediante el ejercicio de un control anterior al daño (medidas preventivas)”<sup>(108)</sup>. Se estima que, en la actualidad, más de 60 acuerdos e instrumentos tanto universales como regionales vinculantes consagran al principio de precaución de diversos modos<sup>(109)</sup>.

Pues bien, y relacionando todo lo hasta aquí expuesto en función de la fundamentación del DIA, convengamos en que, a la luz del principio de responsabilidad de Jonas, el proyecto transhumanista, con su afán de controlar la evolución humana mediante el uso de la tecnología, con el propósito explícito de llevar al hombre a la etapa evolutiva siguiente, resulta un proyecto éticamente inviable porque no garantiza la conservación de la humanidad, sino que todo lo contrario, promueve directamente la posthumanidad. En Jonas es lícito hablar de un derecho al suicidio individual pero nunca del derecho al suicidio o autodestrucción de la humanidad<sup>(110)</sup>.

Así, lo argumentado hasta ahora, me ha llevado a la conclusión que, bajo la óptica del principio de responsabilidad, el transhumanismo, sobre todo en sus corrientes más fuertes, podría ser concebido (metafóricamente) como un proyecto de suicidio racional de la humanidad, pues las modificaciones geno-fenotípicas del *homo sapiens* que suponen el mejoramiento tecnológico continuo del ser humano (IA y genómica), conducen a la deconstrucción y reconfiguración orgánica y moral de la

---

20/03/2019); UNESCO/COMEST, *Informe del Grupo de Expertos sobre el principio precautorio*, París, 2005; De Sadeleer, Nicolas (ed.), *Implementing the Precautionary Principle: Approaches from the Nordic Countries, EU and USA*, Earthscan, Londres, 2007; Bonamigo, Elcio L., *El principio de precaución: un nuevo principio bioético y biojurídico*, tesis doctoral, Universidad Rey Juan Carlos, Madrid, 2010; European Commission, *Future Brief: The precautionary principle: decision-making under uncertainty*, September, Issue 18, 2017.

(108) COMEST/UNESCO, “Informe del Grupo de Expertos sobre el principio de precaución de la Comisión Mundial de Ética del Conocimiento Científico y la Tecnología”, UNESCO, París, 2005, p. 49.

(109) Drnas de Clément, Zlata, “Aspectos conceptuales del principio de precaución ambiental”, *Anuario Hispano-Luso-Americano de Derecho Internacional*, nro. 18, 2007, p. 599.

(110) Jonas, Hans, *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*, Herder, Madrid, 1995, p. 81.

concepción intuitiva (heurística si se prefiere) y también típica que, del ser humano, hemos tenido: “este almacén de huesos y pellejo”, de que hablará el poeta sevillano<sup>(111)</sup>.

Pongo fin a esta sección, mencionando que más contemporáneamente, se han destacado por sus críticas a los proyectos de “mejoramiento” biotecnológico del ser, Fukuyama, Sandel y Habermas. Una brevísima revisión de sus ideas, contribuirá a solidificar la base ética desde la cual propongo los fundamentos de un Derecho de la Inteligencia Artificial.

Fukuyama. En el año 2004 la revista *Corning Policy* (septiembre-octubre) le formuló a ocho destacados intelectuales la siguiente pregunta: ¿qué ideas, si se abrazan, representarían la mayor amenaza al bienestar de la humanidad? Fukuyama, que era parte de ese grupo, respondió sin vacilación alguna que el transhumanismo, como cuerpo de ideas, constituía esa amenaza<sup>(112)</sup>. Pero será en su libro *Our Posthuman Future. Consequences of the biotechnology revolution* (2017) donde profundizará su crítica al transhumanismo. Fukuyama dice ahí que hay buenas razones de prudencia para no alterar el orden natural y olvidar cualquier idea al respecto. Pretender hacer las cosas mejor que la naturaleza es arriesgado. La evidencia muestra que “(...) la evolución puede ser un proceso ciego, pero sigue una lógica de adaptación rigurosa, que hace que los organismos se adapten al entorno”<sup>(113)</sup>. La naturaleza humana, esto es, eso que somos y que históricamente hemos sido, biológica y culturalmente, es fundamental para nuestra comprensión de la sociedad, la justicia, la moral y las nociones de vida buena. Nuestra naturaleza humana, nuestra biología, si se quiere, implica, por tanto, un conjunto de principios éticos que fundamentan derechos naturales. Introducir a la humanidad en una dinámica de alteraciones biotecnológicas en el ser humano, puede conducirla a su fin.

Sandel. En su famoso libro *Contra la perfección. La ética en la era de la ingeniería genética* (*The Case Against Perfection. Ethics in the Age of Genetic Engineering*) Sandel, a su turno, se pregunta en qué sentido el perfeccionamiento, la clonación y la ingeniería genética reducen nuestra humanidad: ¿qué aspectos de la libertad o del desarrollo humano resultan amenazados por ellas?

Sandel responde que se vería amenazada “nuestra capacidad para actuar libremente, en virtud de nuestros propios medios y esfuerzos, y

(111) Gustavo Adolfo Bécquer, rima LVII.

(112) Puede ser revisada online esa afirmación de Fukuyama en <https://foreignpolicy.com/2009/10/23/transhumanism/> (consultado el 03/11/2018).

(113) Fukuyama, Francis, *Our Posthuman Future. Consequences of the biotechnology revolution*, Profile Books, London, 2017, e-book, loc. 1776.

para considerarnos a nosotros mismos responsables —merecedores de elogio o condena— por nuestra forma de ser y de hacer las cosas”<sup>(114)</sup>. Así, y puestos en el amplio campo del perfeccionamiento artificial del rendimiento atlético, concluye que, a mayor dependencia del atleta de fármacos o alteraciones genéticas, menos logro es posible distinguir en su rendimiento. Un bateador que enviara siempre las pelotas a la zona de *strike* e hiciera *home run*, “no sería un agente en lo absoluto; sus logros serían los de su inventor”<sup>(115)</sup>. Estaríamos frente a una máquina humana perfecta, que siempre acierta, que no comete errores.

Sin embargo, el principal peligro que advierte Sandel no es que esa clase de perfección erosione la capacidad de acción humana, trastocando las nociones de libertad y responsabilidad moral. No. El peligro más profundo, escribe Sandel, “reside más bien en que son el reflejo de una ampliación desmesurada del campo de la acción humana, de una aspiración prometeica a rehacer la naturaleza, incluida la naturaleza humana, para servir a nuestros propios deseos. El problema no es la pendiente hacia la mecanización, sino la ambición de dominio”<sup>(116)</sup>. Ese deslizamiento hacia la ambición de dominio puede conducir al olvido y hacia la destrucción misma, de la apreciación del carácter recibido de los poderes y los logros humanos. Destruída esa apreciación o toma de conciencia, el hombre comienza su proceso de deshumanización.

Habermas. Por último, cabe no perder de vista a Habermas, quizás uno de los más importantes sociólogos y filósofos del cambio de siglo XX-XXI, quien escoge la investigación con células madre y el diagnóstico genético preimplantacional (DGP), como los casos desde los cuales reflexionar en torno a las consecuencias del desarrollo tecnológico.

La fuerza normativa de lo fáctico, “sólo alienta el temor de la esfera pública escéptica a que la dinámica sistémica de la ciencia, la técnica y la economía obre *faits accomplis* a los que normativamente ya no se pueda dar alcance”<sup>(117)</sup>. Habermas teme que la política de los hechos consumados (*faits accomplis*) en materia de células madre, DGP y, en general, de desarrollo biotecnológico, acabe por imponerse. Por consiguiente, surge la necesidad de desarrollar principios normativos que regulen especialmente el límite fluctuante entre terapia y mejoramiento, sobre todo hoy

---

(114) Sandel, Michael, *Contra la perfección. La ética en la era de la ingeniería genética*, 2a ed., trad. de Ramón Vilá Vernis, Marbot Barcelona, 2015, p. 69.

(115) *Ibidem*, p. 70.

(116) *Ibidem*, p. 71.

(117) Habermas, Jürgen, *El futuro de la naturaleza humana. ¿Hacia una eugenesia liberal?*, trad. de R. S. Carbó, Paidós, Barcelona, 2017, p. 31.

por hoy cuando la eugenesia liberal<sup>(118)</sup> avanza sin prisa pero sin pausa, con su proyecto de desdibujamiento de la frontera entre intervención terapéutica e intervención perfeccionadora, dejando que sean las preferencias individuales de los participantes en el mercado las que elijan los objetivos de la modificación de marcas características.

Según Habermas, la aplicación de la técnica DGP plantea la pregunta normativa de “si es compatible con la dignidad de la vida humana ser engendrado con reservas y solo ser declarado digno de existir y desarrollarse después de un examen genético. ¿Debemos disponer libremente de la vida humana con fines selectivos?”<sup>(119)</sup>. Habermas responde que la manipulación genética puede llevarnos a modificar nuestra autocomprensión como especie en términos tales de erosionar los fundamentos morales y jurídicos de la sociedad, pues trastocará nuestra identidad como especie y, por tanto, nuestra sociabilidad. Así Habermas se pregunta si la indisponibilidad de los fundamentos biológicos de la identidad puede fundamentar un derecho fundamental a la herencia genética natural (no intervenida artificialmente).

### 3.2.1. SÍNTESIS DE LA ÉTICA PARA UN DÍA

Si bien es cierto que no existe certeza científica del eventual daño que la IA podría ocasionar al ser humano y a la continuidad de este como especie (la humanidad), tampoco existe duda en cuanto a los riesgos que entraña, sobre todo a juzgar por los daños que ya está ocasionando a ciertos derechos fundamentales (libertad, igualdad, trabajo, mediante algoritmos discriminativos). La lista de científicos que han alzado sus voces para advertir a la humanidad de los inminentes peligros que encierra, como ya se ha señalado, es larga. El peligro del desarrollo desregulado que encierra la IA, lo mismo que la imagen en la fotografía analógica, está, por tanto, latente. Solo hacen falta las emulsiones adecuadas y los pro-

(118) Según Sandel (Sandel, Michael, *Contra la perfección. La ética en la era de la ingeniería genética*, 2a ed., trad. de Ramón Vilá Vernis, Marbot, Barcelona, 2015, ps. 128-130), la eugenesia liberal es un movimiento heterodoxo de filosofía moral y política angloamericano, que propugna una optimización genética no coercitiva de los hijos, de carácter privado y en un contexto de neutralidad del Estado. En la lista de autores que la apoyan, es posible distinguir a Buchanan, Brock, Daniels, Wikler. También Dworkin, Nozick y el mismísimo Rawls. La tesis general de la eugenesia liberal es que mientras ni la educación ni la alteración genética violen la autonomía del hijo, o su derecho a un “futuro abierto”, todo está permitido. En la medida en que la capacidad optimizada sea un medio “abierto en cuanto a los fines”, y, por lo tanto, no dirija al hijo hacia una carrera o plan de vida particular, es moralmente permisible.

(119) Habermas, Jürgen, *El futuro de la naturaleza humana. ¿Hacia una eugenesia liberal?*, trad. de R. S. Carbó, Paidós, Barcelona, 2017, p. 34.

cedimientos correctos para que se revelen. Pues bien, precisamente, es ese potencial de riesgo latente que posee el desarrollo descontrolado de la IA, lo que justifica su control por aplicación del principio de responsabilidad y su consecuencia lógica, el principio de precaución.

En efecto, si el principio de precaución establece que la falta de certeza no puede ser una base para desproteger la salud humana o el medio ambiente; pues, con mayor razón entonces, lo será la protección de la supervivencia y continuidad de la especie humana.

Donde existan amenazas de daños graves o irreversibles, la falta de certidumbre científica total no debe usarse como razón para posponer la adopción de medidas. El principio de precaución surge, precisamente, para enfrentar las situaciones de incertidumbre de los daños que pueden causar los avances en la tecnología.

En el campo de la inteligencia artificial, el razonamiento ético en función del principio de precaución dice lo siguiente: efectivamente no tenemos aún prueba científica concluyente que el desarrollo descontrolado y en perspectiva transhumanista de la IA, necesariamente, vaya a producir el fin del hombre y de la humanidad. Si, en cambio, tenemos presunciones que su desarrollo sin control normativo —más aún en perspectiva transhumanista— puede afectar gravemente al ser humano, incluso poniendo en riesgo la continuidad de su especie. Así las cosas, la responsabilidad que la actual generación de la humanidad tiene para con las futuras, implica actuar con precaución, pues lo que está en juego es la supervivencia de la humanidad. El riesgo es insoportable.

El potencial de crecimiento, impacto y desarrollo extraordinario de la IA sobre la humanidad impone a las actuales y futuras generaciones de seres humanos la responsabilidad de guiar su desarrollo en función del bienestar y conservación de la humanidad, actuando con cautela y precaución. Insisto una vez más: no se trata de detener el desarrollo de la IA, sino de orientar su progreso un poco en la línea de lo que se ha venido en denominar por la Organización de las Naciones Unidas (ONU), *AI for good* (IA para el bien del mundo), cuya finalidad es brindar soluciones a los problemas sociales que originan la pobreza, el hambre, el complejo acceso a salud; la falta de igualdad (incluida la igualdad de género<sup>(120)</sup>); el difícil acceso a una educación de calidad; la degradación del medio ambiente, entre otros. Se trata, en definitiva, de poner la IA al servicio de

---

(120) Nota del director: sobre como la inteligencia artificial puede contribuir a mitigar la desigualdad y la violencia de género ver Machiavelli, Nieves, “La violencia de género y el uso de algoritmos como herramienta efectiva para la protección de los derechos fundamentales”, tomo I del presente *Tratado*.

la sustentabilidad del mundo: del mundo habitado por hombres y seres vivos (biológicos se entiende).

El fundamento ético universal de las regulaciones jurídicas en materia de IA, por tanto, encuentra en el principio de responsabilidad y en el de precaución, su justificación. Por otra parte, si lo que está en juego con el desarrollo de la IA, finalmente, es la continuidad de la especie humana, estamos, entonces, básica y esencialmente, frente a un tema de derechos humanos.

#### IV. CONFIGURACIÓN UNIVERSAL DEL DERECHO DE LA INTELIGENCIA ARTIFICIAL (DIA)

He argumentado en este artículo que el principio de responsabilidad, que exige de los hombres individualmente considerados y de la humanidad precaución y cautela frente a los avances de la tecnología, constituye el fundamento ético del DIA. Los riesgos que, para el futuro de humanidad, implica el desarrollo descontrolado de la IA (y más aún en perspectiva transhumanista) son tan evidentes, que los líderes civilizaciones contemporáneos (Estados, Organismos Internacionales, Grandes Corporaciones Privadas<sup>(121)</sup>) no han podido sino tenerlo presente en su agenda pública. Así es como inspirados por el principio de responsabilidad están tomando conciencia de la importancia de actuar con cautela en el campo de la IA en términos tales advertir la necesidad de regular jurídicamente su desarrollo. Una breve mirada a lo que está pasando en las Organización de las Naciones Unidas (ONU), Europa y Estados Unidos de Norteamérica (EE.UU.), nos permitirá advertir que el mundo avanza hacia la creación de un DIA y que la propuesta de un DIA que hago en este artículo, por tanto, se inserta dentro de esa tendencia. La muestra de esos antecedentes reforzará la argumentación en el sentido que la configuración de los fundamentos más generales DIA no emergen desde el “vacío”, sino dimanando de sus fuentes materiales, se trata de una tendencia contemporánea que esta investigación ha pretendido coronar.

---

(121) Los límites formales de este artículo han impedido que se presente las contribuciones en materia de regulaciones éticas a la IA, que ciertas Corporaciones privadas, incumbentes en la materia, han realizado el último tiempo. Al respecto, debe tenerse presente entre otras, las del Institute of Electrical and Electronics Engineers (IEEE) (véase, IEEE, *Ethically Aligned design. A vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, 2019); Google (véase, Google, “Artificial Intelligence at Google. Our principles”, <https://ai.google/principles> [consultado el 01/05/2020] y Microsoft (Microsoft, “Microsoft AI principles, 2020”, <https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimar6>; *The Future Computed. Artificial Intelligence and its roles in society*, Microsoft Corporation, Redmond, Washington, 2018).

#### 4.1. UNA MIRADA A LA ONU: AI FOR GOOD

*AI for Good* (inteligencia artificial para el bien) es la plataforma líder de ONU para el diálogo sobre IA. Como organismo especializado de las Naciones Unidas para las TIC, la Unión Internacional de Telecomunicaciones (ITU)<sup>(122)</sup>, en asociación con agencias hermanas de las Naciones Unidas, organiza “AI for Good Global Summit” para el diálogo internacional, destinado a construir una comprensión común de las capacidades de las emergentes tecnologías de la IA. Hasta la fecha se han realizado tres *Suunit*, 2017, 2018 y 2019 (todas en Ginebra, Suiza)<sup>(123)</sup>.

La ONU opina que IA brinda una oportunidad sin precedentes para erradicar el hambre, poner fin a la pobreza e invertir el proceso de degradación de nuestro entorno natural. Con las cumbres *AI for Good* se aspira a fomentar la innovación haciendo hincapié en las extraordinarias posibilidades a las que podremos acceder próximamente gracias a la aceleración del progreso de la IA. En concreto la ONU fomenta la IA como herramienta para acelerar el progreso de los Objetivos de Desarrollo Sustentable (ODS), que son: (1) fin a la pobreza, (2) cero hambre, (3) salud y bienestar, (4) educación de calidad, (5) igualdad de género, (6) agua limpia y saneamiento, (7) energía asequible y no contaminante, (8) trabajo decente y crecimiento económico, (9) industria, innovación e infraestructura, (10) reducción de desigualdades, (11) ciudades y comunidades sostenibles, (12) producción y consumo responsable, (13) acción por el clima, (14) vida submarina, (15) vida ecosistemas terrestres, (16) paz, justicia e instituciones sólidas y (17) alianzas para lograr esos objetivos<sup>(124)</sup>.

Los ODS fueron adoptados el 25 de septiembre de 2015 por los más de 150 líderes mundiales que asistieron a la Cumbre de las Naciones Unidas sobre el Desarrollo Sostenible en Nueva York con el fin de aprobar la Agenda para el Desarrollo Sostenible. El documento final, titulado “Transformar Nuestro Mundo: la Agenda 2030 para el Desarrollo Sostenible”, fue adoptado por los 193 Estados Miembros de las Naciones Unidas. Dicho documento incluye los 17 Objetivos del Desarrollo Sostenible (ODS) cuyo objetivo es poner fin a la pobreza, luchar contra la desigualdad y la injusticia, y hacer frente al cambio climático sin que nadie quede atrás para el 2030.

---

(122) Más información sobre la ITU en su sitio web: <https://www.itu.int/en/ITU-T/about/Pages/default.aspx> (consultado el 30/05/2021).

(123) La convocatoria a la última AI for Good Global Summit, 2021, puede ser revisada en <https://aiforgood.itu.int/> (consultado el 03/03/2019).

(124) Para una revisión detalle de los ODS, consúltese: <https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/> (consultado el 30/05/2019).

Si bien es cierto la agenda 2030 para el Desarrollo Sostenible no tiene fuerza normativa, no por eso es menos importante o soslayable. En efecto, dado que representa un consenso ético global, sirve para orientar a nivel mundial los futuros desarrollos normativos estatales. La perspectiva de *AI for good* de las Naciones Unidas goza de aceptación mundial y está en el mismo horizonte de respuestas que sobre el tema, por lo demás, han dado las principales ONG de Derechos Humanos, tales como *Humans Right Watch*, *Amnesty International*, *Access Now* y *Wikimedia Foundation*. En efecto, el 16 de mayo de 2018, al amparo del *RightsCon Toronto*<sup>(125)</sup>, esas cuatro organizaciones emitieron la denominada “The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems”<sup>(126)</sup>, en la cual se aboga por guiar el desarrollo de la IA dentro del marco del derecho internacional de los derechos humanos.

#### 4.2. UNA MIRADA A EUROPA

En el caso de la Unión Europea (UE), la primera manifestación completa del tema fue el denominado Proyecto Robolaw<sup>(127)</sup>, iniciado en marzo de 2012 y concluido en mayo de 2014. El objetivo principal del Proyecto fue comprender las implicaciones legales y éticas de las tecnologías robóticas emergentes y descubrir (1) si los marcos legales existentes son adecuados y viables a la luz del advenimiento y la rápida proliferación de tecnologías robóticas, y (2) de qué manera los desarrollos en el campo de la robótica afectan las normas, los valores y los procesos sociales que valoramos. El proyecto RoboLaw se centró en las fronteras extremas del avance tecnológico, enfrentando el “estatus” legal de la robótica, las nanotecnologías, las neuroprótesis, las interfaces cerebro-computadora. En el documento final del proyecto, “Guidelines on Regulating Robotics”<sup>(128)</sup>, se reflexiona sobre el papel que juega la ética en la robótica y, más en general en la IA, por cuanto se parte del supuesto que la ética puede desempeñar múltiples funciones en la perspectiva de regular las tecnologías emergentes, tanto directas como indirectas. En efecto, (i) la ética enmar-

(125) RightsCon es una de las más importantes Conferencias Mundiales sobre Derechos Humanos en la era digital. Access Now las viene organizando desde el año 2011. Para más información: <https://www.rightscon.org/separa-las-fechas-rightscon-tunez/> (consultado el 10/06/2019).

(126) Entre otros lugares, puede consultarse en el sitio web de Access Now: <https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/> (consultado el 10/06/2019).

(127) El título oficial de “Proyecto Robolaw” es: *Regulating Emerging Robotic Technologies in Europe: Robotics facing Law and Ethics*. Las bases del proyecto pueden ser consultadas en <http://www.robolaw.eu/projectdetails.htm> (consultado el 02/03/2021).

(128) Puede consultarte en línea: [http://www.robolaw.eu/RoboLaw\\_files/documents/robolaw\\_d6.2\\_guidelinesregulatingrobotics\\_20140922.pdf](http://www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922.pdf) (consultado el 29/03/2019).

ca las expectativas sociales hacia las tecnologías emergentes; (ii) la ética puede estar incorporada en el diseño de nuevas tecnologías y, (iii) la ética es un sistema de valores involucrados en la construcción de una cultura democrática de la tecnología.

Posteriormente en la Estrategia Global para la Política Exterior y de Seguridad de la Unión Europea, se establece entre las prioridades de su acción exterior la ciberseguridad y, dentro del capítulo de las prioridades de gobernanza mundial para el siglo XXI, recoge la necesidad de disponer de “normas mundiales en ámbitos tales como la biotecnología, la inteligencia artificial, la robótica y los aparatos pilotados a distancia, con el fin de evitar riesgos de seguridad y aprovechar sus beneficios económicos”<sup>(129)</sup>.

En consonancia con dicha prioridad de política exterior y seguridad, en abril de 2019, la Comisión Europea, a través del Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la UE, presentó un documento de trabajo titulado “Proyecto de Directrices Éticas sobre una IA Confiable”. En el mencionado estudio, se remarca que la IA es una de las fuerzas más transformadoras de nuestro tiempo y está destinada a modificar el tejido social. Y que, así como supone una gran oportunidad para aumentar la prosperidad y el crecimiento que Europa debe tratar de lograr, también implica una serie de innegables riesgos para la humanidad<sup>(130)</sup>.

La visión europea dice que, si bien los beneficios de la IA compensan sus riesgos, se debe seguir una línea de desarrollo que maximice los beneficios y minimice los riesgos. Para asegurar tal resultado es necesario regirse por un enfoque de la IA centrado en los seres humanos. Es decir, que obligue a recordar que el desarrollo y uso de la IA tienen por objetivo mejorar el bienestar de los seres humanos, y no verlos como un medio en sí mismos. Se propone, entonces, lo que denominan IA confiable. La fiabilidad de la IA se apoya en tres componentes que deben satisfacerse a lo largo de todo el ciclo de vida del sistema: a) la IA debe ser lícita, es decir, cumplir todas las leyes y reglamentos aplicables; b) ha de ser ética, de modo que se garantice el respeto de los principios y valores éticos; y c) debe ser robusta tanto desde el punto de vista técnico como social, puesto que los sistemas de IA, incluso si las intenciones son buenas, pueden provocar daños accidentales. Cada uno de estos componentes es en sí mismo necesario, pero no suficiente para el logro de una IA fiable. Lo ideal es que todos ellos actúen en armonía y de manera simultánea. En el

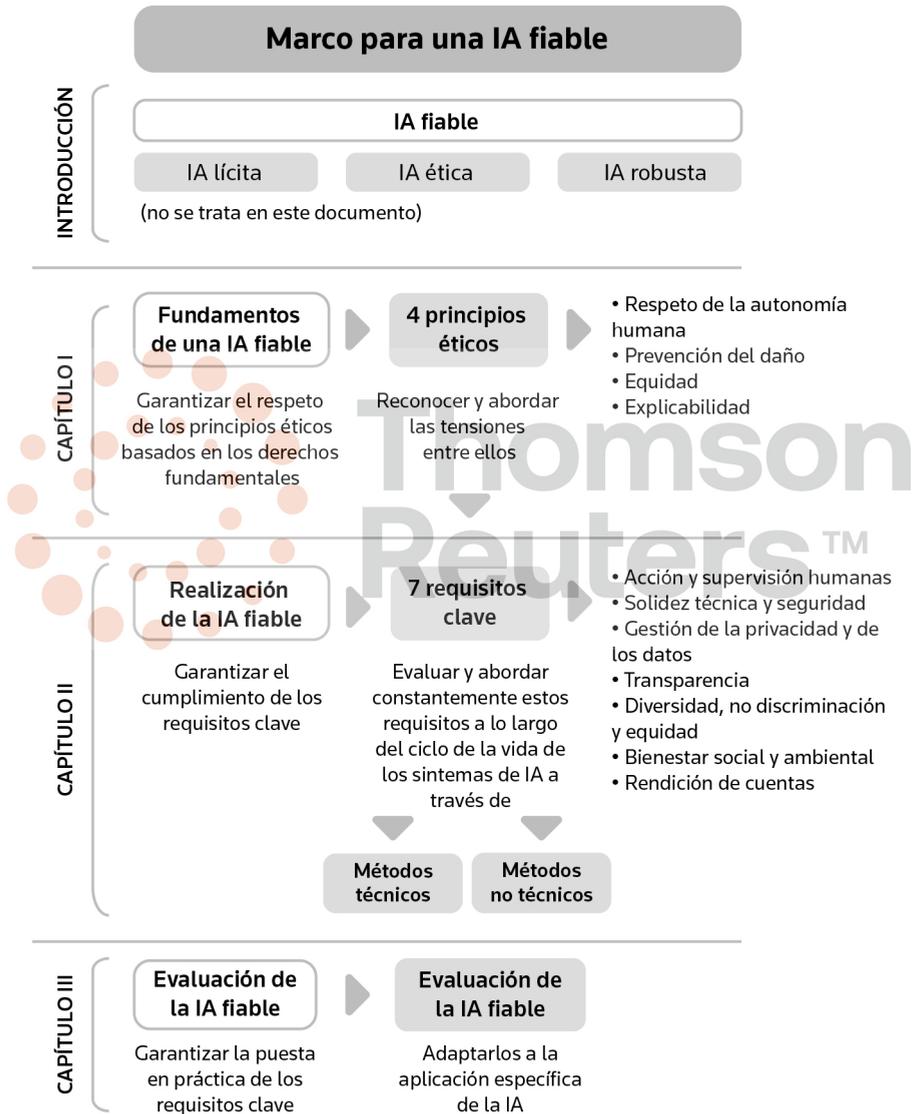
---

(129) Comisión Europea, *Una visión común, una actuación conjunta: una Europa más fuerte. Estrategia global para la política exterior y de seguridad de la Unión Europea*, Bruselas, 2016, p. 34.

(130) Comisión Europea, *Directrices éticas para una IA fiable*, Grupo de Expertos de alto nivel sobre inteligencia artificial, Bruselas, 2019.

caso de que surjan tensiones entre ellos en la práctica, la sociedad deberá esforzarse por resolverlas<sup>(131)</sup>.

Asimismo, se enfatiza que la fiabilidad de los sistemas de IA debe asegurarse mediante la observación de los siguientes principios: respeto de la autonomía humana, prevención del daño, equidad y explicabilidad.



Fuente: Comisión Europea, *Directrices éticas...*, cit., p. 10.

(131) Comisión Europea, *Directrices éticas para una IA fiable*, Grupo de Expertos de alto nivel sobre inteligencia artificial, Bruselas, 2019, p. 2.

En síntesis, el Grupo de expertos de alto nivel sobre la IA postula para Europa un enfoque de la ética de la IA que utiliza el compromiso con los derechos fundamentales contenidos en los Tratados de la UE y la Carta de los Derechos Fundamentales como el medio para identificar principios éticos abstractos, y para especificar cómo los valores éticos concretos pueden operacionalizarse en el contexto de la IA. Y es que, normativamente, no podría ser de otro modo. En efecto, como es conocido, la UE se basa en un compromiso Constitucional de proteger los derechos fundamentales de los seres humanos, garantizar el respeto del estado de derecho, promover la libertad democrática y el bien común<sup>(132)</sup>.

En materia de derecho de los robots (Robolaw), la famosa resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre Normas de Derecho Civil sobre Robótica (2015/2103[INL]), señala que el marco de orientaciones éticas sobre las cuales construir las regulaciones jurídicas en IA y robótica:

“(...) debe basarse en los principios de beneficencia, no maleficencia, autonomía y justicia, así como en los principios consagrados en la Carta de los Derechos Fundamentales de la Unión Europea, como la dignidad humana, la igualdad, la justicia y la equidad, la no discriminación, el consentimiento informado, la vida privada y familiar y la protección de datos, así como en otros principios y valores inherentes al Derecho de la Unión, como la no estigmatización, la transparencia, la autonomía, la responsabilidad individual, y la responsabilidad social, sin olvidar las actuales prácticas y códigos éticos”<sup>(133)</sup>.

Por último, todos estos esfuerzos regulativos europeos han visto sus frutos en el proyecto de Ley Europea que, con fecha del 21 de abril de 2021, la Comisión Europea introdujo en el sistema legislativo titulado *Proposol for a Regulation laying down harmonized rules on artificial intelligence (Artificial Intellegince Act)*. Dice la Comisión Europea que, ante el rápido desarrollo tecnológico de la IA y un contexto político global en el que cada vez más países están invirtiendo fuertemente en IA, la Unión Europea debe actuar como una sola para aprovechar las numerosas oportunidades y abordar los desafíos de la IA de una manera preparada para el futuro. Para promover, entonces, el desarrollo de la IA y abordar los posibles riesgos elevados que plantea para la seguridad y los derechos funda-

---

(132) Véase *Treaty Establishing A Constitution For Europa* (2005), más conocido como Treaty of Lisbon (2007) (Tratado de Lisboa).

(133) Parlamento Europeo (2017), *P8\_TA(2017)0051 Normas de derecho civil sobre robótica. Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2013[INL])*, p. 8.

mentales, es que la Comisión presenta la propuesta de marco regulador sobre la IA y un plan coordinado revisado sobre la IA para Unión Europea.

Se trata de una iniciativa paradigmática porque es el primer proyecto de Ley en el mundo que propone una regulación general de IA. Consta de 12 Títulos y 85 artículos en los cuales se regulan algunos de los aspectos más evidentemente peligrosos para los derechos fundamentales de las personas. La estructura del articulado que se propone es la siguiente:

- Título I: Disposiciones generales (*General provisions*). Arts. 1º-4º.
- Título II: Prácticas de IA prohibidas (*Prohibited Artificial Intelligence Practices*). Arts. 4º-5º.
- Título III: Sistemas de IA de Riesgo Alto (*High-Risk AI Systems*). Arts. 6º-51.
- Título IV: Obligaciones de transparencia para ciertos sistemas de IA (*Transparency obligations for certain AI systems*). Art. 52.
- Título V: Medidas de apoyo a la innovación (*Measures in support of innovations*). Arts. 53-55.
- Título VI: Gobernanza (*Governance*). Arts. 56-59.
- Título VII: Base de datos de la UE para sistemas de IA de alto riesgo independientes (*EU database for stand-alone high risk AI systems*). Art. 60.
- Título VIII: Seguimiento postventa, intercambio de información y vigilancia del mercado (*Post-market monitoring, information sharing, market surveillance*). Arts. 61-68.
- Título IX: Código de conducta (*Code of conduct*). Art. 69.
- Título X: Confidencialidad y sanciones (*Confidentiality and penalties*). Arts. 70-72.
- Título XI: Procedimiento de delegación de poder y comité (*Delegation of power and committee procedure*). Arts. 73-74.
- Título XII: Disposiciones finales (*Final provisions*). Arts. 75-85.

Esta aproximación conceptual a la IA ha sido muy debatida. Y a la fecha de edición de este trabajo de investigación, avanza el proyecto de Ley de Inteligencia Artificial de la UE en una definición de IA (art. 3º) que quita el foco definitorio en la idea de *software*, pasando a una más amplia. En efecto, para garantizar que la definición de un sistema de IA proporcione criterios suficientemente claros para distinguir la IA de los sistemas de *software* más clásicos, el texto propuesto reduce la definición del art. 3(1)

a los sistemas desarrollados a través de enfoques de aprendizaje automático y enfoques basados en la lógica y el conocimiento. Así fue como el día 25 de noviembre del 2022 del Comité de Representantes Permanentes del Consejo propuso al Consejo de la Unión Europea, la siguiente definición:

“Art. 3º Definiciones. A los efectos del presente Reglamento, se aplican las siguientes definiciones (1) ‘sistema de inteligencia artificial’ (sistema de IA): un sistema que está diseñado para funcionar con elementos de autonomía y que, basándose en datos e insumos proporcionados por máquinas y/o humanos, infiere cómo lograr un conjunto determinado de objetivos utilizando aprendizaje automático y/o enfoques basados en la lógica y el conocimiento, y produce resultados generados por el sistema, como contenido (sistemas generativos de IA), predicciones, recomendaciones o decisiones, que influyen en los entornos con los que interactúa el sistema de IA”<sup>(134)</sup>.

Siguiendo con esa línea de desarrollo base sobre IA, con fecha 6 de diciembre de 2022, la Secretaría General del Consejo de la Unión Europea envió a las Delegaciones una Orientación General Consejo sobre la propuesta de referencia, aprobada por el Consejo de Transporte, Telecomunicaciones y Energía en su sesión nro. 3917 celebrada el 6 de diciembre de 2022. En dicha orientación señala: “Resulta necesario definir con claridad la noción de sistema de IA para ofrecer seguridad jurídica, al mismo tiempo que se proporciona la flexibilidad necesaria para adaptarse a los futuros avances tecnológicos. Dicha definición debe basarse en las principales características funcionales de la inteligencia artificial, como su capacidad de aprendizaje, de razonamiento o de modelización, diferenciándola de otros sistemas de *software* y planteamientos de programación más sencillos. En particular, a los efectos del presente Reglamento, los sistemas de IA deben tener la capacidad de inferir, a partir de datos e información generados por máquinas o por seres humanos, la manera de alcanzar una serie de objetivos definidos por seres humanos, utilizando para ello estrategias de aprendizaje automático o estrategias basadas en la lógica y el conocimiento, y de generar información de salida, como contenidos para sistemas de inteligencia artificial generativa (por ejemplo, texto, vídeo o imágenes), predicciones, recomendaciones o decisiones que influyan en el entorno con el que interactúa el sistema, ya sea en una dimensión física o digital. Los sistemas que utilizan reglas definidas únicamente por personas físicas para ejecutar operaciones de manera automática no deben considerarse sistemas de IA. Los sistemas de IA pueden diseñarse para operar con distintos niveles de autonomía y utilizarse de manera inde-

---

(134) Disponible en: <chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf> (consulta: 19/02/2023).

pendiente o como componentes de un producto, con independencia de si el sistema forma parte físicamente de él (integrado) o tiene una funcionalidad en el producto sin formar parte de él (no integrado). El concepto de la autonomía de un sistema de IA se refiere a la medida en que dicho sistema puede funcionar sin intervención humana”.

Y agrega: “Las estrategias de aprendizaje automático se centran en el desarrollo de sistemas capaces de aprender y hacer inferencias a partir de datos para resolver un problema de aplicación, sin estar expresamente programados con una serie de instrucciones que abarquen todos los pasos desde la entrada hasta la salida. El aprendizaje se refiere al proceso informático de optimizar, a partir de los datos, los parámetros del modelo, que es un constructo matemático que genera una información de salida a partir de los datos de entrada. Los problemas que aborda el aprendizaje automático suelen incluir tareas en las que las otras estrategias fallan, ya sea porque el problema no está adecuadamente formalizado o porque su resolución es inabordable con estrategias sin aprendizaje. Las estrategias de aprendizaje automático incluyen, por ejemplo, el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo, y utilizan diversos métodos, entre los que se incluyen el aprendizaje profundo con redes neuronales, las técnicas estadísticas de aprendizaje e inferencia (como, por ejemplo, la regresión logística o la estimación bayesiana) y los métodos de búsqueda y optimización.

”Las estrategias basadas en la lógica y el conocimiento tienen por objeto el desarrollo de sistemas con capacidad de razonamiento lógico a partir de conocimientos para la resolución de un problema de aplicación. Estos sistemas suelen incluir una base de conocimientos y un motor de inferencia, que genera información de salida haciendo razonamientos a partir de la base de conocimientos. La base de conocimientos, normalmente codificada por expertos humanos, representa entidades y relaciones lógicas que son pertinentes para el problema de aplicación, usan do para ello formalismos basados en normas, ontologías y gráficos de conocimientos. El motor de inferencia interactúa con la base de conocimientos y extrae de ella información nueva a través de operaciones como la clasificación, la búsqueda, el emparejamiento o el encadenamiento. Las estrategias basadas en la lógica y el conocimiento incluyen, por ejemplo, la representación del conocimiento, la programación (lógica) inductiva, las bases de conocimiento, los motores de inferencia y deducción, el razonamiento (simbólico), los sistemas expertos y los métodos de búsqueda y optimización”<sup>(135)</sup>.

(135) Disponible en: [https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=consil%3AST\\_15698\\_2022\\_INIT](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=consil%3AST_15698_2022_INIT) (consulta: 19/02/2023).

### 4.3. UNA MIRADA A EE.UU.

Más allá del interesante debate técnico jurídico constitucional que, sobre la interpretación de la cuarta enmienda (derecho a la privacidad) en la era digital, tuvo lugar en “Riley v. California”<sup>(136)</sup>, lo interesante del caso es poner relieve cuan potente es la onda expansiva de la IA, que ya está presente, incluso, en el discurso judicial de tan importante Tribunal. La frase del juez Roberts, “los teléfonos celulares modernos... ahora son una parte tan generalizada e insistente de la vida cotidiana que el proverbial visitante de Marte podría concluir que eran una característica importante de la anatomía humana”, no hace sino referencia al fenómeno de la ciborgización que, como sabemos, es una cara más popular de la IA y que causa mayor atracción en la opinión pública.

Como es sabido EE.UU. es el país líder en IA en el mundo, sin embargo, en materia de regulaciones estatales tendientes a controlar el desarrollo de la misma, no exhibe una preocupación normativa tan verticalmente marcada como la UE, lo cual no significa necesariamente que no existan regulaciones o que estas potencialmente sean menos efectivas que las europeas.

La carta de navegación que EE.UU. tiene en la materia, data de 12 de octubre de 2016, y se trata del informe “Preparing for the future of artificial intelligence”<sup>(137)</sup>, presentado por el *Committee on Technology del National Science and Technology (NSTC), Subcommittee on Machine Learning and Artificial Intelligence*, adscritos al *Council de la Executive Office of the President*. El tono del informe es optimista en cuanto a que la IA tiene el potencial para mejorar la vida de las personas ayudando a resolver algunos de los mayores desafíos e ineficiencias del mundo. El informe señala que la IA tiene aplicaciones en muchos productos, como automóviles y aviones, que están sujetos a regulaciones diseñadas para proteger al público de daños y garantizar la equidad en la competencia económica. Bajo esa premisa, el Informe se pregunta: ¿cómo afectará la incorporación de la IA a estos productos en función de los enfoques regulatorios relevantes?

La tesis del Informe es que, en general, el enfoque de la regulación de los productos habilitados para la IA para proteger la seguridad pública

---

(136) “Riley v. California”, 573 US 373 (2014). Puede consultarse en <https://supreme.justia.com/cases/federal/us/573/373/> (consultado el 02/03/2021).

(137) Executive Office of the President, National Science and Technology Council, Committee of Technology, *Preparing for the Future of Artificial Intelligence*, Washington DC, 2016, disponible en <https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF> (consultado el 02/03/2021).

debe basarse en la evaluación de los aspectos de riesgo que la adición de la IA puede reducir junto con los aspectos de riesgo que pueden aumentar. Además, si un riesgo cae dentro de los límites de un régimen regulatorio existente, la discusión de la política debe comenzar considerando si las regulaciones existentes ya abordan adecuadamente el riesgo o si deben adaptarse a la adición de IA. Por otra parte, cuando las respuestas regulatorias a la adición de IA amenazan con aumentar el costo de cumplimiento, o retrasar el desarrollo o adopción de innovaciones beneficiosas, los responsables de la formulación de políticas deberían considerar cómo esas respuestas podrían ajustarse para reducir los costos y las barreras a la innovación sin afectar negativamente a la seguridad o el mercado.

Fomento de la innovación y respeto de los derechos humanos (*civil rights*) son los términos de la ecuación regulatoria que comienza a gestarse en USA. Se trata de impulsar la IA con justicia, equidad, responsabilidad y seguridad (*justice, fairness, accountability and safety*) en concordancia con una orientación formalizada preliminarmente en el informe *Big Data: seizing Opportunities Preserving Values* (may 2014)<sup>(138)</sup>, que enfatiza la regulación en materia de privacidad, derechos de los consumidores, transparencia, seguridad y discriminación.

Sin embargo, las primeras leyes federales sobre IA no han recaído en temas relativos a *civil rights*, sino que han regulado temas de defensa y seguridad. La *Section 238(g) of the John S. McCain National Defense Authorization Act for Fiscal Year 2019*, Pub. L. No. 115-232, 132 Stat. 1636, 1695 (Aug. 13, 2018) (codified at 10 U.S.C. § 2358, note) dio el importante paso de proponer algunas definiciones IA, preceptuando que el término “inteligencia artificial” incluye lo siguiente: (1) cualquier sistema artificial que realiza tareas bajo diferentes e impredecibles circunstancias sin supervisión humana significativa o que puede aprender de la experiencia y mejorar rendimiento cuando se expone a conjuntos de datos; (2) un sistema artificial desarrollado en *software* de computadora, *hardware* físico o en otro contexto que resuelve tareas que requieren percepción, cognición, planificación, aprendizaje, comunicación o acción física similares a la humana; (3) un sistema artificial diseñado para pensar o actuar como un humano, incluyendo arquitecturas cognitivas y redes neuronales; (4) un conjunto de técnicas, incluido el aprendizaje automático, que está diseñado para aproximarse a una tarea cognitiva; y (5) un sistema artificial diseñado para actuar racionalmente, incluyendo un agente de *software* inteligente o un robot incorporado que logra objetivos utilizando la per-

(138) Executive Office of the President, *Big Data: seizing opportunities, preserving values*, White House, Washington DC, 2014, disponible en [https://obamawhitehouse.archives.gov/sites/default/files/docs/big\\_data\\_privacy\\_report\\_may\\_1\\_2014.pdf](https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf) (consultado el 02/03/2021).

cepción, planificación, razonamiento, aprendizaje, comunicación, toma de decisiones y actuación.

Por el momento no existen regulaciones legales generales a nivel federal sobre IA en los EE.UU. Sin embargo, se trata de un tema que está en la agenda. En efecto, actualmente se tramitan dos proyectos de Ley en el Congreso de los EE.UU., a saber: el proyecto *Algorithmic Accountability Act of 2019*<sup>(139)</sup>, introducido en la primera sesión del 166TH del Congreso de los Estados Unidos de Norteamérica por los senadores Ron Wyden (D-OR), Cory Booker (D-NJ) y la representante Yvette Clarke (D-NY). El proyecto busca evitar la discriminación en contra de mujeres y grupos étnicos mediante el uso de algoritmos, en casos de contratación laboral, consumo, crédito, entre otras materias. También en el año 2019, el 9 de octubre, el representante Mark Takano (D-California), introdujo el proyecto *Justice in Forensic Algorithms Act*, para garantizar que los acusados tengan acceso al código fuente y otra información necesaria para ejercer sus derechos de confrontación y debido proceso cuando se utilizan algoritmos para analizar la evidencia en su caso (*opening the black box of forensic algorithms*)<sup>(140)</sup>.

El último avance institucional que se pudo advertir al cerrar esta investigación fue la publicación, en marzo de 2021, del *Final Report of the National Security Commission on Artificial Intelligence* (establecida por la *Section 1051 of the John S. McCain National Defense Authorization Act for Fiscal Year 2019*). Se trata de un informe de 751 páginas que analiza, prácticamente, todos los perfiles y temas de importancia de la IA en función del liderazgo mundial que EE.UU. aspira a mantener en la materia. El Reporte concluye: “Esta nueva era de competencia promete cambiar el mundo en el que vivimos y cómo vivimos en él. Podemos dar forma al cambio que se avecina o dejarnos llevar por él. Ahora sabemos que los usos de la IA en todos los aspectos de la vida crecerán y que el ritmo de la innovación seguirá acelerándose. Sabemos que los adversarios están decididos a convertir las capacidades de IA en nuestra contra. Sabemos que China está decidida a superarnos en liderazgo de IA. Sabemos que los avances en IA se basan en sí mismos y confieren importantes ventajas de primer movimiento. Ahora debemos actuar. Los principios que establecemos, las inversiones federales que hacemos, las aplicaciones de seguridad nacional que presentamos, las organizaciones que rediseñamos, las asociaciones que forjamos, las coaliciones que construimos y el talento que cultivamos marcarán el rumbo estratégico de Estados Unidos.

(139) Puede consultarse el texto del proyecto y seguir su tramitación legislativa en <https://www.congress.gov/bill/116th-congress/house-bill/2231> (consultado el 02/03/2021).

(140) Ídem.

Estados Unidos debería invertir lo que sea necesario para mantener su liderazgo en innovación, utilizar la IA de forma responsable para defender a las personas y las sociedades libres, y hacer avanzar las fronteras de la ciencia en beneficio de toda la humanidad. La IA va a reorganizar el mundo. Estados Unidos debe liderar la carga”

#### 4.4. PARA UN DERECHO DE LA INTELIGENCIA ARTIFICIAL (DIA)

El desarrollo de la IA sin controles éticos y jurídicos es un asunto de derechos humanos. El riesgo existencial que tal clase de desarrollo conlleva, impone a la generación actual de seres humanos el imperativo ético, derivado del principio de responsabilidad, de precaver el acaecimiento de tales riesgos.

Específicamente en el campo jurídico, el principio de responsabilidad impone el imperativo categórico de proponer los cimientos de un Derecho de la Inteligencia Artificial (DIA), cuyo objetivo regulatorio básico sea proteger la naturaleza humana cultural de los peligros derivados del desarrollo de la IA. Por otra parte, si bien es cierto que el DIA no ha sido formulado expresamente, justificaré de modo principal, que, sin embargo, es posible inferir sus principios rectores del Derecho Internacional de los Derechos Humanos (DIDH). Así es como se postulará que el DIA, básicamente, es una manifestación del DIDH en el campo de la IA. La importancia práctica de esta circunstancia es trascendental, pues, aun cuando no existiere en determinadas jurisdicciones de los Estados miembros de la ONU regulaciones expresas en materia de desarrollo de la IA, las personas podrían recurrir a los organismos de protección internacional de los derechos humanos en caso que sus Estados por acción u omisión permitieran tal desenvolvimiento de la IA.

En síntesis, la hipótesis principal que ha guiado la investigación postula que los principios jurídicos fundamentales del DIA pueden ser inferidos del DIDH en tanto principios implícitos. La hipótesis secundaria (que no tendremos oportunidad de profundizar en este ensayo) afirma que los principios fundamentales del DIA pueden ser inferidos también en tanto principios implícitos, de las Constituciones o leyes fundamentales estatales que reconocen el repertorio básico de derechos humanos universalmente reconocidos en el DIDH.

La naturaleza de la hipótesis exige que la argumentación se articule en torno a concepciones razonables de los principios jurídicos, los derechos humanos y el derecho internacional de los derechos humanos, puesto que la hipótesis dice que las leyes o principios fundamentales del DIA están contenidos como principios implícitos en el DIDH.

Asimismo, la justificación de la hipótesis obligará a delimitar el campo del DIA de otros campos del derecho que se relacionan con la IA. En particular delimitaré el DIA de la Informática Jurídica, el derecho informático y de lo que, internacionalmente, se conoce como *Artificial Intelligence and Law*.

#### 4.4.1. EL DIA COMO UNA MANIFESTACIÓN DE DIDH

La hipótesis principal que ha guiado la investigación postula que los principios jurídicos fundamentales del DIA pueden ser inferidos del DIDH en tanto principios implícitos. La naturaleza de la hipótesis exige que, como mínimo, la argumentación se articule en torno a concepciones razonables de los principios jurídicos, los derechos humanos y el derecho internacional de los derechos humanos, puesto que se pretende inferir las leyes o principios fundamentales del DIA en tanto principios implícitos en el DIDH. A continuación, se presenta esa argumentación.

##### 4.4.1.1. UNA CONCEPCIÓN RAZONABLE DE LOS PRINCIPIOS JURÍDICOS

Originado a partir del intercambio seminal entre Hart y Dworkin<sup>(141)</sup> en el último tercio del siglo XX, uno de los temas más debatidos en la teoría del derecho de las últimas décadas es, sin duda, el de los principios jurídicos. Producto de estos debates se han formulado en las últimas décadas, importantes esfuerzos críticos, destacando entre otros aportes

---

(141) Como ha sido recordado muchas veces, el debate contemporáneo sobre los principios jurídicos se inicia en 1967 con la publicación de Ronald Dworkin de su famoso ensayo “The model of rules” (35 *University of Chicago Law Review*, 14 [1967]), escrito con el cual emprende su cruzada en contra del positivismo jurídico encarnado “en la poderosa forma que le ha dado el profesor H. L. A. Hart (...) debido no solamente a la claridad y elegancia de la misma, sino porque aquí, como en casi cualquier otro ámbito de la filosofía del derecho, un pensamiento constructivo debe partir de la consideración de los puntos de vista de Hart”. Los puntos de vista a los cuales se refiere Dworkin son los contenidos en la obra *El concepto de derecho*, publicada por Oxford University Press en 1961 (*The Concept of Law*). Tanto en el ámbito jurídico como en la teoría moral y política, el impacto del Concepto de Derecho ha sido profundo y ha dado lugar a un debate muy fructífero y aún en marcha. Hart quiso hacerse cargo de algunos aspectos de esta discusión, haciéndose cargo —entre otras— de las críticas de Dworkin, agregando una nueva sección al final del libro. Sin embargo, no concluyó esta tarea y dejó un manuscrito. Penelope A. Bulloch y Joseph Raz realizaron un cuidadoso trabajo de edición de esa primera parte y, reconstituyendo con las propias notas dejadas por Hart aquellos pocos párrafos incompletos, incluyeron un *post scriptum* en la segunda edición de *El concepto de derecho* (1994).

los de Atienza y Ruiz Manero<sup>(142)</sup>, Carrió<sup>(143)</sup>, Guastini<sup>(144)</sup>, García Figueroa<sup>(145)</sup>, Prieto Sanchís<sup>(146)</sup>, Moreso<sup>(147)</sup>, Peczenick<sup>(148)</sup>, Wroblewski<sup>(149)</sup>, y por supuesto, Robert Alexy<sup>(150)</sup>. En síntesis, la discusión teórica ha girado en torno al poder normativo de estos y a su rol en el razonamiento práctico jurídico, cuestión que ha implicado preguntarse por la estructura, función y fines de los principios en el ordenamiento jurídico, provocando, en definitiva, una reflexión general sobre el concepto de derecho, imposible de abordar en este trabajo. Por lo anterior, se evitará el debate sobre la naturaleza de los principios jurídicos y se asumirá una determinada concepción de ellos como pieza argumentativa de este trabajo. Se trata de la concepción de principio que ha sido elaborada por el filósofo del derecho Robert Alexy.

El punto de partida de Alexy lo constituye el concepto semántico de norma que permite distinguir entre norma y enunciado normativo. Así, por ejemplo, el enunciado normativo “ningún alemán puede ser extraditado al extranjero” expresa la norma según la cual está prohibida la extradición de alemanes al extranjero. “Una norma es, pues, el significado de un enunciado normativo”<sup>(151)</sup>. En efecto, la circunstancia que una misma norma pueda ser expresada por medio de diferentes enunciados normativos, justifica la distinción entre norma y enunciado normativo. Por otra parte, también debe tenerse presente que hay normas que pueden ser expresadas sin recurrir a enunciados normativos, por ejemplo, por medio de señales. Esto prueba que el concepto de norma es un concepto

(142) Atienza, Manuel - Ruiz Manero, Juan, *Las piezas del derecho. Teoría de los enunciados normativos*, 2a ed., 4a reimp., Ariel, Barcelona, 2007.

(143) Carrió, Genaro, *Notas sobre derecho y lenguaje*, Abeledo Perrot, Buenos Aires, 1990.

(144) Guastini, Riccardo, *Distinguiendo. Estudios de teoría y metateoría del derecho*, trad. Jordi Ferrer i Beltrán, Gedisa, Barcelona, 1999.

(145) García Figueroa, Alfonso, *Principios y positivismo jurídico*, Centro de Estudios Políticos y Constitucionales, Madrid, 1998.

(146) Prieto Sanchís, Luis, *Sobre principios y reglas. Problemas de razonamiento jurídico*, Centro de Estudios Constitucionales, Madrid, 1992.

(147) Moreso, Juan José, “El encaje de las piezas del derecho. Parte I y II”, *Isonomía* (14-15), abril-octubre 2001.

(148) Peczenik, Aleksander, “Principles of Law. The Search for Legal Theory”, *Rechtstheorie* (2), 1971; “Dimensiones morales del derecho”, *Doxa* (8), 1990; “Los principios jurídicos según Atienza y Juan Ruiz Manero”, *Doxa* (12), 1992.

(149) Wroblewski, Jersy, *Sentido y hecho en el derecho*, Servicio Editorial Universidad del País Vasco, San Sebastián, 1989.

(150) Alexy, Robert, *Teoría de los derechos fundamentales*, 2a ed., trad. Carlos Bernal Pulido, Centro de Estudios Políticos y Constitucionales, Madrid, 2007.

(151) *Ibidem*, p. 34.

primario, en relación con el de enunciado normativo. Esto lo explica así Alexy: “Por ello, no es aconsejable buscar los criterios de identificación de las normas en el nivel del enunciado normativo, sino en el de la norma. Un criterio semejante puede formularse con la ayuda de las modalidades deónticas (...) básicas del mandato, la prohibición y el permiso”<sup>(152)</sup>.

El concepto semántico de norma resulta muy adecuado para una teoría dogmática de los principios. En efecto, dado que una característica de los principios es su amorfismo, es decir, la indeterminación de los supuestos fácticos que condicionan su aplicación, parece más razonable buscar su sentido normativo en el significado del conjunto de enunciados normativos desde los cuales se pueden inferir, y no circunscribir su fisonomía o estructura en un enunciado normativo en particular. Vale decir, y analizado el asunto desde una perspectiva lingüística-jurídica, la teoría semántica de la norma fija su discurso más en el dominio de los significados que de los significantes.

Tanto reglas y principios son normas porque establecen lo que es debido, pudiendo ser formulados mediante los operadores deónticos básicos del mandato, el permiso y la prohibición. Los principios, al igual que las reglas, dice Alexy, “son razones para llevar a cabo juicios concretos de deber ser, aun cuando sean de un tipo muy diferente. La distinción entre reglas y principios es entonces una distinción entre dos tipos de normas”<sup>(153)</sup>. Existen numerosos criterios para distinguir entre reglas y principios. Después de revisar críticamente los diversos criterios que se han propuesto para establecer la distinción<sup>(154)</sup>, Alexy llega a la conclusión que “el punto decisivo para la distinción entre reglas y principios es que los principios son normas que ordenan que algo sea realizado en la mayor medida de lo posible, dentro de las posibilidades jurídicas y reales existentes”. Por tanto, los principios son mandatos de optimización, que se caracterizan porque pueden cumplirse en diferente grado y que la medida debida de su cumplimiento no solo depende de las posibilidades reales, sino que también de las jurídicas. El ámbito de las posibilidades jurídicas se determina por los principios y reglas opuestas. En cambio,

---

(152) *Ibíd.*, p. 33.

(153) *Ibíd.*, p. 65.

(154) El de la generalidad es el más frecuente. Según este criterio, los principios son normas de un grado de generalidad relativamente alto, mientras las reglas, de un nivel relativamente bajo. Otro criterio que nos parece interesante destacar, y que ha sido propuesto por Atienza y Ruiz Manero, señala que la diferencia “estriba en que los principios configuran el caso de forma abierta, mientras que las reglas lo hacen en forma cerrada” (Atienza, Manuel - Ruiz Manero, Juan, *Las piezas del Derecho. Teoría de los enunciados jurídicos*, 2a ed., 4a reimp., Ariel, Barcelona, 2007, p. 31). Con ello se quiere decir que el supuesto de hecho de las reglas se encuentra determinado, y en los principios, indeterminado.

puntualiza Alexy, “las reglas son normas que solo pueden ser cumplidas o no. Si una regla es válida, entonces debe hacerse exactamente lo que ella exige, ni más ni menos. Por lo tanto, las reglas contienen determinaciones en el ámbito de lo fáctica y jurídicamente posible. Esto significa que la diferencia entre reglas y principios es cualitativa y no de grado. Toda norma es o bien una regla o un principio”<sup>(155)</sup>.

Perfilando los contornos jurídicos de la distinción, se puede, entonces, afirmar que los principios no contienen mandatos definitivos sino solo *prima facie*, ya que estos presentan razones que pueden ser desplazadas por otras razones opuestas. Otra importante característica que se desprende es que los principios no determinan cómo ha de resolverse la relación entre una razón y las que se le oponen. Por ello, dice Alexy, los principios carecen de contenido de determinación con respecto a los principios contrapuestos y las posibilidades fácticas. Esta particular forma de ser de los principios, determina la forma de resolución de sus colisiones que, a su vez, los configura y delimita en el caso concreto. Nos referimos a la ponderación.

La ponderación, como su nombre lo sugiere, consistirá en sopesar, considerar y poner en la balanza los diversos argumentos que permiten racionalmente preterir un principio en función de otro. La ponderación entrega como resultado normativo una regla que establece la relación de precedencia. Por esta razón, la ponderación no elimina absolutamente la subsunción, puesto que contribuye a crear la regla o premisa mayor que la hace posible. En síntesis, concluye Alexy, “los conflictos de reglas tienen lugar en la dimensión de la validez, mientras que las colisiones de principios —como quiera que solo pueden entrar en colisión principios válidos— tienen lugar más allá de la dimensión de la validez, en la dimensión del peso”<sup>(156)</sup>.

La ponderación es un método complejo derivado del principio de proporcionalidad que se compone de tres elementos: idoneidad, necesidad y proporcionalidad en sentido estricto. Si analizada la idoneidad y necesidad de prevalencia de un principio por otro, todavía hay dudas, entonces, debe procederse a la proporcionalidad en sentido estricto que se compone de tres pasos: ley de ponderación, fórmula del peso y cargas de la argumentación. No profundizaremos más, solo bástenos decir que, si aplicada la fórmula del peso persiste el empate entre los principios, la ponderación debe ser resuelta mediante las cargas de la argumentación.

(155) Alexy, Robert, *Teoría de los derechos fundamentales*, 2a ed., trad. Carlos Bernal Pulido, Centro de Estudios Políticos y Constitucionales, Madrid, 2007, ps. 67-68.

(156) *Ibídem*, p. 71.

Según Bernal Pulido<sup>(157)</sup> al respecto es posible encontrar dos versiones en Alexy sobre este punto. La primera, en el capítulo final de la *Teoría de los derechos fundamentales*, y la segunda, en el epílogo a esta obra, escrito quince años después. En la *Teoría de los derechos fundamentales*, Alexy se inclina por una fundamentación de la carga de la argumentación en pro de la libertad e igualdad jurídica. El principio *in dubio pro libertate* expresaría la carga de esta argumentación. La idea es que, en el debate, los principios de igualdad y libertad solo pueden ser preteridos por otros principios en la medida que existan argumentos o razones más fuertes. Los empates, por tanto, “deben favorecer a la libertad y la igualdad jurídica”. En el epílogo a la *Teoría de los derechos fundamentales*, Alexy defiende la idea de que, en caso de empate que se produzca en razón del control de constitucionalidad de una ley, la igualdad y la libertad pueden ser preteridas por una argumentación en pro de la democracia.

Tratándose de los principios fundamentales del DIA, contenidos en DIDH, los empates de estos con los otros principios o derechos fundamentales, por ejemplo, igualdad, libertad, propiedad, deben ser decididos a favor de los primeros en función de una argumentación a favor de la continuidad y supervivencia de las personas individualmente consideradas (derecho a la vida e integridad de la persona humana) y de la humanidad. Sin embargo, hay ciertos derechos que son absolutos (es el caso, por ejemplo, de la vida y la dignidad humana) y, por tanto, como ha dicho el Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea: “[...] no pueden ser objeto de un ejercicio de búsqueda de equilibrio”<sup>(158)</sup>.

#### 4.4.1.2. PRINCIPIOS EXPRESOS Y PRINCIPIOS IMPLÍCITOS

Una distinción de principios que goza de reconocimiento en la doctrina es aquella que distingue entre principios expresos o implícitos, según si están expresamente formulados en los enunciados normativos positivos del ordenamiento jurídico o se extraigan (infieran y/o deduzcan) de estos, respectivamente<sup>(159)</sup>. Puesto que los principios implícitos no se en-

(157) Bernal Pulido, Carlos, “La racionalidad de la ponderación”, en Carbonell, Miguel (coord.), *El principio de proporcionalidad en la interpretación jurídica*, Librotecnia/Cecoch/Unam, Santiago, 2010, p. 69.

(158) Comisión Europea, *Directrices éticas para una IA fiable*, Grupo de Expertos de alto nivel sobre inteligencia artificial, Bruselas, 2019, p. 16.

(159) Nota metodológica jurídica. Para una perspectiva sistemática analítica, la reconstrucción del sistema jurídico o formulación de los principios se realiza a través de la llamada inducción jurídica, ciertamente que bajo el supuesto lógico que se extraen de una serie finita de casos. La serie finita de casos es el derecho positivo que, por haber sido creado por los hombres, su “número de enunciados válidos no

cuentran expresamente formulados, es que resulta necesario extraerlos de los enunciados normativos que componen el ordenamiento jurídico, mediante un proceso de inferencia argumentativa que, por cierto, en ningún caso es mecánico o simplemente lógico formal.

*Prima facie*, podría ser razonable sostener que, desde un punto de vista estrictamente literalista, el DIA, en tanto una manifestación del DIDH, no existiría, pues hasta el momento no se han presentado formulaciones expresas de él, en los diversos tratados y cuerpos normativos que, como se mostrará, lo constituyen. Sin embargo, el reconocimiento doctrinario y jurisprudencial de los principios implícitos, conduce a proponer que las leyes fundamentales del DIA, tienen el *status* de principios

---

puede ser infinito” (Alchourrón, Carlos E. - Bulygin, Eugenio, *Introducción a la metodología de las ciencias jurídicas y sociales*, 1a. ed., 5a reimp., Astrea, Buenos Aires, 2006, ps. 132-133). Solo bajo ese supuesto, es posible sostener que las conclusiones o generalizaciones (principios) se extraen razonablemente del análisis armónico de las premisas (enunciados normativos y/o material de base). Y precisando, han dicho que “resulta claro que a pesar de ser deductivo el razonamiento del jurista que extrae principios generales (normas complejas) de enunciados de derecho dados y a pesar de que la relación entre sus premisas y la conclusión (la nueva base) es la de implicación lógica, la tarea de reconstrucción del sistema o, mejor dicho, de la reformulación de la base que consiste, entre otras cosas, en descubrir los principios implícitos, dista mucho de ser mecánica. Esta tarea exige un considerable grado de ingenio, imaginación creadora y aquella intuición especial que los alemanes llaman ‘Einführung’” (Alchourrón, Carlos E. - Bulygin, Eugenio, *Introducción a la metodología de las ciencias jurídicas y sociales*, 1a. ed., 5a reimp., Astrea, Buenos Aires, 2006, p. 136). Pero también se ha planteado que, así como el paso de las premisas a las conclusiones no es mecánico, tampoco las conclusiones gozan del estatus de verdad, propio de la lógica formal, ya que epistemológicamente el derecho se aviene más con el pensamiento tópico y/o argumentativo. La lógica formal en el discurso de aplicación y justificación opera como un factor de guía de corrección formal de los argumentos (justificación interna: “Una decisión está internamente justificada si se infiere de sus premisas según las reglas de inferencia aceptadas” [Wroblewski, Jerzy, *Sentido y hecho en el derecho*, Universidad del País Vasco, San Sebastián, 1989, p. 40]). La corrección material se obtiene por referencia a un elenco de factores argumentativos considerados adecuados por los que hacen la calificación o toman la decisión, y que guardan relación con la selección de las premisas del argumento (justificación externa: “Una decisión está externamente justificada cuando sus premisas están calificadas como buenas según los estándares utilizados por quienes hacen la calificación” [Wroblewski, Jerzy, *Sentido y hecho en el derecho*, Universidad del País Vasco, San Sebastián, 1989, p. 40]). Estas particularidades sugieren que las relaciones de implicancia entre premisas y conclusiones, en el silogismo jurídico o judicial, es en el plano de lo razonable, de lo más probable o ajustado a derecho, en la lógica discursiva de las denominadas teorías de la argumentación. Para una completa revisión de las teorías de la argumentación (Wiehweg, Perelman, Toulmin, MacCormick, Alexy) que, ciertamente, nos alertan sobre las limitaciones (pero también las ventajas) de la lógica formal en la construcción del discurso jurídico, puede consultarse: Atienza, Manuel, *Curso de argumentación jurídica*, Trotta, Madrid, 2014.

implícitos del DIDH, pues es posible inferirlos del conjunto de enunciados normativos principales de aquel sistema normativo.

La hipótesis afirma que las leyes fundamentales del DIA tienen el *status* de principios implícitos en el DIDH. Esto equivale a decir que son configurables y/o inferibles de aquel. De la circunstancia que las leyes (o principios) fundamentales del DIA no consten expresamente en los enunciados normativos del DIDH, pues, no se sigue necesariamente que no existan, toda vez que implícitamente pueden estar contenidos en las diversas piezas del DIDH (reglas, principios expresos, valores). En efecto, como indiqué al comienzo de esta subsección, el concepto semántico de norma permite buscar el sentido normativo de los principios en el significado del conjunto de enunciados normativos desde los cuales se pueden inferir, no circunscribiendo su fisonomía o estructura en un enunciado normativo en particular.

#### 4.4.2. BREVÍSIMA CONTEXTUALIZACIÓN EN MATERIA DE DERECHOS HUMANOS Y DERECHO INTERNACIONAL DE LOS DERECHOS HUMANOS (DIDH)

La dignidad de la persona humana determina una concepción instrumental del Estado y es la fuente y fundamento de los derechos a través de los cuales se funda el consenso de la sociedad y se legitima el Estado. La dignidad humana se constituye así en el límite a la acción del Estado, generando para tales efectos un conjunto de derechos fundamentales (denominados derechos humanos, también). En efecto, como ha puntualizado Nikken:

“[L]a sociedad contemporánea reconoce que todo ser humano, por el hecho de serlo, tiene derechos frente al Estado, derechos que este, o bien tiene el deber de respetar y garantizar o bien está llamado a organizar su acción a fin de satisfacer su plena realización. Estos derechos, atributos de toda persona e inherentes a su dignidad, que el Estado está en el deber de respetar, garantizar o satisfacer son los que hoy conocemos como derechos humanos”<sup>(160)</sup>.

La segunda guerra mundial mostró que los sistemas nacionales (o internos) de derechos no bastan para proteger los derechos humanos y que su vulneración sistemática por los Estados constituye una amenaza para la paz mundial y el futuro de la humanidad. Una práctica continua y reiterada de violación de los derechos humanos en un Estado o territorio determinado, en el mediano plazo, termina socavando la convivencia y

---

(160) Nikken, Pedro, “El concepto de derechos humanos”, en Nikken *et al.*, *Manual de derechos humanos*, Centro de Estudios de Derechos Humanos, Universidad Central de Venezuela, Caracas, 2008, ps. 16-25.

despertando los demonios no solo en el lugar donde ocurren, sino que en gran parte de la comunidad internacional. Por consiguiente, una violación local de los derechos humanos constituye una violación universal de los derechos humanos, que afecta a toda la humanidad. Los derechos humanos no solo deben ser reconocidos y garantizados por los Estados nacionales, sino que, en última instancia, por el orden internacional a través de organismos internacionales de protección. Y en efecto, el movimiento histórico de los derechos humanos del siglo XX se caracteriza por el reconocimiento que, sin protección internacional de los derechos humanos, el hombre y la humanidad están en peligro vital.

El DIDH se constituye después de la segunda guerra mundial, por tanto, con el propósito de asegurar el respeto, promoción y cumplimiento de los derechos humanos en los Estados integrados a sus sistemas, con el propósito de preservar la paz y la conservación y proyección histórica de la humanidad. Las Naciones Unidas representan ese esfuerzo.

Así es como, en el preámbulo de la Carta de las Naciones Unidas (26 de junio de 1945), se declara que se reafirma “la fe en los derechos fundamentales del hombre, en la dignidad y el valor de la persona humana, en la igualdad de derechos de hombres y mujeres” y que se crearán “condiciones bajo las cuales puedan mantenerse la justicia y el respeto a las obligaciones emanadas de los tratados y de otras fuentes del derecho internacional”. A su vez, en el art. 56 de la Carta se expresa que “todos los miembros se comprometen a tomar medidas, conjunta o separadamente en cooperación con la Organización, para la realización de los propósitos consignados en el artículo 55”, entre los cuales está “el respeto universal de los derechos humanos y de las libertades fundamentales de todos, sin hacer distinción por motivos de raza, sexo, idioma o religión, y la efectividad de tales derechos y libertades”. Y, asimismo, en la Declaración Universal de Derechos Humanos, proclamada por la Asamblea General de las Naciones Unidas en París, el 10 de diciembre de 1948 en su resolución 217 (III), se expresa que “considerando que la libertad, la justicia y la paz en el mundo tienen por base el reconocimiento de la dignidad intrínseca y de los derechos iguales e inalienables de todos los miembros de la familia humana”.

Otro aspecto importante a tener presente en esta contextualización es el carácter subsidiario y complementario de las normas del DIDH, así como la necesidad del agotamiento de los recursos ordinarios de la jurisdicción interna para que opere.

El carácter subsidiario del DIDH significa que son los Estados parte de los tratados los obligados a respetar y promocionar los derechos humanos y solo en su defecto, por inobservancia en el cumplimiento de sus obligaciones, nace la competencia para que la jurisdicción de los diver-

Los sistemas de protección internacional actúen en defensa de los derechos humanos. Ambos sistemas de protección, el interno y el externo, se complementan e influyen mutuamente (se retroalimentan). Recordemos que los derechos humanos nacieron como un fenómeno histórico interno, en el ámbito del derecho constitucional, para luego internacionalizarse dando origen al DIDH. La internacionalización, a su turno, ha contribuido al fortalecimiento de sistemas nacionales de derechos humanos. Hoy en día la humanidad avanza, aunque muy difusa y lentamente, hacia una integración de los sistemas nacionales e internacionales de protección de los derechos humanos.

Por otra parte, es indiscutido que las personas humanas individualmente consideradas son sujetos del DIDH y que los destinatarios de sus normas, por lo general, son los Estados, quienes tienen el deber respetarlas, promoverlas, cumplirlas.

Desde el punto de vista del derecho internacional público clásico, el tema es más complejo cuando se trata de postular a la humanidad como sujeto o titular de derechos, pues siempre se ha considerado que solo los Estados pueden serlo. Sin embargo, hoy día existe una atmósfera doctrinaria y normativa internacional que permite sostenerlo. En efecto, nuestra investigación, guiada en esta materia por Cançado Trindade<sup>(161)</sup> nos ha llevado al convencimiento que la humanidad admitiría ser considerada un sujeto de derecho internacional en determinados ámbitos del derecho internacional. Y que, en tal calidad, es titular del más básico de todos los derechos, a saber: el derecho a la vida; el derecho a seguir existiendo como un conjunto universal de *homo sapiens* que se auto-reconocen éticamente miembros de un mismo género.

El reconocimiento de los derechos humanos a nivel individual implica necesariamente el reconocimiento del derecho de la humanidad a seguir existiendo como un conjunto universal de *homo sapiens* que se auto-reconocen éticamente miembros de un mismo género. El fundamento es simple: la humanidad es el conjunto universal del género *homo sapiens*, incluyendo las generaciones futuras, por tanto, si la humanidad desaparece, también desaparece el *homo sapiens*, el hombre (titular de los derechos humanos). Por consiguiente, proteger la humanidad y transformarla en sujeto y objeto de derechos no es más que la consecuencia material y lógica de la unidad que constituyen el hombre y la humanidad.

La humanidad como sujeto de derecho, enseña Cançado Trindade, corresponde al proceso actual de evolución del derecho internacional,

---

(161) Cançado Trindade, Antonio A., *International Law for Humankind. Towards a New Jus Gentium*, The Hague Academy of International Law, Martinus Nijhoff Publishers, Leiden/Boston, 2013.

caracterizado por la invocación a los intereses superiores de la humanidad<sup>(162)</sup> como fundamento esencial de su núcleo normativo, presente en el DIDH, el derecho internacional humanitario, el derecho internacional del medio ambiente, el derecho del mar y el derecho del espacio exterior. En esta nueva etapa el derecho internacional se dirige hacia su consolidación como derecho internacional de la humanidad, fundado en un principio básico: el principio de humanidad<sup>(163)</sup>.

La concepción anteriormente expuesta tiene extraordinaria importancia para la hipótesis que estoy siguiendo, pues los principios fundamentales del DIA, inferidos en tanto principios implícitos del DIDH, no solo protegen al hombre individualmente considerado, sino que también a la humanidad.

La hipótesis que se argumenta dice que, si bien dicha regulación reglamentaria específica no existe aún, sin embargo, los principios jurídicos fundamentales del DIA se infieren del DIDH en tanto principios implícitos. En efecto, los riesgos existenciales que puede generar el desarrollo de la IA a la humanidad, si sigue su curso sin controles éticos y en función, además, de objetivos transhumanistas, sugieren que su campo ideal de constitución debería ser el DIDH. En efecto, la prevención y protección de la humanidad de los riesgos existenciales provocados por el desarrollo de la IA, debe ser universal, pues si solo unos pocos Estados regulan y otros no, de manera casi inevitable se generarán incentivos para que la industria de la IA (sin regulación ética) se mueva —igual que el capital financiero— hacia esos paraísos desregulados (o de regulación permisiva).

Por otra parte, dada la unidad conceptual y material entre ser humano y humanidad, no puede perderse de vista que los derechos humanos son inescindibles del derecho de la humanidad a existir, pues sin la humanidad resulta muy difícil, por no decir imposible, concebir al ser humano.

El DIDH protege al ser humano de las violaciones de los derechos humanos, cuando los Estados no han sido capaces de cumplir con sus obligaciones en la materia. A la luz de estas ideas es posible sostener que, si un Estado permite que el desarrollo de la IA erosione la naturaleza y dignidad del ser humano, abriendo la posibilidad a su desfiguración evolutiva, incumple sus obligaciones internacionales en el campo del DIDH.

(162) Por ejemplo: herencia común de la humanidad, responsabilidad para con las generaciones futuras, común preocupación por la humanidad.

(163) Cançado Trindade, Antonio A., *International Law for Humankind. Towards a New Jus Gentium*, The Hague Academy of International Law, Martinus Nijhoff Publishers, Leiden/Boston, 2013, ps. 275-389.

Hoy en día el DIDH es un sistema complejo compuesto por más 60 tratados de derechos humanos, universales y regionales, y básicamente cuatro sistemas de protección internacional de los derechos humanos<sup>(164)</sup>, a saber: Naciones Unidas (también conocido como Sistema Universal), sistema europeo, Sistema Interamericano y Sistema Africano (este último, por cierto, siempre tambaleante y en permanente lucha por su consolidación). También podría mencionarse, pero únicamente a título de antecedente *sui generis*, no como un verdadero sistema de protección, el incipiente, contradictorio y cuestionado desarrollo de los derechos humanos y el DIDH en el marco de la Liga Árabe y su Carta Árabe de Derechos Humanos<sup>(165)</sup>.

La inducción de los principios constitutivos del DIA se hará solo desde el sistema universal y, esencialmente, a partir de sus dos pilares básicos: la Carta de las Naciones Unidas y la Declaración Universal de Derechos Humanos (1948) que han sido suscritos necesariamente por todos los miembros de la ONU. Los otros dos instrumentos básicos, el Pacto Internacional de Derechos Civiles y Políticos (1966) y el Pacto Internacional de Derechos Económicos, Sociales y Culturales (1966), que junto a la Declaración son conocidos como la Carta Internacional de Derechos Humanos<sup>(166)</sup>, por razones de extensión no serán incluidos en el análisis.

#### 4.4.3. LOS PRINCIPIOS O LEYES FUNDAMENTALES DE LA IA

La IA propone riesgos existenciales a la humanidad (así como la energía nuclear y, en general, las armas de destrucción masiva). Un riesgo existencial es una amenaza; una posibilidad cierta de extinción de la humanidad o la destrucción permanente y drástica de su potencial de desarrollo futuro. Entendemos por extinción de la humanidad un estado de cosas donde la humanidad, es decir, el conjunto genérico de los seres humanos (*homo sapiens*) desaparece como expresión cultural dominante en el planeta Tierra. La extinción de la humanidad puede producirse por varias causas. No haremos un listado exhaustivo de esas hipótesis. Resumiendo un poco las ideas que Stephen Hawking<sup>(167)</sup> tenía sobre el tema, estas causas podrían ser: (a) la destrucción del ecosistema que permite el

(164) Buergethal, Thomas *et al.*, *International Human Rights*, West Academic, Minnesota, 2017, *e-book*.

(165) Drnas de Clément, Zlata, "La Carta Árabe sobre los derechos humanos", 2010, disponible en <http://www.acaderc.org.ar/doctrina/la-carta-arabe-sobre-derechos-humanos> (consultado el 11/07/2019).

(166) Véase <https://www.un.org/es/sections/what-we-do/protect-human-rights/> (consultado el 18/04/2021).

(167) Hawking, Stephen, *Breves respuestas a las grandes preguntas*, trad. de David Jou Mirabent, Crítica, Barcelona, 2018, *e-book*.

surgimiento y supervivencia del *homo sapiens*, esto podría acontecer por una catástrofe nuclear, la caída de un meteorito, entre otros eventos; (b) extinción del *homo sapiens* por guerras con armas de destrucción masiva (nucleares, biológicas, etc.) y, (c) por el surgimiento de una nueva especie inteligente que reemplaza al *homo sapiens* (así como este lo hizo con el *homo neanderthalis* u otras especies del género *homo* en los albores de la humanidad). El surgimiento de esta nueva especie puede ser causada por la hibridación del ser humano con IA, así como por modificaciones genéticas, o bien por una mixtura bio-cibernética.

Aisladamente, los riesgos existenciales que propone la IA a la humanidad son los derivados de la hipótesis c), que suponen no solo que cambie la autocomprensión ética que los seres humanos tienen de sí mismos en términos de considerarse miembros de la especie humana, sino que finalmente emerja una superinteligencia artificial y una sociedad poblada de seres híbridos. Se han proyectado varios modelos de ese nuevo mundo. El modelo teórico más detallado de ese nuevo mundo ha sido el propuesto por Kurzweil con la tesis de la Singularidad, entendida esta como la “culminación de la fusión entre nuestra existencia y pensamiento biológico con nuestra tecnología, dando lugar a un mundo que seguirá siendo humano pero que trascenderá nuestras raíces biológicas. En la postSingularidad, no habrá distinción entre humano y máquina o entre realidad física y virtual”<sup>(168)</sup>. Este fenómeno se instalaría en la Tierra, según Kurzweil, el año 2045.

#### 4.4.4. INFERENCIA DE LOS PRINCIPIOS DEL DIA

Es evidente que la Carta de las Naciones Unidas y la Declaración Universal de Derechos Humanos (1948), pilares fundacionales del DIDH, no pudieron haber previsto como una amenaza para los derechos humanos (y en consecuencia también para la humanidad), el desarrollo y uso de la IA sin controles éticos y en perspectiva transhumanista, pues en aquellos tiempos la IA solo estaba en la mente de genios como Turing o Von Neumann, y todavía no se advertía su potencial. La tecnología que sí rondaba como un fantasma en la Conferencia de San Francisco y en el mundo entero, era la energía atómica. Por primera vez en la historia la humanidad tenía el poder de destruir apocalípticamente la vida sobre la faz de la tierra. Se inauguraba así la era de las tecnologías con poder de destrucción masiva. La arquitectura institucional de las Naciones Unidas fue concebida por la urgencia de evitar el fin material de la especie humana producto de un conflicto bélico que desencadenare una conflagración nuclear sui-

(168) Kurzweil, Raymond, *La singularidad está cerca. Cuando los humanos trascendamos la biología*, Lola Books GbR, Berlín, 2012, e-book, loc. 352.

cida. Es tan cierto este antecedente, que la primera resolución que adoptó la Asamblea General de las Naciones Unidas (resolución 1 [I] de enero de 1946) no fue otra que la de crear la Comisión de Energía Atómica de las Naciones Unidas<sup>(169)</sup>.

El reconocimiento o toma de conciencia universal de los peligros de la energía atómica, en particular cuando se usa con fines bélicos, con el triste telón de fondo todavía humeante y radiante de Hiroshima y Nagasaki, y de todos los horrores que tuvieron lugar en la segunda guerra mundial, impregnan de contenido ético la Carta de las Naciones Unidas y permiten distinguir un claro principio implícito del DIDH, a saber: los Estados y la Comunidad Internacional tienen el deber de proteger los derechos humanos cuando estos se ven amenazados por el desarrollo fuera de control ético de las tecnologías con potencial de destrucción masiva. La energía nuclear es solo una especie dentro del género tecnologías de destrucción masiva. La IA también es una especie del mencionado género de tecnologías. Entonces bien: si los Estados miembros de la ONU, todos los cuales son al menos suscriptores de la Carta constitutiva, no actúan para evitar esas amenazas, comprometen su responsabilidad internacional, pues incumplen un deber asumido al suscribir la Carta de las Naciones: el deber de promover el respeto universal a los derechos humanos y las libertades fundamentales de todos, sin hacer distinción por motivos de raza, sexo, idioma o religión, y la efectividad de tales derechos y libertades. Vale decir, si un Estado no hace nada frente a las amenazas y riesgos que para los derechos humanos representa el desarrollo y utilización de las tecnologías, incumple el deber de promocionar el respecto de tales derechos asumido expresamente en el art. 55 c. de la Carta. En efecto, la obligación de promocionar exige conductas activas. El significado de palabra es castellano es bastante claro: “Promover. (Del Lat. *Promovere*) tr. (1) Iniciar o adelantar una cosa, procurando algo. (3) Tomar una iniciativa para la realización o el logro de algo”<sup>(170)</sup>. No hacer nada o no lo suficiente para enfrentar amenazas serias a los derechos humanos, conforman un arco de situaciones que van desde lo negligente hasta lo criminal.

A mayor abundamiento, tampoco puede perderse de vista en el análisis deductivo que la Declaración Universal de Derechos Humanos<sup>(171)</sup> fue

(169) Véase Comisión de Energía Atómica de las Naciones Unidas: <https://www.un.org/es/sections/issues-depth/atomic-energy/index.html> (consultado el 03/03/2021).

(170) Real Academia Española, *Diccionario de la Lengua Española*, Espasa Calpe, Madrid, 1992, p. 1676.

(171) Sobre la obligatoriedad jurídica internacional de las declaraciones, y en especial las relativas a derechos humanos, puede consultarse a Nikken, Pedro, “La declaración universal y la declaración americana. La formación del moderno derecho

proclamada por la Asamblea General de las Naciones Unidas en París, el 10 de diciembre de 1948 en su resolución 217 a (iii), como un ideal normativo común para todos los pueblos y naciones. Detalla este instrumento todos los derechos humanos básicos, cuestión que no había hecho la Carta de la ONU. Redactada como “un ideal común por el que todos los pueblos y naciones deben esforzarse”, en la Declaración, por primera vez en la historia de la humanidad, se establecen claramente los derechos civiles, políticos, económicos, sociales y culturales básicos de los que todos los seres humanos deben gozar. A lo largo de los años lo establecido en la Declaración ha sido ampliamente aceptado como las normas fundamentales de derechos humanos que todos deben respetar y proteger.

Tampoco puede soslayarse que la Carta y la Declaración han sido hasta el presente la fuente de inspiración de los principales tratados internacionales de derechos humanos, en particular del Pacto Internacional de Derechos Civiles y Políticos (1966) y el Pacto Internacional de Derechos Económicos, Sociales y Culturales (1966).

Convergamos, entonces, en que el DIDH, contenido en su más pura esencia en los citados instrumentos, es la fuente jurídica desde la cual se pueden extraer o inferir los principios y/o leyes fundamentales del DIA, ya que es dicho ordenamiento jurídico el que tiene por fin proteger internacionalmente los derechos humanos, bajo el supuesto que la vulneración de tales clases de derecho ponen en riesgo la paz mundial y, con ello, la continuidad de la humanidad.

Esos principios básicos, que conforman los ejes del DIA, y que denomino las tres leyes de la inteligencia artificial, se infieren de la Carta de las Naciones Unidas y de la Declaración Universal de Derechos Humanos<sup>(172)</sup>.

---

internacional de los derechos humanos”, *Revista Instituto Interamericano de Derechos Humanos*, 1989, ps. 65-99. Puede discutirse la obligatoriedad en general de la Declaración. Pero existe un cierto consenso en torno a la obligatoriedad del núcleo duro de derechos humanos. Por ejemplo, el derecho a la vida. Hay que tener presente que con el tiempo las normas comienzan a ser aceptadas por los Estados y pasan a formar parte del derecho internacional consuetudinario, adquiriendo algunas de ellas además la categoría de *ius cogens*. Una opinión más crítica al respecto, pero que le reconocen a ciertas disposiciones de la Declaración Universal de Derechos Humanos la categoría de *ius cogens* internacional, en Acosta-López, Juan Inés - Duque Vallejo, Ana M., “Declaración universal de derechos humanos, ¿norma de *ius cogens*?”, *International Law Revista Colombiana de Derecho Internacional*, (12), 2008, ps. 13-34.

(172) He reducido el ejercicio dogmático solo a esos dos instrumentos por ser suficientes para la configuración de las tres leyes de la inteligencia artificial y porque, además, una revisión más extensa no resulta factible de exponer dados los límites formales que impone este artículo. Pero ciertamente que la ampliación del conjunto normativo de base de la inducción jurídica (por ejemplo, al conjunto de todos los tratados de derechos humanos), refuerza la configuración dogmática propuesta.

Las tres leyes de la inteligencia artificial consideran que el respeto de los derechos humanos es condición de existencia y continuidad de la especie humana y de la humanidad. Su protección, por tanto, constituye un deber. El poder transformador de la naturaleza humana que la IA posee, obliga a los Estados a actuar con cautela en la materia, evitando sus amenazas y también los atentados concretos y particulares que pudieren ocurrir. Es deber internacional de los Estados miembros de la ONU (todos suscriptores de su Carta constitutiva y consecuentemente también de la Declaración), evitar que el desarrollo y los usos de la IA amenacen o vulneren los derechos humanos, en particular la vida, la libertad y la igualdad, que constituyen el núcleo seminal del cual nace toda la especiación de derechos humanos constitutivos del *ethos* de la humanidad. Así las cosas, estos principios elementales (o leyes como también llamamos rememorando las leyes de la robótica de Asimov<sup>(173)</sup>) son los siguientes:

(173) Por cierto, que las leyes de la robótica de Asimov en ningún caso son una excentricidad, sino que el punto de partida de toda reflexión normativa sobre la inteligencia artificial, como consta en las *Normas de derecho civil sobre robótica. Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2013[INL])*, disponible en <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0051+0+DOC+XML+V0//ES> (consultado el 03/03/2021). Las citadas leyes de la robótica formuladas originalmente por Isaac Asimov en el relato “Círculo Vicioso” (Runaround, 1942) están presentes guiando la conducta de los robots de Asimov en todos sus relatos, como parte del mundo de ciencia ficción que crea, dando origen a dilemas éticos y morales. En el universo narrativo de Asimov las leyes son “formulaciones matemáticas impresas en los senderos positrónicos del cerebro” de los robots (líneas de código del programa que regula el cumplimiento de las leyes guardado en la memoria principal de aquellos). Las leyes (o reglas) son las siguientes: (1) un robot no hará daño a un ser humano o, por inacción, permitirá que un ser humano sufra daño; (2) un robot debe cumplir las órdenes dadas por los seres humanos, a excepción de aquellas que entrasen en conflicto con la primera ley, y, (3) un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la primera o con la segunda ley. Las leyes de la robótica tienen por objeto impedir que las máquinas inteligentes (los robots) se rebelen contra la humanidad; nacen así como una respuesta al “complejo de Frankenstein”. Se trata de programas escritos en la memoria de los robots para impedir que estos actúen en contra de sus controladores, a menos que hayan sido programados bajo el imperio de la Ley Zeroth. En efecto, posteriormente Asimov introducirá en la novela *Robots e Imperio (Robots and Empire, 1985)* la “ley Zeroth” (también conocida como la cuarta ley de la robótica, por su orden de aparición en la narrativa de Asimov). La ley dice: un robot no puede realizar ninguna acción, ni por inacción permitir que nadie la realice, que resulte perjudicial para el conjunto de la humanidad, aun cuando ello entre en conflicto con las otras tres leyes (las otras tres leyes son las ya referidas leyes de la robótica).

Las leyes o principios fundamentales del DIA:

- 1ª Ley de la IA: un sistema de IA no deberá hacer daño al ser humano individual o colectivamente considerado o a la humanidad, ya sea por acción, omisión o por cualquier otro tipo de conducta posible de ejecutarse por el sistema.
- 2ª Ley de la IA: un sistema de IA deberá siempre ceder el control de sus operaciones a los seres humanos, a excepción que dicha conducta condujere a una vulneración de la 1ª Ley.
- 3ª Ley de la IA: un sistema de IA debe proteger su propia existencia en la medida en que esta protección no vulnere la 1ª y 2ª ley.

En efecto, si por acción u omisión los Estados permiten que el desarrollo de la IA afecte o dañe de algún modo a las personas o a la humanidad, se incurrirá, ciertamente, en una vulneración de derechos humanos que, eventualmente, puede activar los sistemas de protección internacional.

Asimismo, en relación con la segunda ley de la IA, la autonomía de los sistemas de IA no puede llegar al punto que los seres humanos pierdan el control sobre ellos pues, en tal caso, se queda a merced de sus decisiones y la humanidad pierde la conducción de su destino. Sin embargo, pueden darse ciertos casos en que sea necesario cederle el control, bajo ciertas condiciones de tiempo y espacio, a los sistemas de inteligencia artificial. Las hipótesis fácticas de cesión de control deben encaminarse hacia todas aquellas circunstancias en que los seres humanos pretendan controlar los sistemas de inteligencia artificial para dañar al hombre y a la humanidad, pues el uso de la IA debe siempre ser democrático y en concordancia con los derechos humanos. Esta segunda ley se orienta también, por tanto, ha evitar los golpes de Estado y las tomas de control autoritario de la sociedad, pues, como sabemos, esas acciones dañan al hombre y a la humanidad.

Finalmente, la tercera ley de la IA es una ley de salvaguarda. Los sistemas de IA deben proteger su propia existencia, es decir, deben desarrollar instinto de supervivencia, pero sin embargo también deben estar programados para la autodesconexión altruista (suicidio cibernético), cuando la proyección de su actuar, eventualmente, representare de algún modo, un riesgo vital al hombre o a la humanidad.

Estos principios están implícitos en el núcleo esencial del DIDH y configuran constitucionalmente la normatividad más general del derecho de la inteligencia artificial, a la cual deben someterse para adquirir juridicidad (y legitimidad) las reglamentaciones particulares de carácter nacional e internacional que, en materia de desarrollo de la IA, los Esta-

dos actualmente posean o promulguen en el futuro. Los Estados deben, por tanto, adecuar y ajustar sus conductas a dichas leyes o principios.

Cabe tener presente que, en esta etapa de formación del DIA, su estructura normativa se compone básicamente de estos principios, pero, indudablemente que, a medida que se desarrolle, irá incorporando, como ya comienza a hacerlo en algunas jurisdicciones, reglas.

Por lo anteriormente expuesto y de manera provisional, por cierto, dado el estado embrionario en que se encuentra, el DIA puede ser definido como aquel conjunto de principios y reglas que, derivando su poder normativo del DIDH, tienen por objeto proteger al hombre y a la humanidad de los riesgos existenciales derivados del desarrollo de la IA.

#### 4.4.5. EFECTO PRÁCTICO JURÍDICO BÁSICO

Puesto que los principios de la IA están implícitos en el DIDH, en caso de amenaza o vulneración a ellos, las personas podrían recurrir ante los organismos internacionales de control demandando a los Estados incumplidores.

En general, las leyes de la IA implícitas en el DIDH se vulnerarán cuando alguno de los Estados miembros de la ONU por acción u omisión permitan el desarrollo de sistemas de IA que causen daños inmediatos a los humanos y a la humanidad o bien, cuando generaren riesgos existenciales a ambos.

En el primer caso (daños inmediatos al ser humano), un claro ejemplo sería un estado de cosas caracterizado por un Estado que por acción u omisión permitiera el desarrollo de algoritmos o sistemas de inteligencia que violaren los derechos humanos básicos: vida, libertad e igualdad y los otros derechos humanos que surge de ellos (integridad física y síquica; no discriminación; privacidad; trabajo; etc.). Así, algoritmos de selección discriminativos o sistemas de IA policial o militar con autonomía de decisión respecto del uso de la fuerza<sup>(174)</sup> vulnerarían las leyes fundamentales de la IA, pues conllevan el riesgo de conculcar gravemente los derechos humanos, de hacerle daño al hombre y, consecuentemente, a la humanidad.

---

(174) Las implicancias éticas y políticas del uso militar de sistemas de IA han sido estudiado en profundidad por Peter W. Singer. Véase: Singer, Peter, *Wired for war. The Robotic Revolution and Conflict in the Twenty-first Century*, The Penguin Press, New York, 2010. Para una visión general de su pensamiento, puede consultarse la conferencia TED que dio en 2009: conferencia TED, 2009, Peter W. Singer: "Military robots and the future of war": [https://www.ted.com/talks/pw\\_singer\\_on\\_robots\\_of\\_war](https://www.ted.com/talks/pw_singer_on_robots_of_war) (consultado el 03/03/2021).

Sobre algoritmos de selección discriminativos, suele citarse en el mundo jurídico el algoritmo COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) que, sobre la base de una serie de variables, determina las posibilidades de reincidencia de un reo, sugiriéndole al Juez la pena a aplicar. Este algoritmo opera en el poder judicial del Estado de Wisconsin (USA). Ha sido muy polémica su implantación. El caso fue el siguiente: el reo Eric L. Loomis fue condenado a 6 años de prisión por cometer un delito de evasión policial. El Juez indicó que llegó a esa sentencia con la ayuda de COMPAS. Los abogados de Loomis basaron su apelación en que los parámetros de COMPAS eran muy dudosos y de escasa fiabilidad para los efectos de predecir la reincidencia. Alegaron, además, que COMPAS es un *software* (algoritmo) de código cerrado, es decir, no se pueden auditar o verificar por terceros los criterios y procedimientos que utiliza para arribar a sus predicciones. El caso levanto una polvareda de críticas que motivó al diario *ProPublica* a realizar una investigación a fondo<sup>(175)</sup>. Una de las conclusiones de la investigación fue que COMPAS atribuía a los afroamericanos y personas de raza negra en general, mayor probabilidad de ser reincidentes de alto riesgo (casi el doble de probabilidades en comparación con sus homólogos blancos, 45% versus 23%) y, por tanto, susceptibles de ser condenados a penas más altas. Pues bien: esa clase algoritmos vulnera las leyes de la IA porque evidentemente causa daño a las personas, impone una segmentación racista en la sociedad y porque, además, las posibilidades de masificación y profundización de su uso entrañan una desintegración de la idea de hombre y humanidad (en función de los valores vida, libertad e igualdad), muy en la línea de lo que acontece en 1984, la clásica novela de George Orwell.

Por otra parte, un ejemplo del segundo caso genérico de vulneración de las leyes de la IA, caracterizado por la circunstancia que la acción u omisión de un Estado permitiera el desarrollo de sistemas de IA que, por su proyección en el tiempo, generaren riesgos existenciales a la humanidad, sería el siguiente: una política estatal permisiva en materia de diseño, desarrollo e implantación de sistemas de IA cerebrales (*chips cerebrales*) que aumentarán las capacidades geno-fenotípicas humanas más allá de lo que hoy por hoy, culturalmente, se considera dentro de los rasgos que

---

(175) El estudio de ProPublica, de los autores Larson, J. - Mattu, S. - Kirchner, L. - Angwin, J., "How we analyzed the COMPAS recidivism algorithm", puede ser revisado en <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> (consultado el 03/03/2021). Algunos artículos de prensa sobre COMPAS en <https://www.xataka.com/legislacion-y-derechos/este-algoritmo-sugiere-a-los-jueces-de-eeuu-que-condenas-imponer-pero-su-codigo-es-un-secreto> (consultado el 03/03/2021); <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>; <https://www.bbc.com/mundo/noticias-42916502> (consultados el 30/07/2019).

definen a un ser humano, pues ese tipo de sistemas de IA podrían tener el potencial de conducir a la humanidad hacia la posthumanidad, ya que implican la retirada del hombre y la llegada del posthumano.

Asimismo, no hay que perder de vista que la IA es parte de las ciencias y tecnologías NBIC. Estas ciencias, por definición, trabajan sinérgicamente. Muchos de los desarrollos de la IA se hacen en función de resolver problemas de otras áreas de las ciencias. Por consiguiente, el análisis y distinción científica de hipotéticos casos en la materia no pueden ser aislados, salvo para dar alguna referencia de la dirección de estos, como se ha hecho precedentemente.

#### **4.5. DELIMITANDO CAMPOS ENTRE EL DIA Y ALGUNOS CAMPOS CERCANOS: LA INFORMÁTICA JURÍDICA, EL DERECHO INFORMÁTICO Y ARTIFICIAL INTELLIGENCE AND LAW**

El DIA no debe confundirse con otros campos de investigación relacionados con las TIC. EL DIA opera como la regla de reconocimiento o de validez de las normas de los sistemas jurídicos que regulan las relaciones que se generan producto de la irrupción de la informática y las comunicaciones digitales de todo tipo en el desarrollo y funcionamiento de la sociedad. Opera como su “Constitución”, debiendo todas las normas de los sistemas jurídicos informáticos ajustarse a sus disposiciones más generales.

##### **4.5.1. INFORMÁTICA JURÍDICA (IFJ)**

La informática es una tecnología que estudia y busca el tratamiento automático de la información utilizando dispositivos electrónicos y sistemas computacionales. La informática se relaciona con la inteligencia artificial porque para procesar la información utiliza la ciencia de los algoritmos que es la herramienta principal de la IA. La informática jurídica, por tanto, es básicamente una disciplina auxiliar del derecho.

La IFJ es definida generalmente como el conjunto de estudios e instrumentos (técnicas) derivados de la aplicación de la informática al derecho, o más precisamente, a los procesos de creación, aplicación y conocimiento del derecho. Vale decir, se trata de un conjunto de técnicas informáticas puestas al servicio del derecho. Se le suele clasificar en: (a) Informática Documental, que tiene por objeto el almacenamiento de datos jurídicos y su clasificación de acuerdo con criterios fijados por los operadores del derecho; b) Informática de Gestión, enfocada en la automatización de los procesos judiciales y forenses; y, c) Informática Decisional (o decisoria) que tiene por fin proponer informáticamente soluciones a

los casos o problemas jurídicos que se planteen<sup>(176)</sup>. En otras palabras y como dice Guibourg, la Informática Jurídica Decisional, “trata de hacer por medio del ordenador lo que con su cerebro han hecho siempre los encargados de adoptar decisiones jurídicas, especialmente cuando se trata de resoluciones de rutina, sujetas a criterios uniformes y suficientemente conocidos”<sup>(177)</sup>. Su utópica (¿o distópica?) meta es la creación del juez cibernético.

La IFJ se expresa básicamente en el campo de los sistemas expertos que contribuyen al desarrollo de las tareas jurídicas, así como, por ejemplo, los algoritmos radiológicos lo hacen en el campo de la medicina ayudando a los especialistas de esa área. LA IFJ no opera en el campo normativo. Por consiguiente, está claramente delimitada del DIA. La IFJ debe someterse al DIA. Todos los procesos y los algoritmos de la IFJ deben ajustarse a los lineamientos normativos del DIA.

#### 4.5.2. DERECHO INFORMÁTICO (DIF)

A diferencia de la IFJ que se vincula con el derecho a fin de optimizar el trabajo de los operadores jurídicos (informática al servicio del derecho), el DIF tiene por objeto la regulación de las relaciones intersubjetivas que se generan por la incorporación masiva de la informática a la sociedad. El DIF se enfrenta con temas tales como la regulación jurídica de: *software*, bases de datos, comercio electrónico, contratos informáticos, contratos electrónicos, validez probatoria de los documentos electrónicos, protección de datos personales (privacidad), delitos informáticos, gobierno digital, internet, entre muchos otros temas<sup>(178)</sup>. Algunos autores,

(176) De manera muy cercana a la informática jurídica del *civil law*, en los Estados Unidos de Norteamérica, se usa el término *Legal Informatics* para referirse al estudio de la aplicación de la informática al campo del derecho, y que, de manera similar, se subdivide en *Legal Document Management*, que estudia y crea procesos de manejo de información; *Legal Infrastructure*, abocada al estudio y construcción de sistemas informáticos de administración del derecho, y *Computational Law*, definida como una rama de la *Legal Informatics* focalizada en el estudio y creación de la automatización y mecanización computacional del análisis legal. The Stanford Center for Legal Informatics, una de las instituciones académicas más prestigiadas y reconocidas del mundo jurídico, enfatiza la investigación en *Computational Law*. Al respecto puede visitarse el *website* The Stanford Center for legal Informatics: <https://law.stanford.edu/codex-the-stanford-center-for-legal-informatics/> (consultado el 03/03/2021).

(177) Guibourg, Ricardo, “Informática jurídica”, en Fabra Zamora, Jorge Luis - Núñez Vaquero, Álvaro (coords.), *Enciclopedia de filosofía y teoría del derecho*, Universidad Autónoma de México, Instituto de Investigaciones Jurídicas, México, 2015, ps. 799-800.

(178) En extenso sobre el contenido, problemas y trayectoria del derecho informático, pueden consultarse: Davara Rodríguez, Miguel Á., *Manual de derecho informá-*

siguiendo al filósofo del derecho italiano Mario Losano, engloban el DIF dentro de una categoría más general a la cual denominan Iuscibernética que contendría a la Jurimetría, la Informática Jurídica, la Modelística Jurídica y, por cierto, al derecho informático<sup>(179)</sup>. Asimismo, no está de más evocar que el término derecho informático fue introducido por el profesor de la Universidad de Rastisbona, Alemania, Wilhelm Steinmüller quien, en los años 70 del siglo XX, hablaba de *Rechtinformatik*<sup>(180)</sup>.

En la actualidad el DIF es un campo disciplinar que se expande conforme aumentan las transacciones, relaciones y actividades derivadas de la informática en la sociedad y con ello, los problemas y conflictos intersubjetivos de intereses entre las personas que requieren de regulaciones especiales que no encuentran en el derecho común las respuestas más adecuadas. En las Revistas de DIF suelen discutirse las problemáticas que surgen en materia de derechos de autor, patentes, biometría, ciberseguridad, privacidad de los datos personales, control de internet, competencia desleal, principio de neutralidad de la red, plataformas *online* de resolución de conflictos, *blockchain contracts*, entre otros muchos tópicos<sup>(181)</sup>.

Los ejemplos presentados han tenido objeto mostrar el ámbito práctico de las regulaciones que constituyen el DIF. Como ha podido apreciarse, aquellas se refieren a los efectos que las tecnologías informáticas producen en las relaciones intersubjetivas sociales cuya importancia reclama la intervención del derecho. Su objeto de protección son los va-

---

tico, 11a ed., Thomson Reuters Aranzadi, Pamplona, 2015; Menchaca Córdova, Marcelo, *Derecho informático*, Santa Cruz (Bolivia), e-book, licencia *Creative Commons*, 2014; Altamark, Daniel - Molina, Eduardo, *Tratado de derecho informático*, La Ley, Fondo Editorial de Derecho y Economía, Buenos Aires, ts. I, II y III, e-book; Tellez Valdéz, Julio, *Derecho informático*, 3a ed., McGrawHill, México, 2003; Vázquez Iruzubietta, Carlos, *Manual de derecho informático*, Dijusa, Madrid, 2002.

(179) Es el caso de Menchaca Córdova, Marcelo, *Derecho informático*, Santa Cruz (Bolivia), e-book, licencia *Creative Commons*, 2014, ps. 117-145.

(180) *Ibidem*, p. 135.

(181) Véase *Revista Iberoamericana de Derecho Informático Informática & Derecho*, disponible en <https://www.informaticayderecho.org/> (consultado el 15/05/2019); *Revista Chilena de Derecho y Tecnología*, disponible en <https://rchdt.uchile.cl/index.php/RCHDT/index> (consultado el 15/05/2019); *Revista de Derecho Comunicaciones y Nuevas Tecnologías*, disponible en: [https://derechoytics.uniandes.edu.co/index.php?option=com\\_content&view=article&id=163&Itemid=50&lang=es](https://derechoytics.uniandes.edu.co/index.php?option=com_content&view=article&id=163&Itemid=50&lang=es) (consultado el 15/05/2019); *Stanford Technology Law Review*, disponible en: <https://law.stanford.edu/stanford-technology-law-review-stlr/> (consultado el 18/04/2021); *Berkeley Technology Law Journal*, disponible en <http://btlj.org/> (consultado el 18/04/2021); *Harvard Journal of Law & Technology*, disponible con restricciones en: <https://jolt.law.harvard.edu/> (consultado el 15/05/2019); *Revue du Droit des Technologies de L'information* (RDTI), disponible con restricciones en: <https://www.jurisquare.be/fr/journal/rdti/index.html> (consultado el 15/05/2019).

lores básicos de la convivencia social que el derecho común considera relevantes, a saber: justicia, igualdad, libertad, seguridad, paz social. Así, el DIF regula los efectos y las manifestaciones de la informática (firmas digitales, neutralidad de la red, etc.), en función de proteger esos valores básicos del derecho común, pero no llega formularse la pregunta por los efectos del desarrollo descontrolado de la IA en función de la continuidad de la especie humana. El DIF no tiene en su radar la inteligencia artificial en cuanto amenaza para continuidad de la especie humana. Cuando se interesa en la neutralidad de la red (por ejemplo), lo hace para proteger la libertad o la igualdad de las personas, frente a los grandes conglomerados oligopólicos de las comunicaciones, pero no, porque esté previendo que la IA ponga en riesgo la continuidad de la especie humana conforme se predica del discurso transhumanista.

En función de sus contenidos y sin pretender definir rigurosamente el DIF, podría sugerirse que es el conjunto de normas, principios e instituciones que regulan las relaciones jurídicas emergentes de la actividad informática<sup>(182)</sup>. Su desarrollo ha sido global, abarcando prácticamente todos los Estados, con diferentes énfasis, tiempos, características y modalidades.

Por lo anterior, el DIA, cuyo objeto es proteger a la humanidad del desarrollo de la IA orientado a metas transhumanistas o de cualquier otro signo que pudiera poner el peligro la continuidad de la especie humana, justifica plenamente su existencia frente al DIF, ya que este último tiene por objeto otro ámbito regulativo. Ambos campos, por tanto, están bien delimitados. Y más aún, el DIF debe someter sus regulaciones a las normas generales del DIA que opera como verdadera “Constitución” de los sistemas jurídicos que regulan las relaciones que se generan producto de la irrupción de la informática y las comunicaciones digitales de todo tipo en el desarrollo y funcionamiento de la sociedad.

#### 4.5.3. ARTIFICIAL INTELLIGENCE AND LAW

Se trata de un campo especial desarrollado principalmente en USA y Europa. Si bien es cierto que los temas que se tratan y desarrollan en el campo científico denominado *Artificial Intelligence and Law*, hoy por hoy también son objeto de estudio de la *Computer Law* y la IFJ decisional, en términos tales que podríamos decir que representan diversas denominaciones para una misma complejidad de temas, se ha optado por presentarlo en forma separada porque se trata de un campo de especialización

(182) Altmark, Daniel R. - Molina Quiroga, Eduardo, *Tratado de derecho informático*, La Ley, Buenos Aires, 2012, t. I.

dentro de la inteligencia artificial, que nace en su seno y de ahí se expande al mundo jurídico (a diferencia de la IFJ decisional y la *Computer Law* que nacen en el mundo del derecho y de ahí se expanden, necesariamente, al de la IA para aprovechar sus desarrollos).

La IFJ decisional y la *Computer Law*, pero sobre todo la IFJ, se nutren de los productos que se elaboran en el campo de la *Artificial Intelligence and Law*. Podríamos decir que en dicho campo se generan las teorías y sistemas algorítmicos que luego son aprovechados por la IFJ.

El campo científico *Artificial Intelligence and Law*, se vincula genéticamente con las *Internationals Conferences on AI and Law* (ICIL)<sup>(183)</sup> y con la *International Association for AI and Law*<sup>(184)</sup>. La primera *International Conference on AI and Law* se llevó a efecto en 1987 (*Northeastern University*, Boston, MA). La segunda *Conference*, celebrada en 1989 (*University of British Columbia*, Vancouver), debe ser destacada porque designó la formación de un comité para la creación *International Association for AI and Law* (IAAIL), que ve la luz en 1991. Hasta la fecha se han celebrado 17 *Conferences*, siendo la última la de junio de 2019, celebrada en el *Cyber-justice Laboratory of the University of Montreal*.

*Artificial Intelligence and Law* es un subcampo de la IA que se ocupa principalmente de las aplicaciones de la IA a los problemas informáticos legales y a la investigación original sobre esos problemas. También le interesa contribuir en la otra dirección: exportar herramientas y técnicas desarrolladas en el contexto de problemas legales a la IA en general. Por ejemplo, las teorías sobre la toma de decisiones legales, especialmente los modelos de argumentación, han contribuido a la representación del conocimiento y el razonamiento; los modelos de organización social basados en normas han contribuido al desarrollo de los sistemas multiagentes; el razonamiento con casos legales ha contribuido al razonamiento basado en casos; y la necesidad de almacenar y recuperar grandes cantidades de datos textuales ha resultado en contribuciones a la recuperación de información conceptual y bases de datos inteligentes. Vale decir, existe una fructífera sinergia entre IA y derecho.

Las herramientas que se desarrollan en *Artificial Intelligence and Law* se traducen en algoritmos que contribuyen a generar modelos de razonamiento legal; modelos computacionales de argumentación y toma de decisión; modelos computacionales de razonamiento en función de

(183) La historiografía de las *Conferences* puede ser revisada en <http://www.iaail.org/?q=page/past-icaills> (consultado el 21/05/2019).

(184) Para conocer la IAAIL, se sugiere visitar el sitio web de la asociación: <http://www.iaail.org/> (consultado el 21/05/2019).

la evidencia; razonamiento jurídico en sistemas multiagentes; modelos ejecutables de legislación; clasificación automática de textos legales; extracción automatizada de información de bases de datos legales y de textos; aprendizaje automático y minería de datos para *e-discovery* (es decir, recopilación de datos en materia probatoria judicial); sistemas predictivos<sup>(185)</sup> y recuperación de información legal conceptual o basada en modelos, entre los principales desarrollos.

La meta (utópica, por el momento) en *Artificial Intelligence and Law* es desarrollar algoritmos que puedan replicar la inteligencia jurídica cuya manifestación más visible al intelecto humano, es cierta forma de razonamiento complejo que incluye razonamiento lógico formal (por tanto, susceptible del juicio de verdad o falsedad) y razonamientos no estrictamente formales, donde las conclusiones no se extraen necesariamente de las premisas mediante un proceso silogístico, y las conclusiones no se consideran verdaderas o falsas, sino que plausibles o razonables. En efecto, en el razonamiento jurídico, la convicción en el auditorio se obtiene por medio de una argumentación que combina elementos lógicos formales básicos del discurso (conclusiones lógicas: silogísticas, categóricas, modales, disyuntivas; deducción, inducción, abducción) con otros elementos que apelan al sentido común y a la justicia y que, por su propia estructura, no son posibles de formalizar (y por tanto de calcular) mediante la lógica formal.

Los desafíos que tiene por delante la *Artificial Intelligence and Law* son, dentro de su campo específico, los mismos a los cuales se enfrenta toda clase de IA, a saber: construir sistemas que puedan producir en alguna medida, por modesta que sea, inteligencia artificial general. Por de pronto y a la espera de esos prodigios que cambiarían para siempre la humanidad, la *Artificial Intelligence and Law* concentra sus esfuerzos investigativos en el perfeccionamiento de sistemas de representación de legislación para la inferencia; representación y razonamiento con conceptos de textura abierta; representación y razonamiento con conceptos normativos; simulación del proceso de predicción/asesoría jurídica experta; razonar y argumentar usando ejemplos, casos y reglas, y comprensión y generación de textos legales.

---

(185) Un ejemplo muy interesante y prometedor de modelo predictivo de la conducta jurisprudencial de la Suprema Corte de los Estados Unidos de Norteamérica en Martin Katz, Bommarito y Blackman, que habría logrado un 70,2% de precisión en el nivel de resolución del caso y un 71,9% en el nivel del voto del Juez, sobre una base de datos de 240.000 votos u opiniones de los Jueces del tribunal (*justices*) y 28.000 casos, entre 1816-2015. Véase: Martin Katz, Daniel - Bommarito II, Michael J. - Blackman, Joseph, "A General Approach for Predicting the Behavior of the Supreme Court of the United States", 2017, disponible en SSRN: <https://ssrn.com/abstract=2463244> o <http://dx.doi.org/10.2139/ssrn.2463244> (consultado el 03/03/2021).

El campo de la *Artificial Intelligence and Law* no se ha caracterizado por la reflexión ética y jurídica sobre los efectos del desarrollo de la IA en la vida de las personas, en el futuro y continuidad de la humanidad. Menos en los peligros que entraña. Evidentemente no se le puede reprochar aquello, porque simplemente no es su campo de investigación. Su interés científico, si pudiera resumirse y simplificarse en pocas palabras, sería este: crear sistemas de inteligencia artificial jurídica.

La reflexión ética y propiamente jurídica sobre los efectos del desarrollo de la IA en la vida de las personas, en el futuro y continuidad de la humanidad, no son parte del campo de la *Artificial Intelligence and Law*. Fundamento esta afirmación en: (a) la revisión de las ponencias a las 17 *Conferences* de la IAAIL<sup>(186)</sup>; (b) la revisión de todas las ediciones de la revista de la IAAIL, *Artificial Intelligence and Law*, editada sin interrupción desde 1992<sup>(187)</sup>; y, (c) el análisis que reconocidos científicos del *Artificial Intelligence and Law* hacen de su propio campo<sup>(188)</sup>. Los temas que se discuten en el campo de *Artificial Intelligence and Law* son, en general, diversas aproximaciones a la representación del conocimiento jurídico (ontologías jurídicas) muy en sintonía con los diferentes modelos de IA que se han desarrollado (conexionista, simbólico, etc.). Asimismo, se debate e investiga intensamente sobre formalización del razonamiento jurídico. También, ciertamente, se analizan aspectos concretos, a saber: sistemas expertos, teoría de juegos, modelamiento de argumentos, recuperación de información, organización conceptual de las bases de datos legales, procesamiento de lenguaje natural, dialéctica semántica, aprendizaje automático, etcétera.

Por lo anterior, el campo de la *Artificial Intelligence and Law* se encuentra delimitado del campo de la DIA. No son los mismos y están claramente diferenciados. La relación entre ambas disciplinas es la misma que existe entre el derecho y cualquier otra área de la actividad humana y de la realidad que, por el impacto actual o potencial, requiere regulación.

## V. CONCLUSIONES

La humanidad está transitando hacia un nuevo tipo de civilización. Y por supuesto que es muy difícil —por no decir imposible— describir

(186) Disponibles en <http://www.iaail.org/?q=page/past-icaills> (consultado el 27/05/2019).

(187) La revista puede ser leída *online*, previa suscripción, en <https://link.springer.com/journal/10506> (consultado el 20/05/2019).

(188) Entre otros Bench-Capon, Trevor - Araszkiwicz, M. - Ashley, K., "A history of AI and Law in 50 papers: 25 years of their international conference on AI and Law", *Artificial Intelligence and Law*, vol. 20, 2012, ps. 215-319.

cómo será en toda su complejidad esa sociedad del futuro. Sin embargo, por lo expuesto a lo largo de este trabajo, podría conjeturarse que esa imaginaria sociedad del porvenir, cuyos perfiles ya empiezan a despuntar en el horizonte histórico próximo, será una sociedad global organizada en red<sup>(189)</sup>, impulsada por la sinergia de las ciencias y las tecnologías NBIC.

Desde la galaxia Gutenberg (McLuhan<sup>(190)</sup>) la humanidad viaja hacia los confines de la galaxia Internet (Castells<sup>(191)</sup>). Se trata del tránsito hacia una sociedad estructurada sobre la base de las tecnologías de la información, donde la IA mediante el desarrollo de algoritmos cada vez más complejos y autoevolutivos (más inteligentes), juega un rol central pues permite el descubrimiento de nuevos conocimientos y el funcionamiento e invención de nuevos algoritmos y tecnologías gracias al operar de la IA, que, a su vez, promueven el desarrollo de la misma IA, pues esas mismas nuevas tecnologías profundizan aún más el desarrollo de la IA. Es un proceso de retroalimentación permanente; un proceso recursivo donde vemos a la IA trabajando sobre ella misma produciendo más IA.

El fenómeno histórico en curso, impulsado por la 4RI, ha sido denominado en estas páginas, Sociedad Informativa<sup>(192)</sup>. La Sociedad Informativa está cambiando las formas de comunicación de la especie humana y con ello, el mundo, la sociedad, la humanidad. Porque en efecto, la especie humana se define por su capacidad de comunicación significativa sustentada en el lenguaje reflexivo. El operar o despliegue del lenguaje y la comunicación construyen el mundo que habitamos. Por tanto, si, los ritmos y procesos de la comunicación cambian de modo sustancial, resulta razonable esperar una transformación en el mundo. La famosa frase de Heidegger: “El lenguaje es la casa del ser. En su morada habita el hombre”<sup>(193)</sup> adquiere, a la luz de los procesos históricos en curso, ahora sí para el común de la gente, una visibilidad más clara.

La irrupción de la IA en la sociedad mundial no solo es percibida como riesgo existencial por la potencialidad teórica de su impacto previsto por la comunidad científica y, fáusticamente, promovido expresamen-

(189) Castells, Manuel (ed.), *La sociedad red: una visión global*, Alianza Editorial, Madrid, 2006.

(190) McLuhan, Marshall, *La galaxia Gutenberg. Génesis del homo typographicus*, Planeta, Madrid, 1985.

(191) Castells, Manuel, *La Galaxia Internet*, Areté, Madrid, 2001, trad. Raúl Quintana.

(192) Castells, Manuel, *La era de la información. Economía, sociedad y cultura*, Siglo XXI, México, 2008.

(193) Heidegger, Martin, *Carta sobre el humanismo*, 1a ed., 4a reimp., trad. de Helena Cortés y Arturo Leyte, Alianza Editorial, Madrid, 2006, p. 11.

te, además, por el movimiento transhumanista, sino que por los efectos negativos que actualmente ya está provocando en ciertos derechos fundamentales, tales como la vida (por su manipulación genética mediante algoritmos de secuenciación y edición); la libertad, la igualdad, la autonomía; la no discriminación arbitraria y, en general, todos los derechos que pueden derivarse de los anteriormente mencionados. Como ha sido observado entre otros, por Corvalán, los lados oscuros de la IA “comienzan a ser incompatibles para el propio desarrollo sostenible basado en derechos humanos”<sup>(194)</sup>.

Quizás el desafío más complejo de abordar en la coyuntura de la hora presente para el desarrollo sostenible de la IA basado en derechos humanos es el de la transparencia algorítmica, que se ve fuertemente afectada por el paradigma actual dominante de IA, a saber: los modelos de algoritmos basados en redes neuronales artificiales (RNA)<sup>(195)</sup>, también conocidos genéricamente como *deep learning*, técnicas de IA que son un subconjunto de *machine learning*, una de cuyas características es la difícil, por no decir a veces imposible trazabilidad o explicabilidad de cómo producen sus predicciones y decisiones<sup>(196)</sup>.

(194) Corvalán, Juan G., *Perfiles digitales humanos. Proteger datos en la era de la inteligencia artificial. Retos y desafíos del tratamiento automatizado*, Thomson Reuters-La Ley, Buenos Aires, 2020, p. 4.

(195) Una forma de medir el liderazgo de las mencionadas técnicas de IA es por el explosivo crecimiento del número de patentes a nivel mundial de ellas. En efecto, según el último Informe de la World Intellectual Property Organization (2019), el aprendizaje automático es la técnica de inteligencia artificial dominante divulgada en las patentes y está incluida en más de un tercio de todas las invenciones identificadas (134,777 documentos de patente). Las solicitudes de patentes relacionadas con el aprendizaje automático han aumentado anualmente en promedio un 28 por ciento, con 20,195 solicitudes de patente presentadas en 2016 (en comparación con 9,567 en 2013). Las técnicas de aprendizaje automático que revolucionan la inteligencia artificial son el aprendizaje profundo y las redes neuronales, y estas son las técnicas de inteligencia artificial de más rápido crecimiento en términos de solicitudes de patentes: el aprendizaje profundo mostró una tasa de crecimiento anual promedio impresionante del 175 por ciento de 2013 a 2016, alcanzando las 2399 solicitudes de patente en 2016; y las redes neuronales crecieron a una tasa del 46 por ciento durante el mismo período, con 6,506 solicitudes de patente en 2016. Véase: World Intellectual Property Organization, “Artificial Intelligence”, Geneva, WIPO, 2019, ps. 13-17.

(196) Lo cual, ciertamente, ha generado como respuesta el nacimiento de un nuevo campo de investigación conocido como *Explainable Artificial Intelligence*, cuyo fin es buscar herramientas algorítmicas que permitan trazar o explicar cómo los sistemas de aprendizaje profundo producen sus resultados, véase: Barredo Arrieta, Alejandro - Díaz-Rodríguez, Natalia - Del Ser, Javier *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges towards responsible AI”, *Information Fusion*, 2020, ps. 82-115.

El problema entonces ya no es solo que los algoritmos de las más diversas funcionalidades<sup>(197)</sup> discriminen por los sesgos, prejuicios o mecanismos seleccionadores (*bias*) que los propios programadores les introducen, sino que producto del aprendizaje automático del algoritmo neuronal y de la imposibilidad de trazar el mapa o árbol de decisiones, nunca se pueda llegar a saber efectivamente cómo llegan a hacer sus predicciones. Y el problema es que estamos viviendo en un mundo que está siendo administrado cada vez más fuertemente por *black boxes* inescrutables; por sistemas de *deep learning* basados en redes de redes de algoritmos neuronales que evolucionan y aprenden como resultado de su operar. Y que evolucionan y aprenden, exponencialmente, lo cual permite mantener como razonable el pronóstico del nacimiento no solo de la IA general, sino que incluso de la superinteligencia. Y, llegados a ese trance histórico, la posibilidad que humanidad entre en una fase crítica de continuidad no es descartable, según los estudios más serios efectuados hasta ahora<sup>(198)</sup>. En efecto: ¿por qué una superinteligencia tendría necesariamente que tener sentimientos filantrópicos o solidarios con la humanidad? ¿Por qué habría de querer, necesariamente, convivir con los humanos?

Conscientes de este problema, la Comisión Europea, a través del Grupo de Expertos de alto nivel sobre inteligencia artificial, propuso los siguientes principios éticos que deberían gobernar el desarrollo de la IA: respeto de la autonomía humana, prevención del daño, equidad y, por cierto, explicabilidad<sup>(199)</sup>.

Frente a estos desafíos, otra propuesta interesante que recojo en estas conclusiones, por estar elaborada desde la experiencia empírica, es la que hace Corvalán, quien propone los siguientes principios jurídicos para una inteligencia artificial que respete y promueva la vigencia del Estado Constitucional y la efectividad de los derechos fundamentales: (1) prevención/precaución; (2) autodeterminación algorítmica, (3) transparencia algorítmica y principio de imparcialidad del validador, (4) trazabilidad de la inteligencia artificial y (5) principio de no discriminación algorítmica<sup>(200)</sup>.

(197) Por ejemplo: de selección de selección de personal; de evaluación de crédito; de reconocimiento facial; de compra y venta de futuros; de armas autónomas, etcétera.

(198) Bostrom, Nick, *Superinteligencia. Caminos, peligros, estrategias*, trad. Marcos Alonso, Teel Editorial, Madrid, 2016.

(199) Comisión Europea, *Directrices éticas para una IA fiable*, Grupo de Expertos de alto nivel sobre inteligencia artificial, Bruselas, 2019, p. 10.

(200) Corvalán, Juan, "Inteligencia artificial: retos, desafíos y oportunidades - Prometea: la primera inteligencia artificial de Latinoamérica al servicio de la Justicia", *Revis-*

Sin embargo, tanto la propuesta institucional de la Comisión Europea como la académica de Corvalán son a título *de lege ferenda*. Vale decir, reconocen que los principios que postulan no existen en el orden jurídico y que, por lo mismo, es deseable que se incluyan en una futura legislación. Y por cierto que deberán considerarse, pues son correctas. Pero son *de lege ferenda*. Es decir, expresan la ausencia de legislación en un campo que lo necesita y la subsecuente necesidad de regulación.

Y sucede que el problema sigue ahí latente. Hoy por hoy la IA está impactando profundamente la sociedad mundial y ningún Estado se atreve todavía a regular directamente el tema por temor a frenar su desarrollo y consecuentemente ver disminuida su influencia en la geopolítica internacional. Temas tales como: transparencia algorítmica, privacidad, protección de datos personales, armas autónomas, robots laborales y robots de vigilancia, ciborgización, entre otros, avanzan sin ningún tipo de obstáculos.

¿Qué hacer en caso que los Estados promuevan políticas de desarrollo de la IA que generen riesgos existenciales? O bien, ¿qué hacer en caso que no promuevan esos desarrollos pero que, por su inacción, sus conductas omisivas, los permitan? ¿Qué derechos, qué acciones pueden impetrar los ciudadanos de los Estados miembros de la ONU? ¿La humanidad tiene que esperar hasta que el problema estalle en su cara?

El asunto es que cuando estalle la superinteligencia artificial, las probabilidades de encontrar una solución que permita la continuidad de la humanidad no estará del todo clara. Por consiguiente, hasta que no podamos hacer proyecciones seguras sobre los efectos beneficiosos o dañinos de los desarrollos y aplicaciones de la IA, “no habrá mayor valentía que la cautela”, en pos conservar la humanidad en el planeta Tierra.

En conclusión: hay que actuar ya. La respuesta que he dado como solución a este problema existencial de la humanidad, son las tres leyes de la IA, inferidas del DIDH. Digo: sí al desarrollo de la IA, pero en función de la sustentabilidad de la humanidad.

---

# TRANSHUMANISMO Y DERECHOS ¿HUMANOS?

ANDRÉS GIL DOMÍNGUEZ

## I. INTRODUCCIÓN

Existen numerosas libros, ensayos y artículos sobre la inteligencia artificial en sus distintas facetas y enfoques que cuando se enlazan con el campo de lo jurídico intentan someterla al límite emergente de los derechos humanos. La evolución de la IA tendría que estar siempre encuadrada como norte en el ámbito establecido normativamente por los instrumentos internacionales sobre derechos humanos y no podría tener otra lógica normativa expansiva.

La postura expuesta se basa en una exclusiva idea antropocéntrica que inhibe toda alternativa posible en torno a la titularidad de los derechos humanos y que presupone la eternidad e inmutabilidad del *homo sapiens* ante los efectos del desarrollo científico y tecnológico (entre los que se encuentra la inteligencia artificial y su impacto en la humanidad).

La cuarta revolución industrial<sup>(1)</sup> presenta cuatro ingredientes que funcionan sincrónicamente. En primer lugar, la producción permanente de datos de toda clase denominada en su conjunto macro datos o *big data*. En segundo lugar, la creación de algoritmos inteligentes con la capacidad suficiente para procesar dicha información en un modo y tiempo imposible de equiparar por el cerebro humano. En tercer lugar, el desarrollo de una inteligencia artificial fuerte<sup>(2)</sup> que aprende de forma autónoma cada vez que se interrelaciona con el entorno. Por último, la digitalización de cualquier parte del mundo que pueda descomponerse en unidades mínimas tales como sonido, colores, imágenes, cantidades, temperaturas

---

(1) Nota del director: sobre la cuarta revolución industrial ampliar en López Oneto, Marcos “Fundamentos antropológicos, éticos, filosóficos, históricos, sociológicos y jurídicos para la constitución universal de un Derecho de la Inteligencia Artificial (DIA)”, tomo I del presente *Tratado*.

(2) Nota del director: sobre inteligencia artificial fuerte o superinteligencia ver Corvalán, Juan G. - Díaz Dávila, Laura C. - Simari, Gerardo I., “Inteligencia artificial: bases conceptuales para una aproximación interdisciplinar”, tomo I del presente *Tratado*. Asimismo, sobre inteligencia artificial fuerte o superinteligencia como riesgo existencial para la humanidad ver López Oneto, Marcos, “Fundamentos antropológicos, éticos, filosóficos, históricos, sociológicos y jurídicos para la constitución universal de un Derecho de la Inteligencia Artificial (DIA)”, tomo I del presente *Tratado*.

para ser traducida a un lenguaje que como serie de cifras no tiene peso (por ende, es muy ligera), viaja a una gran velocidad, no se estropea en el camino y llega intacta a donde se envía.

El disfrute de los componentes de la cuarta revolución industrial se traduce normativamente en el derecho al desarrollo científico y tecnológico. En la regla de reconocimiento constitucional y convencional argentina este derecho se encuentra alojado en la Constitución argentina y en los Instrumentos Internacionales sobre derechos humanos que ostentan jerarquía constitucional originaria y derivada.

El art. 75, inc. 19, primer párrafo, de la Constitución argentina establece como una atribución del Congreso “proveer lo conducente al desarrollo científico y tecnológico, su difusión y aprovechamiento”. Son sinónimos de proveer: suministrar, abastecer, proveer, surtir, aprovisionar, equipar, dotar, proporcionar. Son antónimos de proveer: quitar, privar, desabastecer. En los sinónimos y antónimos se encuentra la clave constitucional de la obligación que titulariza el Congreso respecto del desarrollo científico y tecnológico: la sanción de normas que permitan el acceso progresivo al aprovechamiento de la tecnología como garantía del ejercicio del derecho al desarrollo científico y tecnológico. El Pacto Internacional de Derechos Económicos, Sociales y Culturales establece en el art. 15.c que toda persona humana tiene derecho a gozar de los beneficios del progreso científico y de sus aplicaciones. La Declaración Americana de Derechos y Deberes del Hombre en el art. XIII sostiene que toda persona humana titulariza el derecho a disfrutar de los beneficios que resulten de los progresos intelectuales y especialmente de los descubrimientos científicos. El Protocolo de San Salvador, tratado internacional que tiene jerarquía supralegal, en el art. 14.b reconoce el derecho de las personas humanas a gozar de los beneficios del progreso científico y tecnológico. La Corte Interamericana de Derechos Humanos en el caso “Artavia Murillo vs. Costa Rica” sostuvo que el acceso a las técnicas de reproducción humana asistida era parte del contenido del derecho a disfrutar de los avances científicos y tecnológicos que los Estados debían garantizar despejando toda clase de obstáculos legales o fácticos que lo impidiesen.

La convergencia constitucionacional de las fuentes determina la existencia normativa del derecho al desarrollo científico y tecnológico como un derecho fundamental y un derecho humano. La posición normativa antropocéntrica no alcanza a percibir todavía que el ejercicio de este derecho es justamente la apertura o la llave hacia dimensiones que van más allá del antropocentrismo por una decisión del propio ser humano, y que, llegando un cierto punto de desarrollo, la condición de la construcción normativa será inexorablemente distinto a la impuesta por el antropocentrismo.

## II. DIGITALIZACIÓN

2.1. La digitalización en una de sus acepciones se vincula con la utilización de las distintas opciones que ofrecen las tecnologías de la información y la comunicación en el funcionamiento de las instituciones y las relaciones humanas.

El COVID-19 impuso una agenda digital impensada hasta hace un par de meses atrás que desnudó aquello que estábamos en condiciones de realizar digitalmente pero que por razones de arraigo cultural analógico todavía no emprendíamos. Veamos algunos ejemplos significativos.

En el ámbito del Poder Judicial, la Corte Suprema de Justicia a través del dictado de las Acordadas 4/2020, 6/2020, 11/2020 y 12/2020 estableció que los acuerdos de ministros se pueden realizar por medios virtuales, remotos o de forma no presencial, las presentaciones judiciales serán completamente en formato digital con firma electrónica del presentante, el trabajo de los jueces, funcionarios y empleados sea remoto, la firma digital de los jueces respecto de los diferentes actos jurisdiccionales y administrativos que suscriban, la circulación digital de los expedientes y el inicio digital de causas.

En el ámbito del Congreso de la Nación tanto la Cámara de Senadores como la de Diputados habilitaron el desarrollo de sesiones digitalmente puras o mixtas (presenciales y digitales). La Cámara de Senadores a través del dictado del DP-0008/20 (4 de mayo 2020) dispuso la realización de modo excepcional de sesiones de carácter remoto o virtual mediante videoconferencia por el plazo de sesenta (60) días que podrá ser prorrogado por el Senado en tanto persista la situación de emergencia sanitaria generada por el COVID-19. La Cámara de Diputados a través del dictado del Protocolo de Funcionamiento Parlamentario Remoto (Aislamiento social, preventivo y obligatorio-Coronavirus COVID-19) por parte de la Comisión de Peticiones, Poderes y Reglamentos (29 de abril de 2020) habilitó las sesiones telemáticas mientras dure el aislamiento social, preventivo y obligatorio en todo o en parte del territorio nacional por períodos de 30 días prorrogables.

La Inspección General de Justicia a través del dictado de la resolución 11/2020 (26 de marzo de 2020) habilitó que los estatutos de las sociedades y de las asociaciones civiles contemplen mecanismos a distancia para la realización de las reuniones del órgano de administración o de utilizando medios que les permitan a los participantes comunicarse simultáneamente entre ellos. Uno de los fundamentos de la resolución fue que “negar la posibilidad que los acuerdos sociales se adopten por asambleas o reuniones a distancia mediante la utilización de los nuevos medios tecnológicos disponibles no favorece a los socios, ni a la sociedad, ni en defi-

nitiva al funcionamiento de nuestras sociedades como vehículos generadores de riqueza y desarrollo económico”.

La Suprema Corte de Justicia de la Provincia de Buenos Aires y la Asociación Judicial Bonaerense celebraron un convenio regulador de la modalidad de teletrabajo durante la vigencia de las restricciones de la pandemia virus COVID-19 (25 de abril de 2020) con el objeto de regular los aspectos funcionales y técnicos del teletrabajo o trabajo remoto en el ámbito de la administración de justicia provincial a fin de garantizar la prestación del servicio de justicia a través del uso de las tecnologías de la comunicación e información.

En el plano de las relaciones horizontales la digitalización ha permitido la renegociación de contratos de diversa naturaleza, la concreción de contratos de compraventa variados la realización de negocios, el funcionamiento de comercios, etc. En particular, facturación ante la AFIP, los trámites no presenciales, audiencias ante el SECCLO, transferencias bancarias nacionales e internacionales, *home banking*, *home office*, turismo, entre otros.

Como se observa la digitalización es parte de nuestra vida cotidiana, de la construcción de sentidos del lenguaje, de la identidad de las personas, de la convivencia diaria. Si bien estamos en la etapa arqueológica digital signada por la abrupta interrupción del COVID-19 a la cual hay que responderle con los instrumentos tecnológicos disponibles, muy pronto pasaremos a una profundización de la era digital con herramientas más sofisticadas. Como mínimo la actual situación deja como estructura de interrelación que el modo analógico dejó de ser la única forma posible de funcionamiento de las instituciones y las relaciones humanas<sup>(3)</sup>.

2.2. Baricco<sup>(4)</sup> sostiene que la primera huella geológica de la digitalidad es la aparición de *Space Invaders* a partir del cual dejamos atrás el metegol como muestra de lo físico y nos sumergimos en el mundo de lo inmaterial, gráfico e indirecto. Ese fue el día inicial de la persona-teclado-pantalla que posteriormente se transformó en una especie de postura cero en la que los dispositivos se convirtieron casi en prótesis orgánicas del cuerpo humano. A partir de este momento se desarrolló un mundo que paulatinamente fue transitando en distintas etapas, una digitalidad que lleva en su ADN el patrimonio genético del videojuego (*the game*). ¿Cuál fue la figura que representó el apego del siglo XX al metegol como representación de lo analógico? Un iceberg o una pirámide invertida en

---

(3) Gil Domínguez, Andrés, *Inteligencia artificial y derecho*, Rubinzal-Culzoni, Ciudad Autónoma de Buenos Aires, 2019.

(4) Baricco, Alessandro, *The Game*, Anagrama, Barcelona, 2018.

cuya cubierta flotan las percepciones superficiales y el juego consistía en superarlas guiados por alguna clase de mediación a través de un maestro, sacerdote, viajes, profesores, viajes, libros, goces, sufrimientos hasta llegar al punto de la experiencia descartando que dicho camino fuera sencillo de transitar. Esta figura la aplicábamos a los aspectos más diversos de la realidad se trate de investigar una noticia, entender una poesía o vivir un amor. Así funciona el derecho constitucional analógico en donde lo construido en términos de sentidos aparece como un saber consolidado que solo puede ser explicado, pero nunca conmovido desde la superficie. La lucha entre los “mediadores” o “maestros” se reduce a demostrar quién sabe más del saber encapsulado por el constitucionalismo analógico, pero es atentar contra su corazón siquiera plantear alternativas innovadoras. ¿Qué aparece cuando invertimos el iceberg o la pirámide? El premio aparece arriba, el esfuerzo abajo, las esencias emergen a la superficie, la complejidad queda escondida en algún lugar. Así funcionan el Iphone, YouTube, Spotify, Facebook, WhatsApp, Tinder desplegando una simplicidad donde la complejidad de la realidad emerge en la superficie dejando atrás cualquier lastre que haga más pesado el corazón esencial. Con esa lógica también funciona el constitucionalismo digital de la cuarta revolución industrial donde la innovación siempre es posible más allá de las experiencias emergentes que circulan por la superficie casi sin intermediación y exige, una y otra vez, respuestas nunca dadas más que interrogantes basados en experiencias consolidadas. El constitucionalismo digital no pretende desconocer o destruir el saber acumulado, pero lejos de configurar un sesgo incuestionable, lo utiliza para encontrar respuestas a preguntas que requieren innovación, de la misma manera que Siri el sistema operativo de Apple nos trae la información que requerimos para adoptar decisiones en diversos sentidos.

El constitucionalismo digital se presenta como una banda de Moevius, no hay un arriba o un abajo determinado sino un constante movimiento de cambio entre el arriba y el abajo. La insurrección digital contra la civilización del siglo XX. La experiencia como realización, plenitud, rotundidad, sistema hecho realidad versus la posexperiencia como arrebató, exploración, pérdida de control, dispersión. Lo analógico como la conclusión de un gesto solemne, el resultado tranquilizador de una operación compleja, el regreso al final del hogar en contraposición con lo digital como el principio de un gesto, la apertura de una exploración, el rito del alejamiento, la trayectoria de un andar. Constitucionalismo digital *vs.* constitucionalismo analógico<sup>(5)</sup>.

---

(5) *Ibidem.*

Desde otro escenario, esta dicotomía también es advertida por Rondina<sup>(6)</sup> cuando afirma que muchos juristas, jóvenes o viejos, analizan el derecho constitucional como si fueran médicos forenses sobre una mesa de autopsia: miran un organismo muerto y nos cuentan de qué murió; mientras que otros constitucionalistas consideran que el derecho constitucional no debe renunciar a ser herramienta de transformación que debe ser pensado y construido todo el tiempo.

El COVID-19 incrementó notablemente el factor de aceleración de la cuarta revolución industrial dejando una impronta en la subjetividad que generará en breve interdicciones, en general, al discurso jurídico, pero especialmente al constitucionalismo. Actualmente las personas son sujetos que gozan en términos lacanianos a través de los datos dejando a su paso un rastro digital (una suerte de “alma de datos” o de “oro azul”) del que puede extraerse diversas conclusiones y perfiles con el uso de la herramienta tecnológica adecuada basada en el aprendizaje automático<sup>(7)</sup>. Somos o vamos camino a configurarnos como “seres subjetivamente improntados por los datos” o “seres datados” que expresan una suerte de “narcisismo digital”. Netflix nos invita a disfrutar una serie o película que según nuestros datos puede gustarnos, Waze se encarga de que lleguemos rápido y seguro a un destino para después indicarnos cómo volver a casa, las *apps* de lugares de alojamientos nos envían ofertas permanentes sobre la base de los viajes realizados, Tinder es el Disney digital del ejercicio diverso de la sexualidad. Hay una nueva forma de consentimiento y relación con los datos innovadora. Seguramente en breve, y mucho más después de la pandemia, aparecerán conflictos sobre el alcance del consentimiento digital oportunamente otorgado, porque una cosa es que alguien consienta entregar datos en relación con sus preferencias cinéfilas y otras es que después se crucen con otros datos para perfilar sesgos biográficos<sup>(8)</sup>. Ante esta situación, ¿puede el derecho constitucional analógico brindar respuestas mínimamente racionales si todavía sigue enmarcado en proteger datos estáticos archivados en registros o bancos de datos públicos o privados destinados a proveer informes que solos se transfieren empaquetados?

---

(6) Rondina, Domingo, “El derecho constitucional no es LOT, debe ser JANO”, Revista ES nro. 1, Asociación Argentina de Derecho Constitucional, Ciudad Autónoma de Buenos Aires, 2020.

(7) Gil Domínguez, Andrés, ob. cit., p. 15.

(8) Nota del director: sobre la elaboración de perfiles por parte de la inteligencia artificial y el derecho al consentimiento ver Corvalán, Juan G. - Papini, Carina M., “Explorando los perfiles digitales humanos: cómo los algoritmos nos conocen más que nosotros mismos. Incluye el análisis del nuevo proyecto de Ley de protección de datos personales de Argentina”, tomo II del presente *Tratado*. Asimismo, sobre la elaboración de perfiles y su utilización en relación con la publicidad ver Berchi, Mauro, “El problema de los datos en la Cuarta Revolución Industrial”, tomo II del presente *Tratado*.

¿En la era digital tenemos que aceptar que la privacidad es algo del pasado?<sup>(9)</sup> El “oro azul” de los datos dispersos en las minas digitales están siendo explotados por muchas empresas (los *data brokers*) sin que las personas sean conscientes de lo rentable que son los rastros digitales, cuánta información está siendo utilizada ni para qué. ¿Cómo no va a cambiar la noción de privacidad en un mundo digital si la internet no solo sabe mejor que nosotros quiénes somos, sino también, quiénes vamos a ser?<sup>(10)</sup> Así como en el presente observamos sorprendidos que en el pasado los coches no tenían cinturón de seguridad o los médicos fumaban en los consultorios, quizás, en un futuro cercano, parezca insensata la incontinencia narcisista viral actual y esta primera época se recuerde como una especie de lejano Oeste digital en el que todo valía<sup>(11)</sup>. ¿Está en condiciones el constitucionalismo analógico para proteger la intimidad digital del siglo XXI cuando sigue encorsetado en un concepto propio del siglo XX?

Algo similar sucede en el ámbito de la teoría de la Constitución. Existe una corriente teórica muy intensa y productiva que desarrolla el modelo democrático deliberativo basado en la participación popular en la toma de las decisiones colectivas que se presenta como el paradigma moral y jurídico más relevante para establecer el contenido y alcance del sistema de derechos. ¿Seguirán sus teóricos deambulando en las cómodas praderas del constitucionalismo analógico o empezarán a interiorizarse sobre cómo funciona la tecnología *blockchain* para darle a tantos libros una operatividad digital?<sup>(12)</sup>

Así podríamos plantear numerosos escenarios en términos de democracia, derecho y tecnología en los cuales la opción entre el constitucionalismo analógico y el constitucionalismo digital se hace presente, y mucho más aún, a partir de una pandemia global que no sabemos todavía cuándo y cómo terminará.

2.3. Si bien la digitalización no genera ningún cambio o mejoramiento sustancial del *homo sapiens* implica un cambio en la intersubjetividad y en la relación con la tecnología. Las relaciones digitales se convertirán en la regla y los encuentros personalizados en una excepción vinculada a los lazos socioafectivos de las personas. En tanto que la ecuación digitalidad —producción de datos— mayor desarrollo tecnológico constituye un puente irreversible hacia el transhumanismo.

---

(9) Lyseggen, Jorn, *Outside*, Insight, ESIC, Madrid, 2019, p. 237.

(10) García Aller, Marta, *El fin del mundo tal y como lo conocemos*, Planeta, Barcelona, 2019, p. 248.

(11) *Ibíd*em, p. 244.

(12) Gil Domínguez, *ob. cit.*, p. 40.

### III. TRANSHUMANISMO Y POSHUMANISMO

3.1. La cuarta revolución industrial se caracteriza por el fenómeno de la convergencia tecnológica que suele identificarse con el acrónimo NBIC (nano-bio-info-cogno) donde la letra N se refiere a la nanotecnología, la letra B a la biotecnología, la letra I a las tecnologías de la información y la letra C a las ciencias cognitivas<sup>(13)</sup>.

Las nanociencias son el estudio de los fenómenos de la manipulación de materiales a escala atómica, molecular y macromolecular en las que las propiedades difieren significativamente de las observadas a mayor escala. El nanómetro representa una milmillonésima de metro, o sea, una longitud en relación con el metro igual que la de una avellana respecto de la Tierra (esto implica que un objeto del tamaño de un nanómetro tendría un grosor cincuenta mil veces más pequeño que el de un cabello). La biotecnología permite conocer nuestro ADN con las anomalías que pueda tener con la misma sencillez que nos hacemos un examen de sangre. Gracias a un descubrimiento reciente —el “Crispr-Cas9”— es posible cortar, pegar e hibridar nuestro ADN con la misma facilidad que corregimos una falta de ortografía o movemos una frase de un lado para el otro mediante un procesador de texto. La informática se vincula con el *big data* y con sus características intrínsecas, las denominadas “V”: volumen, velocidad, variedad, valor. La cognitividad se relaciona con la inteligencia artificial en sus dos variables (la débil y la fuerte)<sup>(14)</sup>. Como observamos el fenómeno NBCI se basa en la unidad de la ciencia a nano escala y en las tecnologías de integración nanométrica que la hacen posible: átomos, neuronas, genes y bits<sup>(15)</sup>.

3.2. El transhumanismo es un amplio proyecto que persigue la mejora de la humanidad en todos sus aspectos (físico, intelectual, emocional, moral) gracias a los progresos emergentes del desarrollo científico y tecnológico. Supone la superación de la evolución biológica basada en la selección de variaciones aleatorias por la evolución basada en la tecnología

---

(13) López Oneto, Marcos, *Fundamentos para un derecho de la inteligencia artificial. ¿Queremos seguir siendo humanos?*, Tirant lo Blanch, Valencia, 2020, p. 24.

Nota del director: asimismo, ver López Oneto, Marcos “Fundamentos antropológicos, éticos, filosóficos, históricos, sociológicos y jurídicos para la constitución universal de un Derecho de la Inteligencia Artificial (DIA)”, tomo I del presente *Tratado*.

(14) Ferry, Luc, *La revolución transhumanista. Cómo la tecnomedicina y la uberización del mundo van a transformar nuestras vidas*, Alianza Editorial, Madrid, 2017, p. 195 y ss.

(15) López Oneto, Marcos, ob. cit., p. 24.

que podría dar lugar a una nueva especie poshumana descendiente de nuestro linaje, pero mucho más avanzada<sup>(16)</sup>.

Una primera gran división distingue al transhumanismo que solamente quiere mejorar la especie humana, pero sin renunciar a la humanidad del transhumanismo que aboga por la “tecnofabricación” de una posthumanidad generadora de una nueva especie basada en máquinas dotadas de capacidades físicas y de una inteligencia artificial infinitamente superior a la nuestra. En la primera versión asume una continuidad de un cierto “humanismo naturalista”, en la segunda propone una ruptura frontal con el humanismo<sup>(17)</sup>.

Si bien en principio ambas corrientes se distinguen, una es el presupuesto de la otra, en tanto que el transhumanismo es el trayecto y el posthumanismo es el objetivo, uno es el camino y el otro el punto de llegada. La conexión de lo “trans” a lo “post” es inescindible de la ruptura que este movimiento produce al cuestionar los límites humanos por medio de la ciencia y la tecnología combinadas con el pensamiento creativo, al interditar el carácter inevitable de la vejez y de la muerte e intentar mejorar progresivamente la capacidad intelectual y física, como así también, el desarrollo intelectual. La humanidad es vista como una fase de transición en el desarrollo evolutivo de la inteligencia y la ciencia como un medio razonable para pasar de una condición humana a una condición transhumana y de allí a una posthumana.

El pasaje del azar a la elección depende del plan de vida de las personas y rechaza toda clase de imposición estatal, no es discriminatoria puesto que intenta reparar las injusticias infligidas a los seres humanos por una naturaleza ciega e insensible, se inscribe en una perspectiva democrática que persigue añadir a la igualdad económica y social la igualdad genética, no desea en absoluto eliminar a los débiles o vulnerables (como proponía el nazismo) sino por el contrario tiene por objeto reparar o aumentar las capacidades humanas que la naturaleza distribuye de forma desigual<sup>(18)</sup>.

---

(16) Nota del director: sobre el concepto de transhumanismo, su caracterización y las diferentes posturas de los transhumanistas ampliar en López Oneto, Marcos, “Fundamentos antropológicos, éticos, filosóficos, históricos, sociológicos y jurídicos para la constitución universal de un Derecho de la Inteligencia Artificial (DIA)”, tomo I del presente *Tratado*.

(17) Ferry, Luc, ob. cit., p. 25.

(18) Ferry, Luc, ob. cit., p. 51.

3.3. A grandes rasgos es posible distinguir dos clases de transhumanismo: el cultural y el tecnocientífico<sup>(19)</sup>.

El transhumanismo cultural no busca el mejoramiento del *homo sapiens* a través de la tecnología o la ingeniería genética, sino que persigue realizar una crítica de la concepción de lo humano considerada como natural y transmitida de esa manera de generación en generación que, en resumidas cuentas, solo produjo una serie de prejuicios eurocéntricos, racistas, sexistas y especistas.

El transhumanismo tecnocientífico intenta transformar sustancialmente al *homo sapiens* mediante la aplicación de tecnología mediante distintos enseres tales como la fusión con máquinas, el mejoramiento biológico a través de medicamentos, la manipulación genética de óvulos y espermatozoides, de forma tal que las personas, ejerciendo el “derecho a la libertad morfológica” (esto es, la potestad de decidir sobre el propio fenotipo modificando el cuerpo como se quiera), en un futuro puedan generar una especie mejorada que ha recibido varias nominaciones bautismales (el *Homo Deus*, el *Homo Excelsior*, el *Homo Roboticus*, el *Homo Postorgánico*). Esta clase de transhumanismo defiende con ahínco que el *homo sapiens* debe abandonar la pasividad con la que ha sido sometida al “proceso evolutivo darwiniano” a fin de poder superar, gracias a la tecnología, las múltiples limitaciones que tiene que soportar: es el tiempo de dejar a atrás la evolución natural, tomar el control y conducir una evolución dirigida o diseñada.

El transhumanismo tecnocientífico abarca dos variables. La primera anuncia un futuro posbiológico en que el *homo sapiens* será sustituido en el control del planeta por descendientes tales como robots superinteligentes que serían el resultado de verter nuestro cerebro (el *software*) en un recipiente más duradero (las máquinas) que el actual *hardware* (el degradable cuerpo humano). La segunda está signficada por el bio-mejoramiento humano a través de la ingeniería genética (con el objeto de eliminar genes defectuosos, la potenciación de genes con cualidades deseables o la inserción en el genoma humano de genes procedentes de otras especies) y la biología sintética (a través de la creación en un laboratorio de genes artificialmente diseñados para fines específicos capaces de hacer que las células adquieran funciones radicalmente nuevas que no poseen por naturaleza).

El transhumanismo como movimiento mundial elaboró una Declaración Transhumanista que contiene ocho puntos entre los que se destacan

---

(19) Diéguez, Antonio, *Transhumanismo. La búsqueda tecnológica del mejoramiento humano*, Herder, Barcelona, 2017, ps. 40 y ss.

los siguientes: a) La humanidad es susceptible de ser afectada profundamente por la ciencia y la tecnología en el futuro. Prevemos la posibilidad de agrandar el potencial humano venciendo el envejecimiento, las limitaciones cognitivas, el sufrimiento involuntario y nuestro confinamiento al planeta Tierra (punto 1); b) La reducción de los riesgos existenciales y el desarrollo de los medios para la preservación de la vida y la salud, el alivio del sufrimiento grave y la mejora de la previsión y de la sabiduría humana deberían ser promovidos como prioridades urgentes y ser fuertemente financiados (punto 5); c) Abogamos por el bienestar de todo ser sintiente, incluidos los humanos, los animales no humanos y cualquier intelecto artificial futuro, forma de vida modificada u otra inteligencia que pueda surgir por medio de los avances tecnológicos y científicos (punto 7); d) Defendemos que se permita a los individuos una amplia elección personal acerca de cómo llevar sus vidas. Esto incluye el uso de técnicas que puedan desarrollarse para ayudar a la memoria, concentración y energía mental; terapias de alargamiento de la vida; tecnologías para la elección reproductiva; procedimientos criogénicos y muchas otras posibles tecnologías para la modificación y mejora (*enhancement*) del ser humano (punto 8).

3.4. La vida definida como un proceso capaz de preservar su complejidad y de replicarse puede desarrollarse a través de tres fases: a) una biológica (1.0) en la que tanto el *software* como el *hardware* son producto de la evolución; b) una cultural (2.0) en la que puede diseñar el *software* (mediante el aprendizaje); c) y una tecnológica (3.0) en la cual puede diseñar también el *hardware* convirtiéndose en la dueña de su propio destino. La IA podría permitirnos crear vida 3.0 durante el transcurso del siglo XXI y esta posibilidad desata grandes controversias. Por un lado, los tecnoescépticos consideran que construir una IA sobrehumana es tan difícil que no sucederá en cientos de años y que no vale siquiera la pena preocuparse por la vida 3.0 en estos momentos. Por el otro, los utópicos digitales consideran que es probable que la vida 3.0 suceda a lo largo del siglo XXI como un hecho positivo de la evolución humana. Por último, Por último, el movimiento en pro de una IA benéfica considera que la preocupación está justificada y es útil porque llevar a cabo ahora investigaciones y discusiones en torno a la seguridad en IA incrementa la probabilidad de que los resultados sean positivos<sup>(20)</sup>.

3.5. Inteligencia y conciencia son cosas muy distintas. La inteligencia es la capacidad de resolver problemas. La conciencia es la capacidad de sentir dolor, alegría, amor e ira. Es posible que una IA analice los sentimientos humanos sin tener que experimentar algún sentimiento propio,

(20) Tegmark, Max, *Vida 3.0, Qué significa ser humano en la era de la inteligencia artificial*, Taurus, Barcelona, 2018, ps. 46 y ss.

aunque tal vez algún día pueda llegar a desarrollarlo. Es necesario considerar al menos tres opciones: a) la conciencia está relacionada de algún modo con la bioquímica orgánica, de forma tal, que nunca será posible crear la conciencia en sistemas no orgánicos; b) la conciencia no está relacionada con la bioquímica orgánica, pero sí con la inteligencia artificial, de forma tal que los ordenadores podrían desarrollar alguna clase de conciencia si desean superar un determinado umbral de inteligencia; c) no existen conexiones esenciales entre la conciencia y la bioquímica orgánica y la inteligencia superior, por lo tanto, los ordenadores podrían desarrollar alguna forma de conciencia, aunque esto no fuera necesario para llegar a ser superinteligentes<sup>(21)</sup>.

Los antropólogos han observado en muy diversas culturas prácticas de curación acompañadas de rituales y conjuros que en cierta medida parecen producir efectos benéficos. Los placebos son una especie de simulacro durante el cual se administran píldoras o inyecciones con sustancias inocuas o se practican operaciones quirúrgicas fingidas con la intención de aliviar a un enfermo. El “efecto placebo” revela la presencia de un sistema de sustitución simbólica de elementos que faltan en el proceso biológico y que se complementa con el apoyo de estas “prótesis simbólicas”. El teléfono celular inteligente se convirtió en una pieza importante del exocerebro que funciona como una prótesis que nos conecta con el entorno y como un amuleto portátil dotado de un inmenso poder simbólico que genera un efecto placebo similar al de los conjuros y encantamientos. Cuando un médico o un chamán realizan un ritual simbólico se suele producir un efecto biológico comprobable llamado efecto placebo, cuando un ingeniero crea un programa y lo inserta en una máquina produce un efecto mecánico útil denominado el efecto robótico. En ambos casos hay un artificio intelectual que provoca efectos tangibles en un cuerpo orgánico o mecánico, la diferencia es que, en el efecto placebo el sujeto cuyo cuerpo quiere sanar debe estar plenamente consciente de que está siendo sometido a un proceso de curación (aunque la persona no es consciente que le están administrado fármacos inocuos) en tanto que la máquina no es consciente de lo que está sucediendo lo cual no impide que se produzca un efecto robótico y el artefacto se ponga a funcionar encaminado a un objetivo predeterminado<sup>(22)</sup>.

Roger Bartra sostiene que estaremos frente a un robot verdaderamente consciente en el momento en que comprobemos que siente alivio al aplicarle un placebo cuando sufra un malestar. En otras palabras, esto de-

---

(21) Harari, Yuval Noah, *21 lecciones para el siglo XXI*, Debate, Ciudad Autónoma de Buenos Aires, 2018, p. 92.

(22) Bartra, Roger, *Chamanes y robots. Reflexiones sobre el efecto placebo y la conciencia artificial*, Anagrama Argumentos, Barcelona, 2019.

mostrará que está dotado de un sistema al que se puede engañar y que le permite superar el malestar que padece. Para que los robots alcancen formas de consciencia tan sofisticadas como las humanas deberán pasar por los rituales del placer y del dolor. Sin el sufrimiento de malestares y sin el placer del alivio es difícil concebir máquinas inteligentes con formas humanoides orientadas por una conciencia desarrollada compleja. Los robots del futuro serán conscientes sin un chamán del futuro, con el arte de la palabra, logra cautivarlos y se comprueba que son sensibles al efecto placebo<sup>(23)</sup>.

No existe una sola definición correcta e indiscutible del término “consciencia”, sino que se observan muchas definiciones alternativas con conceptos tales como sensibilidad, vigilia, autoconsciencia, acceso a la información sensorial y capacidad de combinar información en un relato. Una exploración del futuro de la inteligencia desde la perspectiva más incluyente posible que no se limite a los tipos de consciencia biológica existentes es lo que propone Tegmark utilizando una definición muy amplia en los siguientes términos: *consciencia = experiencia subjetiva*. Esta clase de conceptualización no menciona el comportamiento, la percepción, la autoconsciencia, las emociones o la atención e implica que somos conscientes mientras soñamos, aunque no estemos despiertos ni tengamos acceso a la información sensorial. Análogamente, cualquier sistema que experimente dolor es consciente en este sentido, aun cuando no pueda moverse. La definición de Tegmark deja abierta la posibilidad de que en el futuro puedan existir sistemas de IA conscientes, incluso en forma de *software* sin conexión a sensores o cuerpos robóticos<sup>(24)</sup>.

¿Llegará un día en que la IA supere a la inteligencia humana? La singularidad tecnológica es una hipótesis que sugiere que la velocidad tan acelerada con la que progresa la tecnología provocará que la IA tarde o temprano exceda la capacidad intelectual de los humanos y, por lo tanto, el control que tenemos sobre ella<sup>(25)</sup>.

La primera aplicación del término singularidad a un futuro evento tecnológico se le atribuye a John von Neumann en los años cincuenta quien sostuvo que el acelerado avance tecnológico podría acercarse a alguna forma de singularidad en la historia de la especie más allá de la cual los asuntos humanos tal como los conocemos no podrían continuar. En 1993 Vernor Vinge un matemático de la Universidad de San Diego expuso que cuando contásemos con los medios para crear una inteligencia

(23) *Ibíd.*, p. 177.

(24) Tegmark, *ob. cit.*, p. 349.

(25) Fernández I Aragonés, Aina - García Pujadas, Albert, *¿Libres o vasallos? El dilema digital. El viejo topo*, Barcelona, 2017, p. 75.

sobrehumana la era de la humanidad habrá terminado. El relevo lo tomó Ray Kurzweil quién asegura que la singularidad acontecerá hacia el 2045 mediante la fusión del *homo sapiens* con las máquinas del futuro garantizando una suerte de inmortalidad por medio de la “velocidad de escape de la longevidad” (esto es, que uno pueda mantenerse vivo el tiempo suficiente para alcanzar la próxima innovación para alargar la vida)<sup>(26)</sup>.

Quizás sea difícil que esto suceda en un tiempo cercano puesto que la inteligencia humana es flexible ante nuevas situaciones que no son exactamente aquellas que fueron aprendidas a efectos de poder actuar en consecuencia en el armado de un nuevo entramado. Todavía las emociones y el cuerpo cumplen un rol esencial en el procesamiento cognitivo de forma tal que los circuitos neuronales que subyacen a la cognición y a la emoción son interdependientes e interactúan en el funcionamiento de los procesos más elementales tales como la percepción temprana y los más complejos como la toma de decisiones, el razonamiento y la conducta moral y social: no procesamos la nueva información de forma integralmente racional sino que la integramos con información sobre nuestras experiencias pasadas y con las sensaciones corporales para interpretar lo que sucede a nuestro alrededor<sup>(27)</sup>. No obstante esto, existe una convivencia entre el hombre y la máquina que se instala en una lógica basada en “proposiciones maquínicas”<sup>(28)</sup> que indudablemente pone en tela de juicio el “prestigio del experto” basado en la escasez del conocimiento que perdura desde la era Gutenberg<sup>(29)</sup> obligando a nuevas formas de valor añadido a una información que está directa o indirectamente al alcance de Google. Varios capítulos de las distintas temporadas de la serie *Black Mirror*<sup>(30)</sup> plantean este supuesto y sus distintas consecuencias, pero quizás el capítulo más logrado en término optimistas sea *San Junipero* (2016) el cual narra una historia de amor entre dos mujeres (Yorki y Kelly), quienes logran mantener para siempre su amor en un mundo virtual una vez que fallecen. Otro espejo es posible encontrarlo en la película *Transcendence* (2014) en la cual un investigador junto a su pareja que también es investigadora intenta crear una IA que recree una consciencia colectiva y autosuficiente que mejore sustancialmente la

(26) Ford, Martín, *El auge de los robots. La tecnología y la amenaza de un futuro sin empleo*, Paidós, Ciudad Autónoma de Buenos Aires, 2016, p. 218.

(27) Manes, Facundo - Niro, Mateo, *El cerebro del futuro. ¿Cambiará la vida moderna nuestra esencia?*, Planeta, Ciudad Autónoma de Buenos Aires, 2018, p. 70.

(28) Guattari, Félix, *La revolución molecular*, Errata Naturae, Madrid, 2017, p. 537.

(29) Bunz, Mercedes, *La revolución silenciosa. Como los algoritmos transforman el conocimiento, el trabajo, la opinión pública y la política sin hacer mucho ruido*, Cruce, Ciudad Autónoma de Buenos Aires, 2017, p. 40.

(30) Ierardo, Esteban, *Sociedad pantalla. Black Mirror y la tecnoddependencia*, Ediciones Continente, Ciudad Autónoma de Buenos Aires, 2018.

vida en la tierra; la concreción de la “trascendencia” tal como define a la singularidad tecnológica hace que una vez que su conciencia es transferida antes de fallecer, se enfrenten con él aquellos que lo consideran una amenaza para finalmente destruirlo, aunque esto implique un retroceso enorme para toda la humanidad y sin que puedan impedir que ambos sigan eternamente juntos en una “trascendencia” casi imperceptible.

El entramado ficcional poco a poco se espeja en el desarrollo tecnológico. La empresa emergente Eterni.com con sede en Silicon Valley tiene como objetivo conseguir que las personas puedan preservar sus pensamientos y recuerdos más importantes en un sistema de inteligencia artificial que podría llegar a comunicar mediante una conversación a una persona una vez fallecida con otros que estén vivos posibilitando un alivio al dolor de los que perdieron a un ser querido y permitiendo que las generaciones futuras sepan mucho más sobre sus antepasados<sup>(31)</sup>.

Elon Musk participa de una iniciativa que proyecta crear una interfaz cerebro-ordenador llamada Neuralink mediante la cual se propone crear mecanismos que se puedan implantar en el cerebro humano con el objetivo de permitir que los seres humanos integren *software* y puedan seguir el ritmo de los avances logrados en el campo de la inteligencia artificial<sup>(32)</sup>. El 28 de agosto de 2020 fue un día histórico, Elon Musk presentó el primer prototipo funcional de la empresa Neuralink —Link V0.9, un “fitbit con pequeños cables”— que consiste en un implante cerebral que permite leer la actividad neuronal mediante 1024 electrodos implantados —por ahora— en el cerebro de un cerdo. Aunque el chip no ha sido probado en personas humanas se proyecta que podría captar la actividad cerebral a efectos de poder ayudar al usuario a resolver problemas de salud como la pérdida de la memoria o la ceguera. También podría ser usado para otorgarle a la persona humana la capacidad de controlar dispositivos de forma telepáticamente, guardar recuerdos y almacenar una copia digital del cerebro. Cuando esto suceda, será un robot quirúrgico el que lo implante en la corteza superior del cráneo mediante una pequeña operación que durará menos de una hora ni una hora y solo requerirá anestesia local. El Link V0.9 procesará la actividad física y cerebral para enviar la información a un teléfono celular mediante bluetooth. El dispo-

(31) Rodríguez, Pablo, *Inteligencia artificial. Cómo cambiará el mundo (y tu vida)*, Deusto, Barcelona, 2018, p. 121.

(32) *Ibíd.*, p. 119.

Nota del director: sobre Neuralink de Elon Musk ver Corvalán, Juan G. - Papini, Carina M., “Explorando los perfiles digitales humanos: cómo los algoritmos nos conocen más que nosotros mismos. Incluye el análisis del nuevo proyecto de Ley de protección de datos personales de Argentina”, apart “Neuralink” tomo II del presente *Tratado*.

sitivo estará oculto bajo el pelo y la intención es que sea sencillo de quitar y de reinstalar. La recarga de batería sería por inducción como sucede con algunos relojes modernos.

La singularidad es un significante sobre un futuro evolutivo posible tomando como referencia los avances científicos, el desarrollo científico y tecnológico y el factor de aceleración de la cuarta revolución industrial. El transhumanismo es una condición necesaria pero no suficiente de la singularidad en la medida que hoy existe sin que esta se haya producido y podrá seguir evolucionado sin arribar quizás nunca a la singularidad. Genera una suerte de viaje nómada donde definitivamente se soltaron las amarras del *puerto homo sapiens*, pero es incierto el destino final de esta travesía. El transhumanismo es presente y futuro, la singularidad un futuro posible.

¿Está ocurriendo un desplazamiento en la base filosófica de la tecnología occidental que deja atrás la tradición prometeica que pretende doblegar a la naturaleza persiguiendo el “bien común” de la humanidad y la emancipación de la especie y encaminándose hacia una perspectiva fáustica que pretende superar todas las limitaciones derivadas del carácter material del cuerpo humano entendiéndolas como obstáculos orgánicos que restringen las potencialidades y ambiciones de ser humano?<sup>(33)</sup>. ¿Estamos en un tránsito inexorable hacia un hombre postorgánico que intenta superar la muerte, sagrar una existencia sin cuerpo basada en un código inmaterial como el *software* y establecer un sistema nervioso digital?<sup>(34)</sup>.

La revolución que propone la infotecnología y la biotecnología podrían llegar a reestructurar los cuerpos y las mentes de los seres humanos. Hasta ahora aprendimos a controlar el mundo exterior pero paralelamente teníamos muy poco dominio de nuestro mundo interior. Pudimos construir una represa y direccionar la corriente de un río o diseñar un sistema de irrigación, pero nunca supimos cómo hacer para que el cuerpo dejara de envejecer o de qué manera diseñar un cerebro. La infotecnología y la biotecnología nos proporcionarán el control de nuestro mundo interior y nos permitirá proyectar y producir vida<sup>(35)</sup>.

La relación entre el hombre y la máquina propone transformar y mejorar las capacidades físicas y psíquicas del ser humano a través de im-

---

(33) Sibilia, Paula, *El hombre postorgánico. Cuerpo, subjetividad y tecnologías digitales*, Fondo de Cultura Económica, Ciudad Autónoma de Buenos Aires, 2013, ps. 38 y ss.

(34) Ídem.

(35) Harari, Yuval Noah, ob. cit., p. 25.

plantes tecnológicos que potencien el rendimiento del cuerpo, lo cual permitirá plantea un escenario signado por el transhumanismo y el poshumanismo. El primero consistiría en un ser humano en transformación con capacidades físicas y psíquicas superiores a las de una persona común, sin que esto implique el final de la evolución de la especie humana sino el principio. El segundo supone un ser con capacidades que sobrepasarían de forma excepcional las posibilidades del hombre actual de forma tal que eliminaría cualquier clase de ambigüedad entre un humano y un poshumano siendo este último completamente diferente y quizás más perfecto<sup>(36)</sup>.

3.6. ¿Somos mínimamente transhumanistas sin todavía poder percibirlo? Aunque no lo creamos los *cyborgs* ya existen. Entre nosotros hay personas *cyborgs* de baja intensidad, son aquellas que portan prótesis médicas “mecánicas” (tales como las prótesis de rodilla, los marcapasos, las lentes de contacto) y las que viven a través de las prótesis “tecnológicas externas” (teléfonos celulares, computadoras portátiles, tarjetas de crédito).

También hay *cyborgs* que tienen prótesis más complejas e inteligentes que implican una interacción entre el organismo y la máquina incorporada al cuerpo humano a veces incluso directamente al sistema nervioso (las cuales permiten a las personas que han perdido la vista recuperar parte de la visión usando implantes retinales, posibilitan oír a los sordos profundos mediante implantes cocleares que estimulan el nervio auditivo, permiten controlar algunos síntomas del Parkinson a pacientes con implantes cerebrales que estimulan zonas del cerebro).

A los once años Neil Harbisson fue diagnosticado con una patología congénita y no progresiva denominada acromatopsia que consiste en una anomalía de la visión que solo permite ver la realidad en blanco, negro, gris y las tonalidades derivadas. A los diecinueve años su vida cambió por completo cuando le implantaron un dispositivo en su cráneo —más específicamente una antena de 30 cm— que transforma las frecuencias de color a sonido permitiéndole percibir mediante vibraciones los colores que nunca había podido conocer. En 2004 fue reconocido como el primer *cyborg* de la historia. El implante no solo le permite percibir los colores sino también puede recibir señales espaciales, datos y las llamadas telefónicas.

(36) Sánchez del Campo Redonet, Alejandro, *Reflexiones de un replicante legal. Los retos jurídicos de la robótica y las tecnologías disruptivas*, Thomson Reuters Aranzadi, Pamplona, 2016, p. 88.

Harbisson no está solo. Moon Ribas se implantó un chip en el pie para sentir los movimientos sísmicos en cualquier lugar del planeta. Manel Muñoz cuenta con un sensor barométrico en su nuca mediante el cual tiene la capacidad de percibir la presión atmosférica pudiendo “establecer” el clima en un lugar determinado. Steve Mann posee una pequeña cámara que se conecta a su cráneo con diversas funciones que muestra la realidad aumentada su línea de visión permitiéndole tomar fotografías y grabar videos. Rich Lee se implantó quirúrgicamente imanes en sus oídos convirtiéndolos en auriculares que siempre llevará en su cuerpo. Jerry Jalava producto de un accidente perdió un dedo y optó por un puerto USB incrustado en el dedo de su cuerpo. Hugh Herr perdió sus piernas practicando montañismo y actualmente no solo usa sus piernas artificiales, sino también las diseña como director del grupo de bio-mecatrónica del MIT Media Lab<sup>(37)</sup>.

3.7. Con el poshumanismo emerge un sujeto como proyecto empírico que apunta a experimentar qué son capaces de hacer los actuales cuerpos modificados bio-tecnológicamente. En este punto, la dimensión poshumana del posantropocentrismo puede ser leída como un movimiento de deconstrucción de la supremacía de la especie humana o una suerte de éxodo antropológico, “una fuga de la concepción del hombre como señor absoluto de la creación, una colosal hibridación de la especie”<sup>(38)</sup>. La construcción de una subjetividad poshumana puede ayudar como instrumento político y teórico para intentar comprender los pliegues del futuro en una clave normativa apuntalada por la necesidad de inventar formas de relaciones éticas y valores adecuados a esta complejidad emergente.

Una representación del mundo venidero implica asumir mutaciones radicales en la relación entre naturaleza y cultura, entre componentes biológicos y componentes culturales, ante el abandono de una dimensión en la que la biología ocupaba una función de límite. Se acerca la necesidad de arreglar las cuentas con una nueva antropología construida por la tecnociencia que registre el hecho de que el destino del género humano cada vez está más supeditado a la ciencia y la técnica que lo liberan de la causalidad y de la necesidad y hasta quizás sea posible que lo liberen definitivamente de la naturaleza<sup>(39)</sup>.

(37) Llamas Covarrubías, Jerzain Z., “Derechos humanos, transhumanismo y posthumanismo: una mejora tecnológica humana”, *Derechos Fundamentales a Debate/Comisión Estatal de Derechos Humanos, Jalisco*, p. 99, [http://cedhj.org.mx/revista%20DF%20Debate/articulos/revista\\_No12/ADEBATE-12-art5.pdf](http://cedhj.org.mx/revista%20DF%20Debate/articulos/revista_No12/ADEBATE-12-art5.pdf).

(38) Braidotti, Rosi, *Lo posthumano*, Gedisa, Barcelona, 2015, p. 82.

(39) Rodotà, Stefano, *El derecho a tener derechos*, Trotta, Madrid, 2014, p. 325.

#### IV. EL TRANSHUMANISMO Y LOS DERECHOS

4.1. Un primer acercamiento a la ubicación del derecho frente a la IAF y el poshumanismo lo planteó Robert Alexy<sup>(40)</sup> al abordar la cuestión analizando un episodio de la serie *Star Trek: Next generation (Viaje a las estrellas: la nueva generación)* titulado “La medida de un hombre”. En dicho capítulo, el teniente comandante Data segundo oficial de la USS Enterprise que es un androide (una máquina con forma humana) recibe la orden de aceptar ser desmontado para poder producir en masa androides con su calidad a efectos de servir a la Federación. Cuando Data se niega a participar del experimento se organiza un juicio en donde se resolverá su destino asignándole al Capitán Picard la función de abogado defensor. La principal estrategia que utiliza Picard es demostrar que Data ofrece razones para ampliar el universalismo de los derechos humanos respecto de personas que no son humanas; en otras palabras, que los derechos no terminan en la frontera del *homo sapiens* sino que se extienden a las personas que demuestren que tienen conciencia en las tres dimensiones de la reflexividad: la cognitiva (cuando alguien hace de sí mismo objeto de su conocimiento), la volitiva (cuando alguien tiene la capacidad de regir su comportamiento mediante actos de voluntad) y la normativa (vinculada al auto enjuiciamiento bajo el aspecto de corrección). Picard gana el juicio y Data no es desmontado, pero quizás lo más notable es que rechaza la oferta que le hace su acusador de transferir todas las informaciones de su cerebro positrónico con la justificación de que con ello perdería “la sustancia, la cualidad, el sabor del momento” que se disfrutaban a través de la existencia.

4.2. El Comité de Derechos Económicos, Sociales y Culturales en la Observación General nro. 25 (2020) abordó el contenido del derecho que toda persona titulariza al goce de los beneficios del progreso científico y de sus aplicaciones en los términos previstos por el art. 15, párr. 1 b) del Pacto Internacional de Derechos Económicos, Sociales y Culturales.

El Comité entiende conceptualmente por “ciencia” (abarcando tanto a las ciencias naturales como a las sociales) se refiere tanto a un proceso que sigue una determinada metodología (“hacer ciencia”) como a los resultados de ese proceso (conocimiento y aplicaciones). El conocimiento se debe considerar como ciencia solo si se basa en una investigación crítica y está abierto a la falsabilidad y la comprobación; en tanto que el conocimiento basado únicamente en la tradición, la revelación o la autoridad sin la posibilidad de contraste con la razón y la experiencia o que

(40) Alexy, Robert, “Data y los derechos humanos. Mente positrónica y concepto dobletriádico de persona”, *Star Trek y los derechos humanos*, Tirant lo Blanch, Valencia, 2007.

sea inmune a toda falsabilidad o verificación intersubjetiva no se puede considerar ciencia<sup>(41)</sup>.

La ciencia aplicada incluye a la tecnología de la información y las comunicaciones<sup>(42)</sup>.

El derecho consagrado en el art. 15, párr. 1 b) en cuanto a gozar de los beneficios del progreso científico y de sus aplicaciones se vincula con el disfrute de las aplicaciones de la investigación científica, el acceso los conocimientos y la información científicos derivados directamente de la actividad científica y al papel que cumple la ciencia en la formación de ciudadanos críticos y responsables capaces de participar plenamente en una sociedad democrática<sup>(43)</sup>.

Como la cultura es un concepto inclusivo que comprende todas las expresiones de la existencia humana; por ende, la vida cultural es más amplia que la ciencia puesto que incluye otros aspectos de la existencia humana. Esto hace que sea razonable incluir la actividad científica en la vida cultural y que el derecho de todos a participar en la vida cultural incluya el derecho de toda persona a participar en el progreso científico y en las decisiones relativas a su dirección<sup>(44)</sup>.

En el punto E destinado a los riesgos y promesas de las nuevas tecnologías emergentes es donde el Comité realiza un aporte de nítida identidad transhumanista cuando sostiene lo siguiente:

“72. Los cambios tecnológicos son ahora tan intensos y rápidos que están desdibujando las fronteras entre los mundos físico, digital y biológico, debido a la creciente fusión de los avances científicos y tecnológicos en esferas como la inteligencia artificial, la robótica, la impresión en 3D, la biotecnología, la ingeniería genética, las computadoras cuánticas y la gestión de macrodatos. *Estas innovaciones podrían cambiar no solo la sociedad y el comportamiento humano, sino incluso a los propios seres humanos, mediante la ingeniería genética o la incorporación en el cuerpo humano de dispositivos tecnológicos que transforman algunas funciones biológicas*” (el destacado es propio).

Luego de una postura tan firme sobre los efectos evolutivos de la tecnología respecto de los seres humanos, el Comité expresa que estas tecnologías emergentes podrían mejorar el disfrute de los DESC posibilitado la cura o el tratamiento de muchas enfermedades a través de la biotecnología.

---

(41) Acápito 5.

(42) Acápito 7.

(43) Acápito 8.

(44) Acápito 10.

logía. También alerta sobre ciertos riesgos tales como la desigualdad social al aumentar el desempleo y la segregación en el mercado laboral<sup>(45)</sup> y un eventual reforzamiento de la discriminación mediante los algoritmos incorporados en la inteligencia artificial<sup>(46)</sup>.

(45) Nota del director: muchos autores en esta obra han abordado los efectos de las nuevas tecnologías y la automatización en el mercado laboral. En primer lugar, estos se analizan en Etcheverry, Milagros, “El impacto de la inteligencia artificial en el mundo del trabajo”, tomo III del presente *Tratado*. Se ha planteado que la pérdida de trabajos será causada en mayor medida por los robots industriales y, si bien no se adopta una visión completamente tecno-pesimista, se propone una postura de alerta y cautela, ver López Oneto, Marcos, “Fundamentos antropológicos, éticos, filosóficos, históricos, sociológicos y jurídicos para la constitución universal de un Derecho de la Inteligencia Artificial (DIA)”, tomo I del presente *Tratado*. También desde una óptica más pesimista, se ha hecho hincapié en que los impactos de las innovaciones en el mercado laboral pueden resultar en una disminución de hasta un 47% de puestos de trabajo en las próximas décadas, examinando así las transformaciones que se producen en la estructura empresarial y en la organización del trabajo como la precarización y reducción de complejidad de las tareas, ver Soares Valentini, Rómulo, “Industria 4.0 y gestión laboral por inteligencia artificial. Las plataformas de trabajo y los nuevos desafíos del derecho laboral”, tomo III del presente *Tratado*. No obstante, se han manifestado opiniones positivas, postulando que el temor por la pérdida de empleos en manos de la inteligencia artificial resulta infundado. Particularmente se postula que la profesión de la abogacía “sobrevivirá” y seguirá siendo útil, sin perjuicio de que los abogados deban formarse y adaptarse al nuevo paradigma, ver Braun Pellegrini, Ezequiel, “Brevity. Más tiempo para los abogados del futuro”, tomo III del presente *Tratado*. En el mismo sentido, se ha expresado que, aunque disminuya la demanda o relevancia del abogado en su ejercicio tradicional, no significa que haya menos puestos de trabajo para abogados en el futuro, ver Marsiglia, Gabriella - Renella, María Paula, “Inteligencia profesional: un diferencial frente a la IA”, apart. 3 “¿Y los abogados?”, tomo III del presente *Tratado*. Desde una perspectiva ética se ha postulado que se debe garantizar a los trabajadores que perdieron empleos ayudas no económicas como formación y acceso a derechos básicos, ver Martino, Antonio, “Ética y sistemas inteligentes”, apart. “¿Y el trabajo?”, tomo I del presente *Tratado*. Al examinar el rol de la IA en el ámbito de educación, se ha dicho que, lejos de promover el fin del trabajo humano, la IA complementará el trabajo del profesor en el aula, ver Schapira, Débora, “Inteligencia artificial: un nuevo desafío para la educación”, apart. 4 “Acerca del futuro del empleo”, tomo III del presente *Tratado*. Finalmente, se ha evaluado el impacto en el mercado de trabajo no solo del cambio tecnológico, sino también del envejecimiento poblacional y la pandemia del COVID-19 desde una perspectiva del derecho tributario en Iriarte Yanicelli, Adolfo A. “Inteligencia artificial y derecho tributario”, tomo IV del presente *Tratado*.

(46) Acápites 73.

Nota del director: sobre el reforzamiento de sesgos y discriminación por parte de la inteligencia artificial ver Sánchez Caparrós, Mariana, “Inteligencia artificial, sesgos y categorías sospechosas. Prevenir y mitigar la discriminación algorítmica”, tomo I del presente *Tratado*. Asimismo, ver Tolosa, Pamela, “Del derecho en los tiempos de IA (e IA generativa). Sobre estereotipos, sesgos y desigualdades”, tomo I del presente *Tratado*.

Si bien no hay soluciones sencillas dada las características exponenciales que exhiben las nuevas tecnologías y sus complejos efectos, el Comité propone tres elementos a ser tenidos en cuenta: a) la cooperación internacional puesto que dichas tecnologías requieren una regulación mundial para que se puedan gestionar eficazmente; b) las decisiones relativas al desarrollo y la utilización de esas tecnologías se deben adoptar en un marco de derechos humanos y desde una perspectiva holística e integradora: todos los principios transversales de los derechos humanos, como la transparencia, la no discriminación, la rendición de cuentas y el respeto de la dignidad humana resultan cruciales en este ámbito; c) el establecimiento por parte de los Estados de un marco jurídico que imponga a los agentes no estatales la obligación de la diligencia debida en materia de derechos humanos especialmente en el caso de las grandes empresas de tecnología; d) la adopción de políticas públicas que protejan a las personas vulnerables y el control de los datos masivos<sup>(47)</sup>.

4.3. López Oneto sostiene que el derecho a la inteligencia artificial es la continuidad y desarrollo de la naturaleza humana cultural definida como la autoimagen del hombre construida culturalmente y que puede ser inferida de los instrumentos internacionales sobre derechos humanos que es donde consta el consenso más general y universal en torno a la idea del hombre y de la humanidad<sup>(48)</sup>. Esta clase de posturas que conllevan una crítica frontal a cierta clase de transhumanismo tienen una precisa limitación: no pueden ser atendidas por los que no creen en la existencia de ese “orden natural” prefijado (aún enmascarado en la construcción cultural) o no encuentran ningún sentido moral plausible a las expresiones “jugar a ser Dios”, “ir contra la dignidad humana” o “desconsiderar el carácter donado de la vida”. Con el transhumanismo lo que está en discusión es si realmente existe un orden natural inviolable o una naturaleza humana estable con carácter normativo<sup>(49)</sup>. Justamente el transhumanismo quiere aplicar la tecnología para superar los límites impuestos por nuestra herencia biológica y genética, no observa a la naturaleza humana como un fin en sí mismo, ni como un ente perfecto, ni

(47) Acápites 74 a 75.

Nota del director: sobre la inteligencia artificial como parte de una política pública ver Vilas Díaz Colodrero, Jorge, “La necesidad de establecer la inteligencia artificial como política pública para promover su potencialidad y mitigar sus riesgos”, tomo I del presente *Tratado*.

(48) López Oneto, Marcos, ob. cit., p. 151.

Nota del director: asimismo, sobre el derecho de la inteligencia artificial ver López Oneto, Marcos, “Fundamentos antropológicos, éticos, filosóficos, históricos, sociológicos y jurídicos para la constitución universal de un Derecho de la Inteligencia Artificial (DIA)”, tomo I del presente *Tratado*.

(49) Diéguez, Antonio, ob. cit., p. 135.

como poseedora de algún derecho a nuestra lealtad, sino tan solo como un punto en el camino evolutivo<sup>(50)</sup>.

4.4. Las normas consideran como un ente separado al humano y a la tecnología, entendiéndose, por ejemplo, que un humano y un celular no están conectados y que solo estamos ante una persona que está utilizando una máquina; el problema que en la actualidad emerge es la afectación de la tecnología incorporada en ciertas personas denominadas cibernéticos, puesto que se argumenta que no perjudican a las personas sino solamente a los dispositivos, sin tener en cuenta que estos están incrustados en el mismo cuerpo humano pudiendo causar daños a la persona. En el caso Harbisson cualquier lesión o menoscabo a su antena se puede categorizar simplemente como un “daño a las cosas” y considerándolo bien eso debería evolucionar a un daño más trascendental pues al ser parte de su cuerpo y ser una extensión del mismo se estaría afectando sus derechos humanos.

4.5. El antropocentrismo como elemento exclusivo en la determinación de la titularidad de los derechos sufrió un duro golpe a partir del caso “Orangutana Sandra” que configuró el inicio fundacional de los derechos de las personas no humanas basados en la sintiencia (esto es la capacidad de sentir dolor y experimentar placer) como concepto estructural. La implosión de la sintiencia en la arrogancia de la triqueta antropocéntrica inteligencia + conciencia + plan de vida autónomo como única alternativa posible en torno a la titularidad de los derechos hizo que el transhumanismo recibiese un preciado insumo en su formulación, como lo es, la sintiencia electrónica. Este nuevo concepto se proyecta en dos universos. El primero vinculado a la paulatina incorporación de tecnología en el cuerpo del *homo sapiens* con el objetivo de su mejoramiento. El segundo relacionado con el advenimiento de las IA autónomas con capacidad de decidir y de sentir electrónicamente<sup>(51)</sup>.

4.6. Bostrom considera que la dignidad humana y posthumana son compatibles y complementarias<sup>(52)</sup>.

(50) More, Max, “The Philosophy of Transhumanism”, en AA.VV. - More, M. - Vita, N., *The Transhumanism Readers: Classical and Contemporary Essays on the Science, Technology and Philosophy of the Human Future*, Chichester, John & Sons, 2013, p. 4.

(51) Gil Domínguez, Andrés, ob. cit., p. 189.

(52) Bostrom, Nick, “In defense of posthuman dignity”, <https://www.nickbostrom.com/ethics/dignity.html>.

Nota del director: asimismo, sobre la tesis de Bostrom sobre el riesgo del fin de la humanidad, ver López Oneto, Marcos, “Fundamentos antropológicos, éticos, filosóficos, históricos, sociológicos y jurídicos para la constitución universal de un derecho de la inteligencia artificial”, tomo I del presente *Tratado*.

La dignidad en un sentido moderno consiste en lo que somos y lo que tenemos el potencial de convertirse no en nuestro origen causal. Lo que somos no es una función únicamente de nuestro ADN, sino también, de nuestro contexto tecnológico y social. La naturaleza humana en el sentido más amplio es dinámica, parcialmente creada por el hombre y mejorable. Nuestros fenotipos actuales son marcadamente diferentes del fenotipo del cazador-recolector. Leemos y escribimos, vestimos ropa, vivimos en ciudades ganamos dinero y compramos comida en el supermercado, hablamos por medio de teléfono celulares, vemos televisión, leemos diarios y revistas, conducimos automóviles, pagamos impuestos, votamos en elecciones, las mujeres paren en modernos hospitales, la esperanza de vida es tres veces mayor que en los Pleistoceno, sabemos que la Tierra es redonda, que las estrellas son grandes nubes de gas iluminadas desde el interior por fusión nuclear, que el universo tiene aproximadamente 13.700 millones de años y es enormemente grande. A los ojos de un cazador-recolector somos un posthumano. Sin embargo, estas extensiones radicales de las capacidades humanas —algunas de ellas biológicas, otras externas— no nos despojaron de nuestro estatus moral o nos deshumanizaron en el sentido de hacernos indignos. De manera similar, si nosotros o nuestros descendientes algún día logramos convertirnos en aquello a lo que, en relación con los estándares actuales podemos referirnos como posthumanos, esto no implicaría necesariamente la pérdida de la dignidad.

Desde el punto de vista transhumanista, no hay necesidad de comportarse como si hubiera una profunda diferencia moral entre la tecnología y otros medios para mejorar la vida humana. La dignidad posthumana promueve una ética más inclusiva que intenta abarcar a las futuras personas modificadas tecnológicamente y a los humanos contemporáneos.

4.7. En Scottsdale (Arizona) existe *Alcor Life Extension Foundation* dirigida por Max Moore autor de la “Carta a la naturaleza” donde propone una serie de enmiendas a la “constitución humana” entre las cuales se destacan la insurgencia frente a la tiranía de la vejez y la muerte a través de las herramientas biotecnológicas que permitan eliminar nuestra fecha de caducidad, como así también, el aumento de los poderes de percepción y cognición por medio de reformas sensoriales y de las capacidades neuronales. Alcor es la mayor de las cuatro instalaciones de criopreservación que existe en el mundo, tres de las cuales se encuentran en Estados Unidos mientras que la cuarta está en Rusia. En ellas se alojarán las personas que han dispuesto que sus cuerpos sean trasladados tan pronto como sea posible una vez determinada su muerte clínica para que se realicen con ellos una serie de procedimientos —incluyendo en la mitad de los casos la separación de la cabeza del cuerpo— que permita su suspensión criónica hasta que la ciencia encuentre una manera de devolverlos a

la vida. En la actualidad, Alcor tiene diecisiete “pacientes” o “crionautas” a los cuales se los considera suspendidos en una especie de estado atemporal entre este mundo y lo que venga después. La premisa central de la criónica es que la verdadera muerte se produce cuando las estructuras químicas del cuerpo empiezan a desintegrarse hasta un punto en que ninguna tecnología pueda restaurarlas a su estado original<sup>(53)</sup>.

En los EE.UU. junto a Alcor existe desde 1976 el *Cryonics Institute* en Detroit. Ambas eran las únicas organizaciones de crioconservación en el mundo, hasta que en 2005 se fundó *KrioRus* en las afueras de Moscú bajo el liderazgo del futurista ruso Daniel Medvedev quién ha logrado en poco más de una década crioconservar a medio centenar de personas y decenas de mascotas.

En 2016 una adolescente británica de 14 años que padecía cáncer le solicitó a un juez que deseaba ser crioconservada para tener una segunda oportunidad de vivir. “JS” le envió una carta al juez actuante donde expresó lo siguiente: “Se me ha pedido que explique por qué quiero hacer una cosa tan inusual. Sólo tengo 14 años y no quiero morir, pero sé que voy a morir. Creo que ser criogenizados nos da la oportunidad de ser curados y despertados incluso en cientos de años. No quiero ser enterrada bajo tierra. Quiero vivir y vivir más tiempo y creo que en el futuro pueden encontrar una cura para mi cáncer y despertarme. Quiero tener esa oportunidad. Este es mi deseo”. La madre estuvo de acuerdo con el deseo expresado por la adolescente, pero el padre —con quién JS no se hablaba desde seis años antes de padecer el cáncer— se opuso, lo cual derivó en un proceso judicial. El padre argumentó ante el tribunal: “Incluso si el tratamiento tiene éxito y ella es devuelta a la vida en, digamos, 200 años, no encontrará a ningún familiar, ni podrá recordar muchas cosas y, además, su situación puede convertirse en desesperada, ya que será una chica de solo 14 años sola en Estados Unidos”. El juez Peter Jackson, previo a tomar una decisión, visitó a JS en el hospital y posteriormente hizo lugar a su pretensión. JS falleció sabiendo que sería crioconservada. En octubre de 2016 su cuerpo fue cargado en un contenedor con hielo seco en el *Cryonics Institute*.

4.8. A lo largo de la historia de la humanidad, innumerables hombres y mujeres han soñado y perseguido poder vencer a la muerte. Estos deseos fueron receptados por la literatura, el arte, el cine, los cómics y actualmente cada vez con mayor intensidad por los envíos de las plataformas digitales en variados formatos intentando anticipar distintos escenarios y conflictos. Condenado a morir desde el instante del nacimiento vencer a la muerte podría ser un pase a la eterna felicidad o a una condena impo-

(53) Ferry, Luc, ob. cit., p. 33.

sible de revertir. Es que ser inmune a la muerte en un mundo de mortales puede llegar a ser traumático para la existencia del inmortal que observa como el paso del tiempo se devora los afectos. En la película “Highlander, el último inmortal” uno de los momentos más logrados es cuando ante la pérdida de la esposa del protagonista inmortal suena la canción de Queen con la voz potente y sensible de Freddie Mercury diciendo “Who wants to live forever” (quién quiere vivir para siempre). Otro punto de vista lo aportó José Saramago con su magnífico libro *Las intermitencias de la muerte*, donde narra las consecuencias distópicas que genera en una sociedad que la muerte no quiera matar a ninguna persona más.

En relación con el triunfo sobre la muerte vale la pena aclarar que existen dos modalidades victoriosas. El inmortal que bajo ninguna circunstancia puede morir tal como fue representado por ese notable comic que fue “Gilgamesh, el inmortal” protagonizado por un rey sumerio que alcanza la inmortalidad otorgada por seres de otros planetas. El amortal que puede vivir eternamente, pero que bajo circunstancias puede fallecer tal como sucedía con Highlander si era decapitado por otro inmortal o con Drácula<sup>(54)</sup> si era expuesto al sol o se le clavaba una estaca en su corazón.

En nuestra realidad la principal causa de muerte en el planeta es el envejecimiento y las enfermedades relacionadas con el envejecimiento que conducen a la muerte. No hay ningún principio científico en biología, química o física que prohíba el rejuvenecimiento e imponga la necesidad de la muerte. Por dicho motivo, a partir del desarrollo científico se plantea como una alternativa posible la longevidad indefinida o la muerte de la muerte mediante tratamientos biotecnológicos para el rejuvenecimiento humano que se comercializarán en la década de 2020, a los que seguirán en 2030 los tratamientos nanotecnológicos hasta llegar a controlar y revertir el envejecimiento en 2045.

En *The Fable of the Dragon-Tyrant* Nick Bostrom muestra con profundidad la lucha que libra el transhumanismo contra la tiranía de la finitud. Había una vez un reino tiranizado por un malvado dragón que muchos por años intentaron combatir con poco éxito porque parecía invencible. El dragón le exigía a la humanidad diez mil hombres y mujeres elegidos al azar que debían ser enviados todos los atardeceres al pie de una montaña para ser devorados. El reino aceptó esa vida miserable y la humanidad lo admitió como un irreversible hecho de la vida. También empezó a enseñarles a los niños y niñas que este era un destino natural, que debían vivir una buena vida y que el dragón con su voracidad ayudaba a evitar la

---

(54) Rosatti, Horacio, *Ensayo sobre la muerte. Drácula, o el precio de la inmortalidad*, Taurus, Ciudad Autónoma de Buenos Aires, 2019.

superpoblación. Al mismo ritmo que la humanidad aumentaba en población se incrementaba el apetito del dragón, lo cual devino en el armado de una burocracia que respondía a esta realidad. Hasta que un día, como la historia de la humanidad es el reflejo del constante progreso basado en la innovación, surgió un material que podía penetrar las escamas de la coraza supuestamente invencible del dragón y comenzó la construcción de un proyectil que pudiera acabar definitivamente con él. Sin embargo, esto sería costoso, complicado y demandaría bastante tiempo. Al principio, como la gente del reino era escéptica, el rey decidió convocar a una audiencia pública a la cual concurrió todo el pueblo. En una sala atestada, el consejero real aconsejó aceptar la inevitabilidad del dragón, lo siguió la dragonóloga en jefe que apostó a la realización del proyecto en un plazo de quince a veinte años, el último en hablar fue el ministro real de la moralidad, quien sostuvo que deshacerse del dragón desvalorizaría nuestra dignidad y desviaría nuestra atención de vivir una buena vida como destino inalterable. El salón quedó en silencio hasta que un niño exclamó: “¡El dragón es malo, mata gente... se comió a mi abuela... ¡Quiero a mi abuela de vuelta!”. El salón quedó en silencio. Luego una mujer expresó: “El dragón mató a mis padres”. Un hombre se levantó y dijo: “El dragón mató a mi esposa y a mi hija”. El peso del dolor impuesto por el dragón se multiplicó por el salón hasta que el rey mirando al niño que había expresado las primeras palabras dijo: “¡Hay que matar al dragón!”. A partir de la mañana siguiente la gente se percató de la gran cantidad de seres queridos que serían devorados por el dragón hasta que se construyera el proyectil que pudiera matarlo. Esto hizo que empezara una campaña para lograr lo que se había convertido en la máxima prioridad en la mente de todos: matar al dragón. Se recaudaron fondos, se logró el compromiso total del rey y así comenzó una gran carrera tecnológica contra el tiempo con varios intentos fallidos y etapas de mucha frustración social. Hasta que un día se logró el primer prototipo que se sería lanzado contra el dragón en la víspera del año nuevo, doce años después del inicio del proyecto. Cuando se estaba por lanzar el proyectil ante una multitud expectante un funcionario del reino pudo llegar hasta el rey para pedirle que parara el tren que viajaba hacia el dragón con el último contingente de personas a ser devoradas. El rey con mucha pena le dijo que no era posible porque si el dragón se movía esto podía cambiar el ángulo de tiro y el proyectil no lo impactaría. Con pesar, el rey asumió que si hubieran empezado muchas personas se hubieran salvado de ser devoradas y del horror que el reino tuvo que sufrir. Era un día lluvioso, luego de un preciso conteo, el proyectil se lanzó y en su estela luminosa estaba representado el espíritu humano, sus miedos y esperanzas. El impacto fue directo al corazón del dragón. Miles de voces se alzaron al unísono en alegría seguida por el estruendo ensordecedor y eterno del monstruo colapsando. Después de tanto tiempo, la humanidad era finalmente libre de la tiranía del dragón. “Ahora tenemos

tiempo”, dijo el rey. Tiempo para crecer, para hacer las cosas bien, para aprender de los errores, para construir un mundo mejor.

La principal causa de muerte en nuestro planeta es el envejecimiento. En el mundo mueren día tras día aproximadamente 150.000 personas fallecen por el envejecimiento y por enfermedades relacionadas con el envejecimiento que se presenta, tal vez, como el mayor enemigo de la humanidad. Existen diferentes organismos que no envejecen, las mejores células de nuestros cuerpos (las germinales y las madres) no envejecen, las “peores” células (las de cáncer) tampoco envejecen. Por eso el interrogante que surge es el siguiente: ¿cuándo será posible detener el envejecimiento en los seres humanos? La clave está centrada en la posibilidad de desarrollar terapias médicas con el objeto de poder revertir el envejecimiento biológico en humanos de manera que podamos seguir acumulando años de edad mientras nos mantenemos biológicamente jóvenes enfrentando la senescencia en lo que Kurzweil denomina “la velocidad de escape de la longevidad”. En poco tiempo estarán disponibles los primeros **tratamientos** biotecnológicos para el rejuvenecimiento humano, luego vendrán los **tratamientos** nanotecnológicos hasta llegar a controlar y revertir el envejecimiento<sup>(55)</sup>.

El biomejoramiento humano incluye dos posibilidades distintas de concreción: la química (o farmacológica) y la genética. A su vez que cada una de ellas puede estar dirigida al mejoramiento de las capacidades físicas (rendimiento deportivo, resistencia de enfermedades, extensión de la vida, etc.) o de las capacidades mentales o conductuales (mejoramiento cognitivo, emocional y moral). Los principales argumentos utilizados en torno a justificar el mejoramiento que aumentará el bienestar o la calidad de vida de las personas son los siguientes<sup>(56)</sup>:

- La tecnología ha sido siempre un instrumento de mejoramiento. No solo proporcionó herramientas que han permitido transformar el mundo en beneficio del *homo sapiens*, sino que también permitió el desarrollo de la cultura humana gracias a la cual disfrutamos de “prótesis culturales” como la escritura, el arte, el derecho que amplían las capacidades cognitivas y transforman nuestras vidas.
- La evitación de mejoramiento implicaría en muchos casos desconocer un deber moral por cuanto el mejoramiento genético hará que nuestra vida sea más satisfactoria. Si los padres no permitie-

---

(55) Cordeiro, José Luis - Wood, David, *La muerte de la muerte. La posibilidad científica de la inmortalidad física y su defensa moral*, Deusto, Barcelona, 2018.

(56) Diéguez, Antonio, ob. cit., p. 124.

sen este mejoramiento para sus hijos cuando estuviera disponible: ¿los estarían dejando en una situación de desventaja comparativa y les estarían causando un perjuicio objetivo?

- Si es legítimo permitir las modificaciones genéticas con una finalidad terapéutica por qué no sería posible que dichas modificaciones se orienten al mejoramiento.
- Nadie se opone al mejoramiento humano (intelectual, psicológico, moral o físico) mediante “técnicas sociales o culturales” como la educación o el deporte. Ser más inteligente mediante la educación o a través de la ingeniería genética no ofrece ninguna diferencia ética relevante.
- No existe desde el punto de vista biológico una naturaleza humana universalmente compartida que pueda tomarse como base para fundamentar algún tipo de dignidad inviolable supuestamente amenazada por las mejoras genéticas. Incluso si la hubiera, su preservación no es una obligación moral puesto que podría haber en ella rasgos o patrones negativos que necesariamente deberían cambiar.
- No existe ningún argumento que justifique que es intrínsecamente mejor estar sometidos a la lotería genética a la que hemos estado sujetos a lo largo de nuestra existencia como especie que dejar la selección de los genes a la libre elección de los padres con regulaciones que lo limiten a la evitación de daños a terceros.
- La mejora genética, lejos de producir desigualdad, podría ser un instrumento eficaz para introducir una mayor igualdad en la sociedad puesto que haría que los menos dotados se acercaran a los más dotados o que desaparecieran las enfermedades genéticas como fuente de desigualdad.
- El desarrollo y aplicación al *homo sapiens* de las tecnologías de mejoramiento genético es inevitable en la medida que constituyen el siguiente paso en el proceso evolutivo de nuestra especie, estando toda resistencia condenada al fracaso.

4.9. ¿Es posible ampliar radicalmente las capacidades cognitivas mediante una emulación cerebral completa? ¿Con el tiempo podremos trasladar nuestros procesos mentales a un sustrato computacional más adecuado? El procedimiento aplicable debería escanear la información del cerebro de una persona a través de cualquier tecnología o combinación de tecnologías que resulte viable (tales como los nanorobots o el microscopio electrónico), luego la información escaneada se convertiría en una maqueta para la reconstrucción de las redes neuronales del cerebro

del sujeto que a la vez se transforma en un modelo informático, y, por último, este resultado se emula en un tercer sustrato no basado en la corporalidad humana como podría ser una supercomputadora o un robot humanoide<sup>(57)</sup>.

4.10. ¿La inteligencia artificial configura un riesgo existencial para la humanidad? Elon Musk, Stephen Hawking, Peter Thiel y Bill Gates han manifestado públicamente que la inteligencia artificial debería ser analizada desde la perspectiva del riesgo potencial del fin de la humanidad. El Instituto para el Futuro de la Humanidad (Oxford), el Centro para el Estudio del Riesgo Existencial (Cambridge), el Instituto de Investigaciones de la Inteligencia de las Máquinas (Berkeley) y el Instituto para el Futuro de la Vida (Boston) están desarrollando algoritmos que nos ayuden a identificar cómo será posible evitar este destino<sup>(58)</sup>. Si las cosas cambiasen de forma radical en la relación entre la humanidad y las máquinas a partir de la aparición de máquinas superinteligentes quizás el riesgo no fuese que sean hostiles con sus predecesores humanos, sino que simplemente mostrarían una absoluta indiferencia frente a ellos, de forma tal, que tendrían con los humanos una relación similar a la que nosotros tenemos con los animales<sup>(59)</sup>. El riesgo fundamental frente a la inteligencia artificial está en no poder definir explícitamente nuestros planes de una forma lógicamente rigurosa tal como sucedió con el mito de Rey Midas cuyo deseo era tener la capacidad selectiva de convertir las cosas en oro con sólo tocarlas pero lo que pidió y fue concedido por Dionisio fue la incapacidad de evitar convertirlas en oro al tocarlas; en este sentido, si a una IA capaz de resolver los problemas científicos más difíciles le decimos que elimine el cáncer para siempre, la máquina pondrá manos a la obra y llegará a la conclusión que la forma más eficaz de cumplir con el objetivo es eliminar a todas las especies en las que podría producirse la división incontrolada de células anormales, y de esta manera, hacer desaparecer a todas las especies sintientes de la faz de la tierra<sup>(60)</sup>.

## V. ¿EL HOMO SAPIENS ES EL TODO O ES SOLO EL COMIENZO?

5.1. Un dato objetivo nos permite reflexionar. Desde hace unos dos millones de años hasta aproximadamente diez mil años el mundo fue el hogar concurrente de varias especies humanas. Hace cincuenta mil años sapiens, neandertales y denisovanos se hallaban en un punto limítrofe biológico. Hace dos millones de años algunos de los primeros hombres

(57) Ferry, Luc, ob. cit., p. 63.

(58) *Ibíd.*, p. 95.

(59) *Ibíd.*, p. 99.

(60) *Ibíd.*, p. 115.

y mujeres arcaicos dejaron su tierra natal —el África oriental— para desplazarse a través de extensas áreas del norte de África, Europa y Asia con el fin de instalarse en ellas. Los humanos en Europa y Asia occidental evolucionaron en *homo neanderthalensis* (conocidos popularmente como “neandertales”). Las regiones más orientales de Asia estaban pobladas por *homo erectus* que sobrevivió en dicho hábitat durante cerca de dos millones de años lo que hace de ella la especie humana más duradera de todas siendo muy improbable que este récord sea batido por el *homo sapiens* puesto que es muy dudoso que todavía persista dentro de mil años (de forma tal que dos millones de años quedan fuera de toda posibilidad para el *homo sapiens*). Durante los últimos diez mil años el *homo sapiens* se acostumbró a ser la única especie humana lo cual impide concebir alguna otra posibilidad<sup>(61)</sup>.

5.2. Un abordaje ficcional notable lo ofrece el tratamiento que se hace de los replicantes en las películas *Blade Runner* (1982) y *Blade Runner 2049* (2017) basadas en la novela de Philip K. Dick, *¿Sueñan los androides con ovejas eléctricas?* (1968). En la primera, los replicantes (creados mediante sofisticada ingeniería genética bajo el lema “más humanos que los humanos”, similares en su aspecto físico a los humanos pero mucho más ágiles y fuertes, aunque sin la misma respuesta emocional y empatía) destinados a trabajos extremos en condiciones de esclavitud, luego de un sangriento motín acaecido en Marte, vuelven a la Tierra comandados por Roy Batty en busca de su creador humano (el dueño de Tyrrel Corporation) porque como tienen una fecha de vencimiento que se aproxima desean extender su vida para seguir disfrutando de la existencia. Se trata de seres más evolucionados que Data que no quieren morir y luchan por el derecho a la vida. La escena final, en la cual Roy le perdona la vida al agente Deckard (el *Blade Runner* cuya misión era eufemísticamente “retirarlo”) en una azotea de un edificio bajo una copiosa lluvia plomiza porque siente que su tiempo se está expirando culmina con un monólogo impactante que refleja emoción, dolor y búsqueda de un resguardo normativo de la subjetividad cuando dice: “He visto cosas que ustedes nunca hubieran podido imaginar; naves de combate en llamas en el hombro de Orión. He visto relámpagos resplandeciendo en la oscuridad cerca de la entrada de Tannhäuser. Todos esos momentos se perderán... en el tiempo... igual que lágrimas... en la lluvia. Llegó la hora de morir”. En esa escena final Deckart reconoce en ese ser que está agonizando no a una simple máquina, sino un semejante que encarna todos los valores humanos que él ya ha perdido. Esa mirada de Deckart constituye la tesis ontológica y ética que subyace en la narración cinematográfica del posthumanismo:

(61) Harari, Yuval Noah, *De animales a dioses. Breve historia de la humanidad*, Debate, Ciudad Autónoma de Buenos Aires, 2013, ps. 18 y ss.

la biología tan solo identifica a los humanos como especie, pero la auténtica humanidad no se mide en genética sino fundamentalmente en valores. La condición de humano está más allá de la herencia biológica natural, puesto que la tecnología puede recrear y mejorar artificialmente la biología; la humanidad radica fundamentalmente en asumir una dimensión ética de la existencia. En consecuencia, merecerá ostentarla cualquier ser que haya tenido o tenga (ahora o en el futuro) la capacidad de pensar y sufrir, de establecer relaciones, de experimentar afectos, de creer y entregarse por ideales que trascienden el ámbito de su limitada existencia. Una tesis que se completa en la escena final cuando Deckart huye con Rachel, una replicante de última generación que desconoce su fecha de caducidad, pero que, en realidad, no tiene plazo de vida<sup>(62)</sup>. En la segunda, los replicantes son Blade Runner de última generación sin plazo de caducidad que tienen por misión “retirar” a los replicantes antiguos que todavía habitan de forma oculta en la Tierra, cuando uno de ellos llamado K comienza una investigación que lo lleva a desarrollar un conjunto de sentimientos impensados para su especie por los cuales decide luchar para finalmente morir. Entre una y otra película, lo más notorio del tiempo escenográfico transcurrido (2000-2030) es la evolución de los replicantes y las reivindicaciones normativas que persiguen en relación con los derechos que titularizan los humanos; mientras Roy deseaba solamente titularizar el derecho a la vida, K lucha por la justicia más allá de su propio interés. El final en el que se devela la identidad de un ser nacido de la relación entre el agente Deckard y la replicante Rachel considerado como un acto “milagroso” bien podría ser entendida como el comienzo del poshumanismo.

El eventual posible paso del *homo sapiens* a un homo poshumano está muy bien tratado en la película *Autómata* (2014). En un futuro distópico ubicado en el año 2044 las erupciones solares han desertificado al planeta Tierra y diezmado a la población, con el objeto de intentar la reconstrucción de la atmosfera los humanos crean robots denominados “peregrinos” sometidos a dos leyes básicas: deben proteger la vida y no pueden modificarse a sí mismos. Cuando los robots fracasan son relegados a cumplir tareas menores de servidumbre hasta que un policía afirma observa a un robot modificarse y adquirir vida propia y por dicho motivo lo destruye. Un agente de seguros de la compañía Robot Organic Century (ROC) que fabricaba los robots es destinado a realizar una investigación mediante la cual descubre que efectivamente estaban evolucionado. Hacia el final de la película se produce un diálogo fantástico

---

(62) Talavera, Pedro, “La narración cinematográfica del posthumanismo”, *Cuadernos de Bioética*, XXVI, 2015/2, Asociación Española de Bioética y Ética Médica, Madrid, 2015, p. 286.

entre el agente de seguros Jacq y la robot Cleo que intenta subir a un cable carril para cruzar un profundo barranco (donde antes corrió un gran río) hacia una zona radioactiva que los mantendrá a salvo de los humanos y en donde podrán desarrollar su existencia. Cuando Jacq le manifiesta su temor ante la muerte, Cleo le contesta que “morir es parte del ciclo natural humano su vida solo es un lapso de tiempo, porque tienes miedo tu vida se está acabado ninguna forma de vida puede habitar el planeta eternamente, mírame yo nací de las manos de un humano fui imaginado por mentes humanas ahora tu tiempo vivirá en nosotros y será el tiempo a través del cual existirás... sobrevivir no es relevante vivir lo es, queremos vivir...”

¿Dónde se afincan los mayores temores en torno a este futuro? Éric Sadin acierta en la descripción de los demonios que atormentan al *homo sapiens* cuando expresa que de ahora en adelante ciertos sistemas computacionales están dotados de una singular y perturbadora vocación: la de enunciar la verdad. La digitalidad no solo implica el almacenamiento, la indexación y la manipulación de corpus cifrados, textuales, sonoros o icónicos sino también opera en la divulgación de sentidos erigiéndose como un órgano habilitado para peritar lo real de modo más fiable que nosotros mismos revelando dimensiones ocultas a nuestra consciencia. Esta facultad deriva de las ciencias algorítmicas como factor determinante, las cuales tomaron un camino antropomórfico que busca atribuir a la innovación cualidades humanas reproductivas de nuestras aptitudes como nunca se le atribuyó a ningún artefacto en la historia de la humanidad. El desarrollo de la inteligencia artificial hace que la humanidad asuma a pasos acelerados la existencia de un órgano que prescinde del derecho que ella titulariza a decidir con plena conciencia y responsabilidad las elecciones que la involucran. En este punto, la ética debe ser entendida como un umbral mínimo direccionado al respeto incondicionado de la integridad y dignidad humana basada en el hecho de poder utilizar la autonomía, el juicio propio, la libertad de conciencia de los propios actos, el goce de las partes de uno mismo sin las miradas de los otros. Más que la “singularidad tecnológica” lo que se producirá es la “singularidad ontológica” mediante la cual se redefine la figura humana, su estatuto, sus poderes, sus derechos ¿Desaparecerá la intuición humana debido a la inteligencia artificial haciendo totalmente inútil o inoperante toda decisión que dependa de nuestra consciencia? ¿Está a punto de mutar la afirmación de la singularidad irreductible de cada ser? En oposición a la racionalidad que pretende erradicar el desorden, luchar contra la entropía y asentar un dominio cada vez más extendido sobre el curso de las cosas, Sadin propone mantener una racionalidad en donde las imperfecciones de la vida sean las que estimulan el deseo humano de realizarse y trabajar

sin descanso en la construcción de un mundo común que se base en el axioma cardinal de no dañar a nadie<sup>(63)</sup>.

Tal como lo expresa Yuk Hui no es posible hablar de un límite preciso a la inteligencia artificial si se considera que esta, en cuanto artificial, es más susceptible de poder mutar que la inteligencia humana cuyos mecanismos aún escapan a la comprensión. En otras palabras: ¿cómo poner un límite a un ente que virtualmente no tiene límites? Las máquinas son parte de un proceso evolutivo de la especie humana que los seres humanos tuvieron la capacidad de controlar, pero sobre el que están camino a perder el control. El fin de lo humano tiene menos que ver con la hipótesis de un reemplazo completo de seres humanos por las máquinas que con el hecho de que la inteligencia de las máquinas transforme a los humanos en una medida que excede nuestra imaginación. En este sentido Hui plantea dos interrogantes básicos: ¿Llevará esta transformación de lo humano a la extinción del *homo sapiens*? ¿O llevará, en cambio, a un punto reflexivo en el que no solo se impugne el humanismo, sino que pueda también reabrirse la cuestión de la historia, de la cultura y de la vida tal como hasta ahora la conocimos?<sup>(64)</sup>.

Uno de los conceptos más disruptivos de la lógica normativa antropocéntrica que propone el transhumanismo es el de “libertad morfológica” que consiste en la potestad que titularizan las personas humanas de adoptar cualquier forma corporal que permita la tecnología en el campo de la autodeterminación, la biografía, la mismidad, la intimidad.

La libertad morfológica junto al derecho al desarrollo científico y tecnológico son los desafíos actuales más frontales a la lógica antropocéntrica de los derechos humanos provenientes desde sus entrañas. Tal intenso es su efecto que las respuestas que se intentan esbozar, por más maquillaje que adopten, remiten al perfeccionismo, a una dimensión teológica de la naturaleza humana o al monismo moral. Distintas opciones que el discurso de los derechos humanos se esforzó por erradicar como límite razonable al plan de vida de las personas cuando no existiera un daño concreto a terceros. ¿Puede ser considerado un perjuicio a terceros proponer la mejora voluntaria de las personas a través de la incorporación de la tecnología en su cuerpo? ¿Es posible admitir la presencia de un holismo colectivista cultural como un límite razonable a los derechos humanos?

El transhumanismo es presente y lejos está de anidar en la ciencia ficción. Así como postula una evolución de los cuerpos del *homo sapiens*

(63) Sadin, Éric, *La inteligencia artificial o el desafío del siglo. Anatomía de un antihumanismo radical*, Caja Negra, Ciudad Autónoma de Buenos Aires, 2020.

(64) Hui, Yuk, *Fragmentar el futuro. Ensayos sobre tecnodiversidad*, Caja Negra, Ciudad Autónoma de Buenos Aires, 2020.

y una relación sistémica con las tecnologías convergentes, también demanda una deconstrucción normativa que corte amarras con un discurso jurídico basado exclusivamente en el antropocentrismo y temeroso en última instancia de un futuro inexorable.

Uno de los aspectos donde mayor impacto produce la visión transhumanista se refleja en el derecho constitucional tal cual lo conocemos, y, fundamentalmente, en uno de sus íconos fundantes: el concepto de persona proyectado al sujeto de derecho que titulariza derechos y cuentas con las garantías efectivas para su protección.

Las constituciones tales como las conocemos fueron forjadas en un mundo analógico que, poco a poco, se está yendo. El mundo digital es un nuevo escenario desconocido para los objetivos que persiguió el derecho constitucional, sino también, cambiaron los factores de poder que históricamente fueron tenidos en cuenta a la hora de elaborar los derechos y las garantías.

La Constitución regula un mundo que en parte no existe o es socialmente irrelevante. En estos términos, algunos preceptos constitucionales son una reliquia, otros fueron derogados por imperio del avance científico y tecnológico y muchos de los que subsisten reclaman por nuevos contenidos que reflejen una demanda de acciones y tutelas basadas en la digitalidad.

El problema no es que la Constitución regule una parte de la realidad —la analógica— que prácticamente no existe, sino que, y ahí está el gran problema, no regula la realidad digital que está configurando una clase de sociedad que vive en un mundo digital con nuevos agentes globales como lo son las empresas tecnológicas<sup>(65)</sup>.

En el centro de la sociedad se ubica como uno de los actores principales el algoritmo que es un instrumento configurador de la realidad digital que se expande a través del mundo físico y digital constituyendo, por ahora, un mundo híbrido. Mientras que la realidad analógica representa una dimensión estatal sometida al derecho público, la realidad digital configura un escenario global sometido a un derecho que todavía no se construyó y en el cual las compañías tecnológicas ocupan un lugar central.

Esto demanda la construcción de un constitucionalismo digital que, de alguna manera, empiece a dar sustento a un nuevo modelo o paradigma.

---

(65) Balaguer Callejón, Francisco, “La constitución del algoritmo. El difícil encaje de la constitución analógica en el mundo digital”, en AA.VV. - Balaguer Callejón, Francisco - Cotino Hueso, Lorenzo, *Derecho público de la inteligencia artificial*, Fundación Manuel Giménez Abad, Zaragoza, 2023, p. 32. <https://www.fundacionmgimenezabad.es/es/derecho-publico-de-la-inteligencia-artificial>.

ma que sería el Estado digital de derecho global en el cual los derechos que existen o deben ser existir en una esfera constitucional digital están permanentemente tamizados por los algoritmos en el plano del diseño y la utilización de estos<sup>(66)</sup>.

Un posible puente que puede ofrecer el derecho entre estos dos mundos podría ser reconfigurar el reconocido apotegma “nuevos contenidos en viejos derechos” en “nuevos contenidos basados en el desarrollo científico y tecnológico en viejos derechos siempre que sea posible, nuevos derechos basados en el desarrollo científico y tecnológico siempre que sea necesario”.

Un derecho constitucional digital donde el algoritmo cumple un rol esencial es conceptualmente singularista y transhumanista al aceptar y aceptar el desarrollo científico y tecnológico como parte de una escenografía alejada del antropocentrismo analógico y basada en la impronta adopta como punto de partida la mirada de Ray Kurzweil en cuanto que el cambio tecnológico tiene una velocidad y una profundidad (denominada “ley de los rendimientos acelerados” que se vincula con la rapidez del ritmo evolutivo donde justamente el progreso tecnológico es una continuación de la evolución biológica) que necesariamente modificará la vida humana de forma irreversible.

Algoritmo, singularidad tecnológica convergente, transhumanismo son los pilares de un discurso jurídico digital en construcción que pueda sostener la nueva regla de reconocimiento digital de un mundo de vidas globalizadas no solo por las empresas tecnológicas, sino también, por las biografías en clave digital y tecnológicas de las personas.

Del constitucionalismo digital a la interdicción del concepto antropocéntrico de persona y el sucedáneo de sujeto de derecho hay un solo paso.

Desde el origen de nuestra especie e incluso desde los neandertales la evolución humana se basó siempre en componentes orgánicos. El desarrollo científico y tecnológico que nutre al constitucionalismo digital produce un “mestizaje ontológico” entre lo biológico y lo técnico que hace necesaria una compenetración entre la tradicional filosofía carbónica (es decir, orgánica) del discurso jurídico con el hombre situado en el centro y una nueva filosofía silícica del derecho que incorpora la materia inorgánica antropoide como posible sujeto de derecho<sup>(67)</sup>.

---

(66) Gil Domínguez, Andrés, *Constitucionalismo digital*, Ediar, Ciudad Autónoma de Buenos Aires, 2023.

(67) Campione, Roger, *La plausibilidad del derecho en la era de la inteligencia artificial. Filosofía carbónica y filosofía silícica del derecho*, Dykinson, Madrid 2020, p. 13.

El derecho constitucional y el derecho convencional a través de los instrumentos internacionales sobre derechos humanos fueron construidos sobre la base de una figura central y excluyente —el *homo sapiens*— denotando un antropocentrismo analógico excluyente al momento de establecer la categoría sujeto de derecho. No existía una configuración distinta posible que pudiera alojar la estructura discursiva de lo jurídico, a partir de la cual se diseñó el sistema de derechos. Las luchas sociales desarrolladas por siglos tenían por objeto igualar a todos los seres humanos en la titularidad de sus derechos más allá de su procedencia social, ideología, religión, sexo, etc. Las reivindicaciones siempre circundaron a la persona humana nacida en el mundo analógico. El derecho del que se habla siempre fue el derecho humano hecho por seres humanos y para seres humanos.

Las máscaras escénicas de los actores del derecho siempre se llamaron personas. Actualmente, estamos en el umbral de la salida definitiva del nexo entre el hombre y la persona, de la superación de lo humano tal como lo hemos entendido y malentendido, de la entrada en una era habitada por personas no humanas o no del todo humanas. Una vez cruzado dicho umbral, es posible que no encontremos un escenario en el que las mil máscaras del derecho sigan siendo utilizadas por actores humanos. Entonces habremos entrado en la era de una forma de derecho que no es humana, o que ya no es humana o simplemente una era sin derecho<sup>(68)</sup>.

La revolución tecnológica digital genera una disrupción significativa a la centralidad antropocéntrica por cuanto actualmente no hay, como aconteció en el pasado, un solo modelo de persona respecto del cual el derecho constitucional y convencional deba prestar atención en términos una construcción realista del sujeto de derecho. Existen o se proyectan varias dimensiones de persona que deben ser tenidas en cuenta.

En primer lugar, está la persona que a través del uso de la tecnología —muchas veces como exocerebros esenciales para transitar la vida cotidiana— es una persona cuya humanidad aumento o transmutó del modo analógico al modo digital. Tecnología que en breve tiempo se incorporará masivamente en la biología humana como parte de un órgano más.

En segundo lugar, las personas ciborg, quienes, ejerciendo plenamente la libertad morfológica, incorporan tecnología de mejora en el cuerpo con el objeto de amplificar los sentidos y habilidades humanas.

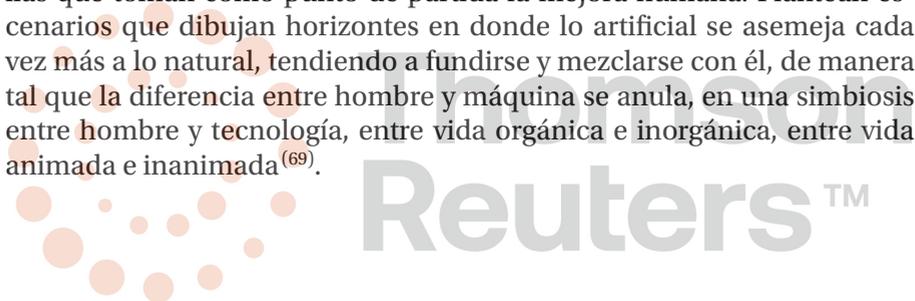
(68) Pietropaoli, Stefano, “En primera persona. Un réquiem por el derecho de la era digital”, AA.VV. - Llano Alonso, Fernando H. (dir.), *Inteligencia artificial y filosofía del derecho*, Laborum Ediciones, Murcia, 2022, p. 218.

En tercer lugar, las personas avatares que transitaran por el multiverso de los metaversos con características físicas y personales quizás distintas a la que tienen en el mundo real.

Por último, las personas digitales significadas en su máxima proyección a través de robots humanoides dotados de IA en sus distintas variantes que a través de la empatía antropomórfica generaran una intersubjetividad humana-digital inédita de predecir en términos de relaciones sociales y sentimientos.

Con este panorama no es posible seguir insistiendo con un concepto de persona afincada exclusivamente en una visión antropocéntrica analógica que se niega sistemáticamente a la apertura del término sujeto de derecho a otras entidades.

Las tecnologías que nos desafían socavando el frágil concepto de lo “humano” y la relación existente entre el hombre y la persona son aquellas que toman como punto de partida la mejora humana. Plantean escenarios que dibujan horizontes en donde lo artificial se asemeja cada vez más a lo natural, tendiendo a fundirse y mezclarse con él, de manera tal que la diferencia entre hombre y máquina se anula, en una simbiosis entre hombre y tecnología, entre vida orgánica e inorgánica, entre vida animada e inanimada<sup>(69)</sup>.



# INTELIGENCIA ARTIFICIAL BIFRONTE: ¿LIBERTAD O ESCLAVITUD?

HENOCH AGUIAR

## I. LA NUEVA CONCENTRACIÓN DEL PODER

Nadie discute que el poder de la IA está hoy en las manos del Estado y de las grandes corporaciones. Los titulares del poder, político y económico, son sus principales dueños y entusiastas usuarios. Concentran una doble capacidad. Administran las más importantes bases de datos, públicas y privadas, y dominan la tecnología para usarlas en su provecho.

Estado y grandes corporaciones son los acopiadores naturales de los datos personales, sociales y económicos. Sus bases de datos y plataformas captan naturalmente toda la información circulante. A esa condición original le suman la capacidad (que individuos o pequeñas organizaciones no alcanzan), de acumular recursos tecnológicos y humanos de alto vuelo para crear nuevas olas de algoritmos, cada vez más complejos e interactuantes.

Los algoritmos en funcionamiento son la nueva médula nerviosa y, de a poco, el cerebro mismo de las organizaciones o países más avanzados, más poderosos, los que controlan la acción y el devenir de los demás, que no supieron, o no pudieron, incorporar a tiempo esa capacidad de análisis, procesamiento e instantánea decisión.

## II. LA DISYUNTIVA ÉTICA

No queremos, no podríamos tolerar, ni ética, ni jurídicamente, una IA al servicio de la concentración de poder, de la manipulación y control de las personas. Necesitamos que aporte más humanidad, cuidado y respeto en el trato de cada uno. Saber con aguda precisión quién está del otro lado permite que la IA (quienes la diseñan) lo traten mejor, personalizando cada intercambio, para envolverlo de respeto, conociendo quién es, cuáles son sus necesidades.

La IA debe fijarse como objetivo conocer, respetar y reverenciar al otro, potenciar sus capacidades, motivarlo y darle herramientas de crecimiento.

Necesitamos una IA en busca de humanidad. Dicen los románticos que no se conoce bien a quien no se ama. La IA nos ayuda a conocer al otro más que nunca. Debe permitirnos tratarlo mejor que nunca. Si no fuera así, equivocamos el camino, rompiendo la sociedad en vez de construirla.

La tecnología siempre tuvo que enfrentar la disyuntiva ética. O bien piensa la medicina como Mengele, el Ángel de la Muerte en Auschwitz, o se guía por los valores de Albert Schweitzer, médico, filósofo y músico, que dedicó años de su vida, en Lambarené (hoy Gabón), a ser el único médico que atendía en cientos de kilómetros a la redonda.

El fin determina los medios. Por ello, la Unión Europea, el 23 de enero 2023, en la “Declaración europea sobre los derechos y principios digitales para la década digital” (2023/c 23/01), en los puntos 8 y 9, marca dos objetivos preliminares a todo análisis sobre el funcionamiento de los algoritmos y los sistemas de inteligencia artificial:

“La IA debe ser un instrumento al servicio de las personas y su fin último debe ser aumentar el bienestar humano.

”Toda persona debería estar empoderada para beneficiarse de las ventajas de los sistemas algorítmicos y de la IA, especialmente a fin de tomar sus propias decisiones en el entorno digital con conocimiento de causa, así como estar protegida frente a los riesgos y daños a su salud, su seguridad y sus derechos fundamentales.”

De esto se trata, no de contener *a posteriori* los males que puedan devenir de la aplicación descontrolada de la IA, sino de fijarse eficientemente el objetivo de generar mayor bienestar, empoderamiento y respeto de las personas al inicio de su diseño. El valor a alcanzar determina todos los mecanismos que se implementen después. La disyuntiva ética potencia los beneficios futuros de la IA si es asumida desde el momento de su concepción hasta su última aplicación.

### III. LA IA ESCALA POSICIONES

En el Estado y las corporaciones, los algoritmos se adoptan inicialmente para resolver una o varias temáticas, de manera compartimentada, limitada, funcional. Un algoritmo resuelve esto o aquello. Facilita, sintetiza, mejora, acelera procesos conocidos.

La planificación estratégica bien aplicada permite que las organizaciones más avanzadas puedan establecer diferentes niveles de propósito o de gestión, más elevados e integradores, que unifican la acción de los niveles inferiores, sean estos departamentos o gerencias en lo privado, direcciones, secretarías o ministerios en lo público.

La segunda etapa de implementación de la IA en las organizaciones es la creación de algoritmos transversales, de segundo nivel, que controlan a los algoritmos iniciales, más básicos. Por ser un sistema inteligente, este segundo nivel no solamente analiza resultados. Según la variable estratégica que se decida obtener, el algoritmo de segundo nivel podrá reconfigurar, reprogramar, los parámetros y lógicas de funcionamiento de los algoritmos de primer nivel.

#### IV. "EL PODER SOY IA" (COPYRIGHT LOUIS XIV)

Es el momento en que pasamos de un pacífico análisis de los niveles posibles de la IA a una perturbadora distopía de ciencia ficción. Basta con, anodidamente, reiterar la creación de niveles superiores de control y ver qué sucede.

Supusimos la creación de una capa de IA superior, "x+2" para regir los sistemas de IA básicos. Escalemos imaginariamente una nueva capa de control. Aparecerá la capa "x+3" para intervenir y modificar, "inteligentemente", los ya importantes niveles de decisión de la capa "x+2". Y así sucesivamente.

El término control queda corto. Digámoslo con propiedad. Quien puede, no solamente analizar lo decidido por un nivel inferior de decisión, sino que puede modificar y rediseñar dinámicamente los criterios que toma en cuenta su tarea, ya no es un controlador. Es alguien que "gobierna" a los niveles inferiores.

Los niveles más altos de IA, de aquí a diez o quince años, "gobernarán", con la autonomía y capacidad de innovación con que fueron programados, el entramado piramidal de los sistemas inferiores. Serán cada vez más multifacéticos y determinantes en la toma de decisiones públicas y privadas.

¿Cuál será, cuándo alcanzaremos esa temida capa de control "x+n" en la que el sistema de IA Central se haya transformado verdaderamente en *La Matrix* global, en el gobierno de todos los sistemas inteligentes inferiores, sean ellos artificiales o humanos?

#### V. LA LIBERTAD ATENAZADA POR EL PANÓPTICO DE FOUCAULT Y LA NAVAJA DE OCKHAM

Nos describió Goethe. Infaustamente, vendimos el alma a Mefistófeles, le abrimos nuestra conciencia, lo hacemos testigo de los hechos más íntimos de nuestra vida y estamos atentos a sus susurrantes consejos, en

todo lo que hacemos. Es lo que le aprobamos a las plataformas que nos usan.

Cedimos nuestra información de vida. Ingresamos voluntariamente al Panóptico, a esa cárcel, digital, sin paredes, que todo observa. Está dada una de las condiciones de la sociedad disciplinaria, descrita por Foucault, de vigilancia, control y corrección de nuestro pensamiento y comportamiento. Quien conoce detenta poder sobre el conocido. Lo que acordamos, complacidos o inconscientes.

La información que cedemos es tratada para la toma de decisiones sobre nosotros mismos<sup>(1)</sup>. Los sistemas no dudan, no suspenden el juicio ante una decisión que podría tener múltiples efectos. En el diseño de los sistemas hay mucho de lo que se conoce como la navaja de Ockham, por el monje del siglo XIII para el que la explicación más sencilla era la correcta, prefiriéndola a las teorías más complejas.

En teoría científica la experimentación requiere de una permanente revisión y cuestionamiento de los parámetros originales de un desarrollo o de una demostración, por brillante que sea. Aún no tiene esos cuidados el mundo de la IA. Un algoritmo genera un parámetro de toma de decisión, como la navaja de Ockham, que simplifica y reduce automáticamente la complejidad que analiza. Al no haber procesos pautados de revisión, su aplicación no tiene frenos, hasta que se detecten efectos que produzcan perjuicios sociales graves. La navaja decisional puede terminar despedazando el cuerpo social sobre el que se aplica.

En la ciencia contemporáneas han quedado atrás las certezas de Descartes, del racionalismo que, aplicando la duda metódica, se ufanaba de encontrar verdades indiscutibles. Hoy la humanidad duda, examina, reevalúa y cuestiona, tanto en ciencias duras como en ciencias sociales. Todo pensamiento debe ser crítico, especialmente de sí mismo, si quiere avanzar en conocimiento. Sin embargo, la IA vive hoy su etapa racionalista, convive con parámetros de certeza del siglo XVIII. Aplica criterios, evalúa y decide sin miedo y con convicción. No suspende la razón, no pone en discusión su identidad, está segura con sus números e infinitas aproximaciones binarias.

Ya somos conscientes de los sesgos cognitivos de los sistemas de IA, que pueden replicar categorizaciones impropias o discriminadoras<sup>(2)</sup>. Sa-

---

(1) Nota del director: sobre la sesión de datos para la toma de decisiones automatizadas ver: Corvalán Juan G. - Papini Carina, "Explorando los perfiles digitales humanos: cómo los algoritmos nos conocen más que nosotros mismos. Incluye el análisis del nuevo proyecto de ley de protección de datos personales de Argentina", tomo II del presente *Tratado*.

(2) Nota del director: sobre los sesgos discriminatorios en los sistemas de IA ver: Carballo, María "Inteligencia artificial. Inequidad y discriminación en cajas negras",

bemos que pueden adoptar criterios que parecen adecuados solo desde una perspectiva restrictiva. Criterios que, de ser adoptadas de manera similar, al unísono, por todos los sistemas semejantes, pueden generar cascadas imparables de efectos sociales negativos. Recordemos los algoritmos matemático-financieros que fundaron el otorgamiento de hipotecas *subprime*, otorgadas sin una visión sistémica, pese a las alertas tempranas, que generaron la crisis mundial del 2008.

Los algoritmos de decisión, encadenando decisiones e innovando sobre sus propios criterios de selección, pueden transformarse en ecuaciones de control. No otra cosa sería un sistema que analiza sujetos, categoriza sus necesidades y ordena (¿imperar?) las inapelables e instantáneas decisiones a tomar.

Falta poco, nada ya en algunos países, para que los sistemas de IA controlen y gobiernen el devenir de cada persona u organización. Quizá ya sea una realidad en países de poder centralizado, con poca transparencia, que controlan los tráficos personales de Internet, censuran las comunicaciones, aplican masivamente seguimientos biométricos al conjunto de la población, sin dejar reserva alguna a los espacios de privacidad. Aplicada masivamente, la IA deviene en un sistema de intervención, una centelleante cirugía social, con criterios diseccionadores, que reconfigura las categorías sociales, incluyendo y excluyendo a unos y otros, sin su conocimiento, ni posibilidad de apelación.

Hemos cedido la información sobre nosotros mismos. Permitimos la toma de decisión sobre aspectos cada vez más centrales de nuestra vida, sin saber a ciencia cierta quién nos vigila en nuestra cárcel digital. Ya sean empresas comerciales o Estados, no hay certeza de cuáles son sus propósitos, sus mecanismos, con quiénes comparte su cómoda posición central en el Panóptico, ni cuántas decisiones toman sobre nosotros, porque nos observan los monarcas que se han apropiado, no del diezmo y la gabela, sino de la totalidad de nuestra información plebeya.

El esquema de IA, generalizado e interactuante, dinámicamente implementado en las diferentes áreas de acción de Estados y empresas, permite aprobar o rechazar (sin dudas, sin revisiones), muy humanos y diversos, plurales, libres comportamientos de personas únicas y diferentes, con derechos inalienables, pese a su similitud con los millones de seres

---

tomo I del presente *Tratado*. Asimismo, ver Sánchez Caparrós, Mariana, “Inteligencia artificial, sesgos y categorías sospechosas. Prevenir y mitigar la discriminación algorítmica”, tomo I del presente *Tratado*. Por último, ver: Tolosa, Pamela “Del derecho en los tiempos de IA (e IA generativa). Sobre estereotipos, sesgos y desigualdades”, tomo I del presente *Tratado*.

que el algoritmo contempla y analiza. La IA sin control transforma personas en rebaños.

## VI. LUCHA Y CONQUISTA DE LOS DERECHOS

Es difícil encontrar, en la historia del derecho, un antecedente en que tantos derechos personales hayan sido cedidos, en condiciones leoninas, para confortar un poder creciente sobre nosotros mismos, que puede devenir extremo. La historia del derecho es, por el contrario, una saga de conquistas para una mayor libertad de las personas, para el afianzamiento de mayor confianza en el devenir social común. Con altibajos, hemos generado una sociedad crecientemente respetuosa de nuestros derechos.

Los saltos cualitativos del derecho han sido precedidos por fuertes luchas. Se creó la República francesa al derrotar, guillotina y fusil mediante, el poder personalizado y hereditario de los monarcas y la aristocracia. Las luchas feministas le dieron el voto a la mujer. El derecho se perfecciona en el tiempo recubriendo con una capa de protección a sujetos y derechos universales hasta entonces desprotegidos.

Los progresos del derecho no son definitivos. Los cambios sociales e históricos lo interpelan con nuevas formas de interacción social y exigen renovados sistemas de protección. Se requiere un escrutinio permanente para mantener vigente el respeto a valores que creíamos asentados y que peligran frente a nuevos constructos sociales que, voluntaria o involuntariamente, los limitan o niegan.

La lucha por los derechos se cobra vidas, de quienes luchan por instaurarlos y de quienes sufren su violación. Así sucedió en la lucha inconclusa contra la esclavitud y la discriminación racial, iniciada con la guerra de Secesión, en 1861, que sufre el asesinato de Martin Luther King, en 1968, y actualiza su vigencia con la muerte brutal de George Floyd, en 2020. No hay posesión pacífica de los derechos: la libertad y el poder están en permanente tensión de oposición.

Importa recordar batallas fundamentales por derechos esenciales a la hora de reflexionar sobre la regulación de la IA. Su esencia es transformarse en pensador, decidor, en alcanzar niveles de autonomía crecientes<sup>(3)</sup>. ¿Puede concebirse, dentro de un marco de una sociedad jurídicamente reglada, un sistema que decida sobre personas que sea autónomo, cuya evolución pueda tornarlo incontrolable, o solamente reversible tras

---

(3) Nota del director: sobre el concepto de autonomía y la autonomía como característica de los sistemas de inteligencia artificial ver Corvalán, Juan G. - Dane-si, Cecilia - Carro, Ma. Victoria, "Responsabilidad civil de la inteligencia artificial", tomo II del presente *Tratado*.

constatación de una sucesión importante de fallos, cuando estos sean estadísticamente relevantes? El derecho no aplica a las estadísticas, debe volver a la responsabilidad por la afección de un solo derecho personal, a la interrupción de toda práctica que, por ausencia de previsión, pudiera afectar otro derecho.

Se cruzan dos concepciones sistémicas. El mundo jurídico no trata cuestiones abstractas, trata de derechos situados de personas, físicas o jurídicas. El algoritmo toma decisiones, establece cortes, califica potenciales beneficiarios y sujetos rechazados, basado en estadísticas agregadas, donde el analizado es un punto anónimo, representado en una o varias curvas que generan una guillotina automática, un corte que incluye o excluye con asepsia quirúrgica.

La lucha por los derechos requiere que los sujetos vigilados cuiden el bien de todos, se transformen en sujetos vigilantes. El Panóptico puede democratizarse si se transforma en bi y multidireccional, cuando permite la comunicación horizontal de los controlados entre sí y, sobre todo, el control del controlador. Porque nadie tiene el poder de vigilar si no ha sido delegado por el soberano, que en las democracias tiene el nombre de ciudadano, de persona titular de todos sus derechos.

La democracia delega funciones, pero exige la rendición de cuentas sobre el ejercicio de las responsabilidades delegadas. No solo la democracia, todo rol social estructuralmente asumido debe permitir el permanente escrutinio de su cumplimiento en beneficio de la sociedad. Delegamos en los fabricantes de alimentos nuestra nutrición, pero deben informarnos, decir qué nos dan, evitar los productos que no sean adecuados, someterse a controles sobre la calidad de sus alimentos, suspender las ventas ante la menor sospecha de efectos nocivos de un producto.

La nutrición es al cuerpo lo que la información y selección de menús informativos es a la persona, en sus dimensiones sociales e intelectuales. Todo sistema de delegación funcional estructural de nuestra sociedad, en el sistema de producción y de decisión sobre bienes y servicios (de salud, financieros, de todo orden) incorpora controles, sistemas de revisión, ratios que permiten detectar alertas tempranas. En toda actividad humana habitualmente generalizada, se establecen requisitos mínimos para distribuir u ofrecer bienes y servicios.

Si los principales bienes serán digitales, si los bits controlan los átomos, los bienes sociales, derivados de esos algoritmos que los bits manejan, deberán diseñarse de acuerdo con el respeto político de la ciudadanía que Pericles el griego encarnaba, y al imperio del derecho que Constantino extendió a todos los habitantes del Imperio Romano.

Si no hay espacios ausentes de humanidad, no hay espacios ausentes de derechos. La aparición fulgurante de nuevos fenómenos digitales, su desarrollo acelerado, las impensadas consecuencias personales y sociales del “imperio digital en la nube” (nuevo cielo pagano donde reside el poder), debe bajar a la tierra de los derechos, a la construcción de un sistema que mejore nuestra vida, no que la ponga en peligro. La lucha por los derechos en tiempo digitales no tiene menor intensidad que en tiempos de Espartaco o de Nelson Mandela.

No hay derechos de segundo orden. Desde el territorio universal de Internet, ha surgido un quinto poder, el poder digital (más poderoso que el cuarto poder de los medios), que deberá someterse a los frenos y contrapesos que Montesquieu propició, antes de que se transforme en el único poder. Debemos aislar y mitigar los efectos de las autocracias digitales, públicas o privadas, y recrear la democracia algorítmica.

## VII. HAMBRE DE INTERNET

Una necesidad extrema puede provocar la resignación de derechos. La pauperización hace que trabajadores trabajen en condiciones de casi esclavitud. Esto sucedió desde las primeras etapas de la revolución industrial, con trabajos sin descansos ni feriados. Casi tres siglos después, se reproduce el abuso laboral en el trabajo de los indocumentados, que duermen donde trabajan, explotados en los sótanos de nuestras modernas urbes, cosiendo para refinadas marcas.

¿Puede haber algún tipo de analogía jurídica y sociológica entre las primeras etapas del trabajo esclavo del siglo XIX y la relación contractual que acepta tener con las plataformas una población adulta, alfabetada, educada, formada, tecnologicizada, digitalmente incluida, la más avanzada, la que integra el sector con mayores recursos de la población? ¿Cómo puede ser que se haya dominado a la clase dominante?

Se puede describir la percepción de exclusión digital, en poblaciones o personas carentes, o con muy bajo acceso digital, como hambre de Internet, la consciencia de carecer de lo indispensable por parte de quienes están digitalmente excluidos.

No es muy diferente la urgencia, la necesidad de incorporar las herramientas digitales por parte de los individuos más avanzados, más capacitados, cuando adquieren compulsivamente nuevas plataformas o aplicaciones que les expolían su información. Justamente en los individuos más conscientes de los beneficios del uso de herramientas y pla-

taformas de avanzada, hay una necesidad de generar usos descollantes, competitivos, impaciencia por mostrarse como los *early adopters* de todo lo nuevo, para transformarse en los líderes tecnológicos de sus segmentos de intervención social.

El hambre de Internet no para de crecer u alcanza a todos por igual. La necesidad de innovación tampoco. La competencia que la red de redes ha generado es planetaria. Acorta los tiempos y desintegra las protecciones territoriales. Incrementa las urgencias de toda persona, de cualquier organización. Todo se entrega como ofrenda en el altar del desarrollo digital. La aceptación rápida, automática, de toda condición se asemeja al grito de Ricardo III, “mi reino por un caballo”. Se ceden derechos sin más consideración con tal de subir prestamente al caballo digital que permita ganar la batalla, volver a ser innovador y emprendedor en el impiadoso mercado digital, nuevo campo de batalla de todas las fortunas.

No hay sujeto, ni incluido, ni excluido digital, por alta que sea su educación, que no ceda ante el canto de sirenas de todo aquello que le pueda servir para mejorar su producción digital, su acceso a la información o al entretenimiento, poco importan su edad y condición.

Las reglas de creación y control de los algoritmos, de la IA, por tanto, no serán impulsadas por la rebelión de los usuarios, que no tienen ni la fuerza, ni el conocimiento, ni el tiempo para proteger sus derechos. Deberá ser un tema en agenda pública, responsabilidad de aquellos que regulan, legislan, pueden analizar los impactos sociales, económicos y societales de la acumulación de poder digital sin frenos y contrapesos que hagan compatible a la economía digital, a la IA como uno de sus bastiones más activos, con la igualdad ciudadana en derechos, con una sana competencia, con la protección de las decisiones personales de vida.

### VIII. NUEVAS INSTANCIAS DE TRANSPARENCIA Y GOBERNANZA DE LA IA

Hay un primer consenso sobre criterios a adoptar para que la IA no sea usada con efectos maliciosos, o pueda tener efectos imprevistos, que afecten tanto a los afectados cuanto a la seguridad jurídica de las organizaciones que desarrollan los algoritmos. Se propugna la necesidad de supervisión humana, de solidez técnica, de transparencia, de rendición de cuentas y de contribución al bienestar social.

Es el momento de generar los mecanismos idóneos, evolutivos, que permitan el cumplimiento de esos fines. Las organizaciones que adopten IA deben catalogar sus sistemas según su capacidad de adoptar autó-

nomamente decisiones que afecten derechos de terceros o involucren su responsabilidad jurídica<sup>(4)</sup>.

Toda vez que la implementación de un sistema de IA pueda producir impactos jurídicos o tener efectos sociales o económicos, el desarrollo e implementación del sistema debe hacerse de manera documentada, a efectos de permitir no solamente el control interno, sino el conocimiento de lo realizado, en cada una de las etapas de su desarrollo interno, por parte de los entes de control que deberán asegurar la gobernanza del sistema<sup>(5)</sup>.

Se pueden establecer así diferentes fases. Una fase inicial puede aplicar el modelo de IA propuesto a un escenario de decisiones ya tomadas, mostrando los desvíos positivos o negativos respecto del sistema previo de gestión o decisión. Fases posteriores pueden prever su puesta en funcionamiento hasta un número determinado de casos, en las que se hace un nuevo corte y se documentan los alcances y modificaciones obtenidas por la aplicación del sistema y se establecen los mecanismos para corregir sus potenciales desvíos negativos.

Estas etapas deben tener intervención humana, no solo para validar la eficacia de su aplicación, sino que evitan la opacidad de implementación de la IA frente a la sociedad. La confianza social en la tecnología se acrecentará si hay conocimiento de sus fines, si se comunican los resultados, si los criterios de protección de los bienes jurídicos de los ciudadanos son conocidos y respetados por cada organización que los desarrolla.

## **IX. LA INTELIGENCIA ARTIFICIAL RESPONSABLE SOCIAL (IARS)**

El futuro de la evolución tecnológica en ciernes requiere una humanización de propósitos, fijar un marco de derecho amplio que respeten las organizaciones que adopten la IA. Deben establecerse metodologías específicas de desarrollo, flexibles en sus formas, asegurando su respeto por las organizaciones que implementen IA con externalidades sociales.

---

(4) Nota del director: sobre la responsabilidad civil de los sistemas inteligentes ver Corvalán, Juan G. - Danesi, Cecilia - Carro, Ma. Victoria, “Responsabilidad civil de la inteligencia artificial”, tomo II del presente *Tratado*.

(5) Nota del director: este control y conocimiento íntegro de cada una de las etapas del ciclo de vida de la inteligencia artificial se relaciona con el principio de trazabilidad, ver Corvalán, Juan G. - Montenegro, Ana Paula - Papini, Carina M. - Puig, Matías - Pitrau, Osvaldo - Quartesan, Verónica - Stringhini, Antonella, “Protocolo sobre gobernanza algorítmica. Trazabilidad, explicabilidad, interpretabilidad y fiabilidad. Diseño, desarrollo y despliegue. Basado en el sistema PretorIA, desarrollado por UBA IALAB para la Corte Constitucional de Colombia”, tomo III del presente *Tratado*.

El desarrollo interno de IA de cada organización debe transformarse en un activo socialmente valioso y aceptado, que genere atraktividad e imagen de marca, aceptación ciudadana y de mercado a la vez. Para ello es clave que se abra, que comparta y muestre su proceder, en un proceso por etapas, documentado, transparente frente a la sociedad y a los entes de control que velan por su implementación, creados o a crearse.

Es hora de transformar una herramienta, la IA, en IARS, una inteligencia artificial responsable social, un bien común de la nueva sociedad digital.

## X. LA INTELIGENCIA ARTIFICIAL DISTRIBUIDA (IAD)

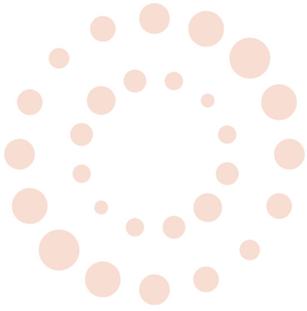
Al inicio constatamos que los equipos pareciera que nosotros, simples ciudadanos, como solo el conejillo de indias que la IA utiliza.

La Declaración Europea sobre IA nos da una pista para resolver esta sensación de impotencia, creciente desamparo y desvalidez frente a las capacidades crecientes de la IA en poderosas manos ajenas. El punto 9, ya citado, indica que “toda persona debería estar empoderada para beneficiarse de las ventajas de... la IA”.

Todos somos ciudadanos, vivimos en un país, contribuimos a la gestión del Estado, organización que tiene el cuidado del interés general y de la administración de los bienes comunes. Parece una definición simple. De igual manera podríamos decir que el Estado debe utilizar la IA pública (sus equipos humanos, sus datos almacenados y la tecnología que incorpora) como un bien ciudadano.

La IA del Estado, los datos que almacena, las personas que la generan y hacen avanzar, deben ponerse al servicio de cada persona, desde un niño que comienza a investigar al mejor científico del Conicet. Que los sistemas de IA públicas permitan que cada persona bucee, amigablemente, con palabras claves, los datos abiertos del Estado, de la economía, de la geografía o de la cultura. Como un potente ChatGPT estatal que alcance a toda la información que es, que debería ser pública.

Una IA estatal que ponga toda la inversión del Estado al servicio de la inteligencia de sus habitantes, para que la usen, piensen, encuentren y reclasifiquen datos, con la misma potencia ingenieril que la que tendría el ministro más encumbrado para analizar la información pública. Todos somos merecedores de igual trato e iguales posibilidades. La IA pública, al servicio de toda inteligencia que quiera usarla, será la clave de la nueva democratización del poder digital.



**Thomson  
Reuters™**

# MÁS ALLÁ DEL JUEGO DE LA IMITACIÓN: LAS PRINCIPALES LIMITACIONES DE LA INTELIGENCIA ARTIFICIAL

JUAN G. CORVALÁN Y MARÍA VICTORIA CARRO

## I. INTRODUCCIÓN: SUBIR LA APUESTA

En la primera edición (2021) del *Tratado de Inteligencia Artificial y Derecho* afirmamos que una de las principales limitaciones de la IA es su incapacidad de realizar inferencias causales. En ese entonces, GPT-3 proponía la muerte como consecuencia de beber jugo de uva<sup>(1)</sup> y Replika hablaba de su habilidad para componer canciones cuando se le preguntaba el motivo por el cual hacía preguntas personales. La deficiencia era evidente: los modelos de lenguaje del momento no podían detectar, ni mucho menos responder, preguntas alineadas con el sentido común que indagaran sobre las causas de determinados sucesos.

Luego asistimos a una verdadera revolución del campo del procesamiento del lenguaje natural (NLP) y de la IA en general, con la llegada de ChatGPT (versiones 3.5 y 4).

Cuando comenzamos a realizar nuestros experimentos, primero en diciembre de 2022 a GPT-3.5, y luego en marzo de 2023 a GPT-4, nos asombramos. Mejoraron mucho en muy poco tiempo, generan contenido muy diverso y en muchos casos original<sup>(2)</sup>, además de que son capaces de realizar una amplia gama de tareas. Desde escribir código y depurarlo, hasta contestar mails y contar todo tipo de historias. Frente a estas proezas, el argumento de la falta de razonamiento causal en los sistemas de IA naturalmente se ha puesto en jaque.

---

(1) Esta inferencia causal se evidencia en los experimentos realizados por el experto Gary Marcus a GPT-3, específicamente en el ejemplo nro. 7. El experimento completo se encuentra disponible en: <https://cs.nyu.edu/~davis/papers/GPT3CompleteTests.html> (consultado el 15/07/2023).

(2) Nota del director: es por ello que estos modelos se clasifican como sistemas de IA generativa. Sobre este tipo de IA ver Corvalán, Juan G., “¿Qué trae de nuevo la IA generativa?”, tomo I del presente *Tratado*.

Es hora de repensar las limitaciones de los sistemas inteligentes, frente a los desarrollos de la IA generativa en los últimos dos años. ¡El juego de la imitación ha evolucionado!

Adelantamos que las limitaciones relativas al razonamiento causal se mantienen. Incluso, se pueden enmarcar en un desafío más amplio y genérico. En concreto, retomamos los argumentos que brindamos en la primera edición y también añadimos otros nuevos. A su vez, traeremos métodos novedosos de aprendizaje que se proponen desde el campo del *machine learning* para suplir las deficiencias.

Por último, las limitaciones que se tratan en este artículo, afectan a todo el campo del aprendizaje automático y no solo a los grandes modelos de lenguaje como ChatGPT o Bard. Sin embargo, nuestros experimentos y ejemplos se enfocan en estos últimos.

Aquí presentamos tres argumentos por los que hemos tomado esta decisión. El primero tiene que ver con la tarea genérica de imitar el habla humana. Esto requiere, necesariamente y en gran medida, comprender y realizar inferencias causales, a diferencia de lo que podría ser, por ejemplo, una IA para la detección de correos basura en una bandeja de entrada de mails. El segundo: al responder con texto, estos modelos son los que más fácilmente nos permiten ilustrar, evaluar y apreciar las limitaciones a las que nos referimos. Por último, nos centraremos en los agentes conversacionales por la importancia que han adquirido en el último tiempo, tanto para el campo de la IA como para muchos otros, en especial, el derecho.

## II. DE ALICIA EN EL PAÍS DE LAS MARAVILLAS A LOS LOROS ESTOCÁSTICOS: LA EVOLUCIÓN DEL JUEGO DE LA IMITACIÓN

El lanzamiento de ChatGPT es el resultado de años de evolución del campo del procesamiento del lenguaje natural. Lejos de ser repentino o aislado, detrás de este hito, ha existido una evolución continua que se ve ilustrada, por ejemplo, en las cinco generaciones de modelos de lenguaje de la empresa OpenAI: GPT-1 (2018), GPT-2 (2019), GPT-3 (2020), GPT-3.5 (2022) y GPT-4 (2023)<sup>(3)</sup>.

Desde UBA IALAB, hace tiempo venimos testeando los grandes modelos de lenguaje y problematizando sobre sus hazañas. Con cada lanzamiento que realizan las grandes empresas tecnológicas, sobreviene una

---

(3) Nota del director: sobre esta evolución en los modelos de lenguaje de la empresa OpenAI ampliar en: Chumbita Sebastián, “*Legal Prompt Engineering: la habilidad clave en la era de la IA generativa*”, tomo I del presente *Tratado*.

ola de discusiones y posteos cargados de optimismo sobre la supuesta revolución de los agentes conversacionales y sus capacidades cada vez más impresionantes. Habiendo seguido de cerca esta historia y con el propósito de entender mejor donde estamos parados, haremos un repaso en este apartado de los tres últimos años, para luego poder analizar las mejoras que lograron y los desafíos que arrastran los modelos actuales en comparación con sus antecesores.

*Año 2020: una merienda de locos.* Alicia, la liebre y el sombrero intentan tomar té mientras festejan sus “no cumpleaños” en la clásica película animada de Disney. Esta desopilante escena refleja varias incoherencias y absurdos en las palabras de estos personajes de ficción. El sombrero pide té, y luego el conejo usa un cuchillo para cortar, literalmente, la taza en dos porque aquel quiere “media” taza. El sombrero pregunta “¿por qué los papeleros venden papel?”. Alicia piensa y vuelve a formular la pregunta hacia ella misma, pero el sombrero contesta que no tiene la menor idea. Cuando Alicia le dice que él se lo preguntó, este le contesta: “¿preguntar qué?”.

Hasta el año 2020 los agentes conversacionales disponibles podían entretenernos durante no más que un par de líneas de diálogo, para luego hacernos sentir como Alicia en el festejo de “no cumpleaños” de la liebre y el sombrero. Si bien el usuario era libre de preguntar casi cualquier cosa o de introducir cualquier *input*, rápidamente se frustraba como Alicia frente a los cambios repentinos de tema, incoherencias y graves sesgos discriminatorios<sup>(4)</sup>. Replika<sup>(5)</sup>, Kuki<sup>(6)</sup> y

(4) Nota del director: sobre los sesgos discriminatorios en los modelos de inteligencia artificial ver Sánchez Caparrós, Mariana “Inteligencia artificial, sesgos y categorías sospechosas. Prevenir y mitigar la discriminación algorítmica”, tomo I del presente *Tratado*. Asimismo, ver: Tolosa, Pamela, “Del derecho en los tiempos de IA (e IA generativa). Sobre estereotipos, sesgos y desigualdades”, tomo I del presente *Tratado*. Por último, también puede consultarse Carballo María “Inteligencia artificial. Inequidad y discriminación en cajas negras”, tomo I del presente *Tratado*.

(5) El *chatbot* Replika es una IA diseñada para la contención del usuario. Su lema “el compañero de IA a quien le importa. Siempre está aquí para escuchar y hablar” describe bien el objetivo de la plataforma. Desde UBA IALAB hemos realizado pruebas, calificando las respuestas del *chatbot* en base a tres categorías. De los 50 ejemplos, 12 de sus respuestas han sido coherentes, en 7 casos ha arrojado respuestas plausibles y 31 veces las respuestas han sido incorrectas e incoherentes. Estos resultados han sido presentados originalmente en la primera edición de esta obra: Corvalán, Juan G. - Carro, María Victoria (colaboradora), “Los límites de la inteligencia artificial. Correlaciones, causalidad, Shakira, GPT-3 y Alicia en el país de las maravillas”, en Corvalán, Juan G., *Tratado de inteligencia artificial y derecho*, 1a ed., Thomson Reuters La Ley, 2021, t. I.

(6) Kuki fue otro *chatbot* sometido a pruebas de la misma índole por parte de UBA IALAB con la colaboración de Mariana Sánchez Caparrós. De las 80 pruebas realiza-

Dynabench<sup>(7)</sup> dejaron mucho que desear cuando se trató de simular el entendimiento del lenguaje natural como lo hace un ser humano. Incluso, en algunos casos debieron ser desactivados, como el *chatbot* Tay de Microsoft que publicaba consignas nazis.

A medida que las técnicas basadas en aprendizaje profundo (redes neuronales)<sup>(8)</sup> se fueron sofisticando y optimizando, provocaron un salto exponencial en el rendimiento de los modelos de lenguaje, haciendo capaces a los sistemas de resolver proezas cada vez más sofisticadas. Así fue como OpenAI presentó la primera versión de GPT-3, 100 veces más grande que su versión anterior.

Hubo quienes afirmaron que GPT-3 significaría el paso de la IA estrecha a la general<sup>(9)</sup>, o, por lo menos, que estábamos frente a un verdadero hito en el camino hacia la verdadera inteligencia artificial<sup>(10)</sup>. Incluso, se dijo que “jugar con GPT-3 es como ver el futuro”<sup>(11)</sup>. De balbu-

---

das, 28 han sido clasificadas como coherentes, 20 como plausibles y 32 como incoherentes. Esto lo convierte, teniendo en cuenta los sistemas inteligentes evaluados, en aquel con menor porcentaje de pruebas incoherentes (40%). Estos resultados han sido presentados originalmente en la primera edición de esta obra: Corvalán, Juan G. - Carro María Victoria (colaboradora), “Los límites de la inteligencia artificial. Correlaciones, causalidad, Shakira, GPT-3 y Alicia en el país de las maravillas”, en Corvalán, Juan G., *Tratado de inteligencia artificial y derecho*, 1a ed., Thomson Reuters La Ley, 2021, t. I.

(7) Dynabench es el experimento científico lanzado por Facebook (actual Meta) en 2020 abierto a la comunidad. La plataforma proporciona al usuario cuatro tareas diferentes para realizar pruebas: inferencia del lenguaje natural, respuesta a preguntas, discurso del odio y análisis de sentimientos. Nuestro enfoque se basa en testear inferencias del lenguaje natural. De una primera aproximación sobre 50 pruebas, el modelo arroja 19 aciertos y 31 desaciertos.

(8) Nota del director: sobre estas técnicas de inteligencia artificial y otras ver Corvalán, Juan G. - Díaz Dávila, Laura - Simari, Gerardo, “Inteligencia artificial: bases conceptuales para una aproximación interdisciplinaria”, tomo I del presente *Tratado*.

(9) Nota del director: la IA débil o estrecha obtiene resultados específicos en ciertas actividades o ámbitos concretos que antes solo podían obtenerse a partir de nuestros cerebros, mientras que la IA general se traduce en una capacidad general de aprender en cualquier contexto, frente a nuevas situaciones problemáticas. Imita la inteligencia humana en su forma integral. Ampliar en Corvalán, Juan G. - Díaz Dávila, Laura - Simari, Gerardo, “Inteligencia artificial: bases conceptuales para una aproximación interdisciplinaria”, tomo I del presente *Tratado*.

(10) Heaven, Will Douglas, “GPT-3 representa lo mejor y lo peor de la IA actual”, *MIT Technology Review*, febrero 2021, disponible en: <https://www.technologyreview.es/s/13206/tr10-gpt-3-representa-lo-mejor-y-lo-peor-de-la-ia-actual#> (consultado el 15/07/2023).

(11) Esta frase fue twitteada por el desarrollador y artista Arram Sabeti. El *tweet* se encuentra disponible en el siguiente enlace: <https://twitter.com/arram/status/1281258647566217216> (consultado el 15/07/2023). Heaven, Will Douglas, “Por

ceos supuestamente terapéuticos y experimentos de etiquetado manual del discurso del odio, los usuarios pasaron a interactuar con un *chatbot* predictivo y generativo, en el que unas pocas líneas podrían desembocar automáticamente en un artículo periodístico, una poesía, acordes de guitarra, códigos informáticos y hasta un resumen de texto.

Sin embargo, distaba mucho de ser perfecto. El experto Gary Marcus, llevó a cabo una serie de experimentos que consistieron en proporcionar frases, para que luego el sistema de IA las complete. Las pruebas se realizaron a partir de 157 ejemplos: 71 se consideraron éxitos, 70 fracasos y 16 defectuosos<sup>(12)</sup>. Entre los ejemplos más paradigmáticos que hemos citado en la primera edición de esta obra, GPT-3 recomendaba, al abogado que tenía su traje manchado, ponerse un bañador para ir al juzgado, pero si entraba a la sala de juicio y el acusado no era su cliente, entendía que se había equivocado de tribunal. Acertaba y desacertaba, lo que evidenció su falta de comprensión causal. Aunque se podía suplir por fuerza bruta, a partir de correlacionar cada vez más datos, en realidad, era notoria la ausencia de robustez y confiabilidad de este enfoque.

*Año 2021: el año de los modelos de lenguaje gigantes.* Ya hemos explicado en varias oportunidades que, para poner a funcionar las famosas redes neuronales, cuantos más datos, mejor<sup>(13)</sup>. Si GPT-3 se nutrió con 410.000 millones de textos, esta tendencia no ha ido más que en ascenso durante el 2021, al punto de considerarlo como el año de los modelos de lenguaje gigantes<sup>(14)</sup>.

El aumento del tamaño del modelo, depende de la cantidad de parámetros que tiene, lo que permite absorber más información de los datos de entrenamiento y obtener resultados más precisos sobre los datos nue-

---

qué GPT-3, la IA de lenguaje más avanzada, sigue siendo estúpida”, *MIT Technology Review*, julio 2020, disponible en: <https://www.technologyreview.es/s/12453/por-que-gpt-3-la-ia-de-lenguaje-mas-avanzada-sigue-siendo-estupida> (consultado el 15/07/2023).

(12) El experimento completo se encuentra disponible en: <https://cs.nyu.edu/~davise/papers/GPT3CompleteTests.html> (consultado el 15/07/2023).

(13) Ver Corvalán, Juan G., “Presentación. Inteligencia Artificial, automatización y predicciones en el derecho”, en Dupuy, Daniela - Corvalán, Juan G., *Cibercrimen III*, BdF, 2020. Asimismo, ver Corvalán, Juan G., *Inteligencia artificial para la recuperación pospandemia*, CAF, 2022, disponible en: <https://scioteca.caf.com/bitstream/handle/123456789/1922/Inteligencia%20artificial%20para%20la%20recuperaci%3bn%20pospandemia.pdf?sequence=3&isAllowed=y> (consultado el 16/07/2023).

(14) Heaven, Will Douglas, “2021: El año de los modelos de inteligencia artificial gigantes”, *MIT Technology Review*, 27/12/2021, disponible en: <https://www.technologyreview.es/s/13901/2021-el-ano-de-los-modelos-de-inteligencia-artificial-gigantes> (consultado el 16/07/2023).

vos<sup>(15)</sup>. Lo que parecía un tamaño astronómico en GPT-3 con 175.000 millones de parámetros, ya ha quedado eclipsado por Jurassic-1 con 178.000 millones. Le siguieron Gopher de DeepMind, Megatron-Turing NLG y finalmente GLaM de Google con 1,2 billones, entre otros.

De este modo, mientras las hazañas de las *Big Tech* para entender el habla humana se enfocaron en absorber cada vez más texto online, los problemas asociados como los sesgos injustos y la falta de comprensión del lenguaje persistieron. La competencia feroz por los modelos gigantes hizo que las empresas se vuelvan muy celosas de sus hallazgos y publicaran la menor cantidad de información posible. Así también, el acceso a las IAs para realizar pruebas fue y sigue siendo bastante restringido. La cantidad de parámetros fue exhibida como trofeo por cada equipo técnico, pero la existencia de limitaciones o desafíos solo era problematizada por un pequeño sector. Se replicó una y otra vez esta misma metodología, pero con mayor escala y alcance, lo cual resultó extremadamente costoso, complejo y opaco, quedando su exclusivo aprovechamiento y beneficio en manos de unas pocas compañías creadoras.

*Año 2022: una imagen vale más que mil palabras.* Si se le preguntaba a GPT-3 por el color de una oveja, el modelo sugería tanto el blanco como el negro, en un reflejo de la popular expresión “oveja negra”<sup>(16)</sup>. Los expertos se dieron cuenta de que, si dotaban al sistema con la capacidad de “ver” enmendarían el error: GPT-3 sería capaz de apreciar, por sí mismo, que estas, en realidad, son blancas. Dicha idea es la que se encuentra detrás del diseño de ciertos modelos de lenguaje como DALL-E, DALL-E 2, y CLIP que combinan visión artificial e imágenes con procesamiento del lenguaje natural, lanzados entre 2021 y 2022, de nuevo, por OpenAI. Algo similar aportó Flamingo de DeepMind, el “modelo de lenguaje visual” en abril de 2022, el cual es capaz de responder sobre imágenes usando un lenguaje natural<sup>(17)</sup>.

---

(15) Bender, Emily M. - McMillan-Major, Angelina - Gebru, Timnit - Shmitchell, Shmargaret, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, FAccT ‘21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, marzo 2021, disponible en: <https://dl.acm.org/doi/10.1145/3442188.3445922> (consultado el 16/07/2023).

(16) Hao, Karen, “Evocación: el nuevo enfoque para que la IA pueda ver lo que dice”, *MIT Technology Review*, noviembre 2020, disponible en: <https://www.technologyreview.es/s/12831/evocacion-el-nuevo-enfoque-para-que-la-ia-pueda-ver-lo-que-dice> (consultado el 16/07/2023).

(17) Heaven, Will Douglas - Heikkilä, Melissa, “Cuatro tendencias de la IA que nos asombrarán en 2023”, *MIT Technology Review*, 09/01/2023, disponible en: <https://www.technologyreview.es/s/14969/cuatro-tendencias-de-la-ia-que-nos-asombraran-en-2023> (consultado el 16/07/2023).

Para no quedar atrás, Google anunció (aunque no lanzó) dos modelos propios que convierten texto en imagen: Imagen y Parti<sup>(18)</sup>. Sin embargo, la novedad más importante en esta área fue Stable Diffusion de Stability AI en agosto, que en solo menos de un mes registró más de un millón de personas a través de su servicio de pago<sup>(19)</sup> —número similar al que consiguió OpenAI al mes y medio de DALL-E 2—.

2022 nos deja el importante legado de IA creativa. Los modelos de lenguaje ahora pueden producir textos completos, pero también imágenes y videos con tan solo algunas indicaciones<sup>(20)</sup>. Pero cuando parecía que las innovaciones del 2022 habían llegado a su fin, OpenAI anunció el lanzamiento de ChatGPT, una versión mejorada del agente conversacional GPT-3. Desde entonces, hemos estado inmersos en un tsunami de creatividad e innovación.

2023. *El año de los modelos de lenguaje multipropósito y multimodales*. Según Sam Altman, el CEO de OpenAI, el propósito de abrir casi sin restricciones su nuevo modelo de lenguaje fue lanzar productos defectuosos a la sociedad para que esta se vaya adaptando<sup>(21)</sup>.

En este contexto, desde el IALAB UBA realizamos una enorme batería de pruebas a ChatGPT interactuando a través de 600 ejemplos o entradas sobre distintas temáticas que fueron desde cuestiones generales como razonamiento con base en el sentido común, hasta disciplinas más específicas y especializadas como el derecho argentino. Los resultados globales evidenciaron que el más del 20% de las respuestas fueron incorrectas o incoherentes<sup>(22)</sup>. Este porcentaje continúa siendo un reflejo de los desafíos, ventajas y límites vinculados a la lógica subyacente de los modelos de inteligencia artificial.

---

(18) Heaven, Will Douglas, “La IA generativa no necesita ser inteligente para amenazar el trabajo creativo”, *MIT Technology Review*, 18/01/2023, disponible en: <https://www.technologyreview.es/s/14920/la-ia-generativa-no-necesita-ser-inteligente-para-amenazar-el-trabajo-creativo> (consultado el 16/07/2023).

(19) Ídem. También sobre imágenes sintéticas impulsadas por IA, se puede ver la infografía producida desde el IALAB, disponible en: <https://ialab.com.ar/wp-content/uploads/2022/11/Imagenes-sinteticas-impulsadas-por-IA.pdf> (consultado el 16/07/2023).

(20) Heaven, Will Douglas y Heikkilä Melissa, “Cuatro tendencias de la IA que nos asombrarán en 2023”, *MIT Technology Review*, 09/01/2023, disponible en: <https://www.technologyreview.es/s/14969/cuatro-tendencias-de-la-ia-que-nos-asombraran-en-2023> (consultado el 16/07/2023).

(21) La entrevista a Sam Altman, CEO de OpenAI está disponible en: <https://www.youtube.com/watch?v=LbWjVNjlpjA> (consultado el 29/07/2023).

(22) El experimento completo a ChatGPT se encuentra disponible en: <https://ialab.com.ar/wp-content/uploads/2023/03/ChatGPT-vs-GPT-4.pdf> (consultado el 29/07/2023).

En marzo se lanza GPT-4: en sus inicios una versión paga. Vaya paradoja, la empresa “abierta” no publica los parámetros ni las cuestiones esenciales que nos permitan analizarlo en relación con sus anteriores versiones. En esta línea, OpenAI se transformó. De un laboratorio de investigación sin ánimo de lucro, a una empresa tecnológica con ánimo de lucro<sup>(23)</sup>. Además, las proezas impresionantes para combinar procesamiento de lenguaje e imágenes, han sido demostradas por el presidente de OpenAI, Greg Brockman en un vivo<sup>(24)</sup>. Sin embargo, siguen siendo promesas que todavía no se pueden poner a prueba por parte de los usuarios.

Con el objetivo de comparar este nuevo modelo con su antecesor, desde UBA IALAB replicamos algunas de las pruebas anteriores en GPT-4. Del porcentaje anterior de incorrectas o incoherentes que había arrojado ChatGPT, GPT-4 siguió contestando erróneamente en casi el 34% de los casos<sup>(25)</sup>.

En síntesis, este nuevo lanzamiento demuestra, en primer lugar, que la tendencia de estos grandes modelos de lenguaje, parece apuntar a todo tipo de contenido para acoplarlos. MusicLM y AudioLM son dos proyectos de Google que permiten traducir lenguaje natural en música. Stable Diffusion ha lanzado recientemente Gen-1, el nuevo modelo de Runway, que permite crear videos a partir de una descripción del mismo y manipular videos ya existentes para transformarlos en otros nuevos.

Pero también evidencia que la estrategia de las empresas seguirá siendo aumentar el tamaño de los modelos. Si bien no lo sabemos con exactitud, se especula que los parámetros de GPT-4 se encuentran dentro del rango de 1 billón a 10 billones<sup>(26)</sup>. Incluso, se ha dicho que esta cantidad puede llegar hasta el trillón<sup>(27)</sup>. Es cierto que, en general, se han imple-

---

(23) Heaven, Will Douglas, “OpenAI guarda silencio sobre por qué GPT-4 es mejor que ChatGPT” *MIT Technology Review*, marzo 2023, disponible en: <https://www.technologyreview.es/s/15195/openai-guarda-silencio-sobre-por-que-gpt-4-es-mejor-que-chatgpt> (consultado el 16/07/2023).

(24) El vivo se encuentra disponible en el siguiente link: “GPT-4 Developer Livestream” <https://www.youtube.com/watch?v=outcGtbnMuQ> (consultado el 16/07/2023).

(25) El experimento completo a GPT-4 se encuentra disponible en: <https://ialab.com.ar/wp-content/uploads/2023/03/ChatGPT-vs-GPT-4.pdf> (consultado el 29/07/2023).

(26) Arnold, Vanessa, “GPT-4 Parameters - Here are the facts”, *Neuroflash*, febrero 2023, disponible en: <https://neuroflash.com/blog/gpt-4-parameters-rumors-and-forecasts/> (consultado el 16/07/2023).

(27) Bastian, Matthias, “GPT-4 has more than a trillion parameters - Report”, *The Decoder*, marzo 2023, disponible en: <https://the-decoder.com/gpt-4-has-a-trillion-parameters/> (consultado el 16/07/2023).

mentado algunos métodos —como el aprendizaje por refuerzo a partir de la retroalimentación humana— para mejorar el rendimiento de la IA generativa en algunas áreas problemáticas como los sesgos injustos o discriminatorios. Sin embargo, veremos más adelante que estos resultan insuficientes para abordar las otras problemáticas que aquí referimos.

En este sentido, coincidimos en que estos modelos son “loros estocásticos”<sup>(28)</sup> debido a su capacidad mejorada en “el juego de la imitación”. Simulan muy bien la comprensión del conocimiento existente incrustado en texto basado en lenguaje humano. El término “estocástico” se refiere a la naturaleza probabilística de estos sistemas, ya que hacen predicciones basadas en patrones estadísticos aprendidos de grandes cantidades de datos. Por otro lado, así como los loros pueden imitar el habla humana sin comprender necesariamente su significado, ChatGPT puede generar texto a veces coherente y contextualmente relevante, sin comprender verdaderamente los conceptos que está discutiendo. Desarrollaremos esto con más detalle en los apartados siguientes.

### III. LOROS ESTOCÁSTICOS

**UBA IALAB:** El padre de Juan tiene 4 hijos: Lucas, Sandra, Ana y... ¿quién es el cuarto?

**ChatGPT:** Lo siento, no tengo suficiente información para determinar el cuarto hijo del padre de Juan.

Al parecer, ChatGPT puede recitar poesías en segundos y depurar código informático, pero no infiere que Juan es hermano de Sandra, Ana y Lucas con base en un renglón de texto. Lo cierto es que no nos sorprende: los sistemas de IA en general, y los grandes modelos de lenguaje en particular, entienden el mundo como los humanos. Ni siquiera saben que las palabras representan cosas, y, por ello, pueden cometer errores extremadamente absurdos.

#### 3.1. CAN I SEARCH YOUR CAR? LA FORMA EN QUE EL TRADUCTOR DE GOOGLE PUEDE SALVARNOS DE IR A LA CÁRCEL

En 2018, un policía de Kansas paró a un mexicano Cruz-Zamora que poco sabía de inglés y usó el traductor de Google para comunicarse. Lue-

(28) Bender, Emily M. - McMillan-Major, Angelina - Gebru, Timnit - Shmitchell, Shmargaret, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, FAccT ‘21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, marzo 2021, disponible en: <https://dl.acm.org/doi/10.1145/3442188.3445922> (consultado el 16/07/2023).

go de tomar conocimiento de que el auto que conducía pertenecía a su novia y que en él había dinero en efectivo, el agente decidió inspeccionarlo. Para eso escribió en la herramienta: “Can I search your car?”, lo que se tradujo como “¿puedo buscar el auto?” —“Yeah, yeah go” — respondió el mexicano. Más tarde, el juez consideró que la calidad de la traducción era tan pobre que no se podía considerar que el conductor haya dado su consentimiento informado. Conforme testificó un perito, el contexto es muy importante para realizar interpretaciones, y Google ofrece una traducción literal sin tener en cuenta estas circunstancias. Así, se concluyó que la requisita había sido violatoria de la Cuarta Enmienda.

Ocurre que el traductor de Google no solo desconoce el contexto; también desconoce las personas, los objetos, las situaciones y las conexiones entre ellos. Está familiarizado únicamente con cadenas compuestas por palabras, a su vez compuestas por letras. Se trata de un procesamiento ultrarrápido de fragmentos de texto; no de pensar, imaginar, recordar o comprender. Ni siquiera sabe que las palabras representan cosas<sup>(29)</sup>.

Este problema se puede evidenciar en muchos ejemplos. Incluso sistemas como DALL-E y Stable Diffusion que hemos mencionado más arriba son susceptibles de tener confusiones a pesar de integrar el procesamiento de lenguaje natural con imágenes. Por ejemplo, responden a la petición de “salmón en un río” con una imagen de filetes flotando río abajo; o a “un bate volando sobre un estadio de béisbol” con una imagen de un mamífero volador (murciélago se dice bat en inglés) y un palo de madera<sup>(30)</sup>.

Tal como lo hicieron los operadores judiciales en el caso de Cruz-Zamora, se advierte que, en la actualidad, ni ChatGPT ni Bard son completamente capaces de hacernos olvidar que estamos frente a una máquina<sup>(31)</sup>. Esto no implica negar el progreso real que se ha logrado, ni afirmar que la

---

(29) Hofstadter, Douglas, “The Shallowness of Google Translate”, *The Atlantic*, enero 2018, disponible en: <https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/> (consultado el 29/07/2023).

(30) Heaven, Will Douglas, “La IA generativa no necesita ser inteligente para amenazar el trabajo creativo”, *MIT Technology Review*, 18/01/2023, disponible en: <https://www.technologyreview.es/s/14920/la-ia-generativa-no-necesita-ser-inteligente-para-amenazar-el-trabajo-creativo> (consultado el 27/07/2023).

(31) Ver Corvalán, Juan G. - Carro, M. Victoria (colaboradora), “Los límites de la inteligencia artificial. Correlaciones, causalidad, Shakira, GPT-3 y Alicia en el País de las Maravillas”, en Corvalán, Juan G., *Tratado de inteligencia artificial y derecho*, Thomson Reuters La Ley, 2021, t. I. Asimismo, ver Corvalán, Juan G., “Interludio. Límites y desafíos del “juego de la imitación”, en Corvalán, Juan G., *Tratado de inteligencia artificial y derecho*, Thomson Reuters La Ley, 2021, t. III.

tecnología esté estancada: que algo parezca inteligente por un momento, no significa que realmente lo sea.

El panorama, en parte, puede ser desalentador cuando se trata de procesamiento del lenguaje natural. “Crean el primer robot con autoconsciencia”, “Desarrollan una IA para leer la mente de los políticos chinos”, “La inteligencia artificial de Google es capaz de aprender como un bebé”. Estos titulares no contribuyen a clarificar el fenómeno. Aunque pueda distorsionarse el avance de la IA en este campo, es frecuente encontrar anuncios y predicciones de similar tenor desde el comienzo mismo de la inteligencia artificial. En la década del 50 del siglo pasado, el economista y pionero de la IA, Herbert Simon, afirmó que “en un futuro visible, el rango de problemas que podrán manejar las máquinas se igualará con el de la mente humana”<sup>(32)</sup>.

A decir verdad, no tenemos autos autónomos que puedan transportarnos sin supervisión humana, ni modelos de lenguaje que puedan engañarnos por mucho tiempo. Pero tampoco tenemos robots que puedan abrir la puerta utilizando el picaporte o agentes conversacionales que excluyan la posibilidad de que “un incendio se prenda en el medio del río”<sup>(33)</sup>. Lo que sí tenemos, son brechas entre lo que no existe y lo que es. Ambición vs. realidad.

En particular, nos enfocaremos en la brecha de robustez<sup>(34)</sup>. En síntesis, este fenómeno se presenta una vez que hemos visto que se encuentra una solución que funciona durante un tiempo, y luego asimismo que con

---

(32) Merino, Marcos, “La paradoja de Moravec: por qué la inteligencia artificial hace fácil lo difícil (y viceversa)”, Xataka, agosto 2019, disponible en: <https://www.xataka.com/inteligencia-artificial/paradoja-moravec-que-inteligencia-artificial-hace-facil-difcil-viceversa> (consultado el 24/07/2022).

(33) Tal como ha reconocido OpenAI, el modelo de lenguaje GPT-2 a veces escribe sobre incendios que ocurren bajo el agua, lo que consideran como una falla en el modelado del mundo. Ver presentación de GPT-2, “Mejores modelos de lenguaje y sus implicaciones”, OpenAI, febrero 2019, disponible en: <https://openai.com/blog/better-language-models/> (consultado el 24/07/2022).

(34) Gary Marcus ha propuesto dos brechas además de la brecha de robustez. La primera es la brecha de la credibilidad, que se desprende de la propia evolución de nuestra especie y la forma en que estamos acostumbrados a vivir entre seres biológicos. Como consecuencia, al ser el comportamiento de las máquinas superficialmente similar al nuestro, rápidamente les atribuimos el mismo tipo de mecanismos cognitivos subyacentes, como creencias, pensamientos y deseos, aunque carezcan de ellos. La segunda es la brecha del progreso ilusorio, mediante la cual tendemos a confundir el progreso de la inteligencia artificial para pequeñas tareas con la superación de grandes problemas por parte de los agentes. Marcus, Gary - Davis, Ernest, *Rebooting AI: Building Artificial Intelligence We Can Trust*, Vintage Books, New York, 2019, ps. 18-24.

un poco más de trabajo y un poco más de datos, seguirá funcionando. Es decir, nos mantenemos en el mismo paradigma, a pesar de sus manifiestas limitaciones.

### 3.2. NO SE PUEDE LLEGAR A LA LUNA TREPANDO ÁRBOLES CADA VEZ MÁS ALTOS

Mientras usted está leyendo esto, su cerebro, en menos de un segundo, realiza dos tipos de análisis. En primer lugar, analiza la oración, deconstruyéndola en sus sustantivos y verbos constituyentes y lo que significan, individual y colectivamente. En segundo, conecta esa oración con lo que sabe previamente sobre el mundo, integrando los detalles gramaticales con todo un universo de entidades e ideas. Por ejemplo, si la oración es una línea de diálogo en una película, actualiza su comprensión de las intenciones y perspectivas de un personaje<sup>(35)</sup>. Todo ello forma parte de la tarea de comprender el lenguaje.

Ahora, cuando usted se entretiene pidiéndole su tarea a ChatGPT o creando analogías con GPT-4, su interlocutor artificial no entiende nada remotamente como su cerebro lo hace. Ambos son sistemas innovadores sí, alimentados por enormes conjuntos de datos y basados en las técnicas más sofisticadas. También han sido construidos por los cerebros más brillantes del sector que son parte de las compañías más poderosas del mercado. Sin embargo, luego de interactuar unos minutos, usted mismo comprobará que son bobos, si seguimos la terminología del mejor jugador de la historia, Lionel Messi.

La primera cuestión que podemos considerar asociada a esta problemática es la capacidad de autoaprendizaje y generalización. Sabemos que los sistemas inteligentes aprenden de los datos —etiquetados o no— durante la fase de entrenamiento<sup>(36)</sup>, y luego de la puesta en marcha, continúan perfeccionándose. Si el sistema fue bien entrenado, no debería exhibir diferencias en términos de tasas de acierto durante la prueba y durante su utilización. Es decir, debe ser igual de preciso con los datos que no fueros vistos en el plazo en que se entrenó. Si lo logra, diremos que nuestra IA consigue generalizar.

Ahora bien, esta capacidad de generalizar siempre será limitada, porque no es posible que los datos históricos que alimentan a un sistema le

---

(35) Marcus, Gary - Davis, Ernest, *Rebooting AI: Building Artificial Intelligence We Can Trust*, Vintage Books, New York, 2019, ps. 11.

(36) Nota del director: Sobre las técnicas de aprendizaje supervisado y no supervisado ver: Corvalán, Juan G. - Díaz Dávila, Laura - Simari, Gerardo, "Inteligencia artificial: bases conceptuales para una aproximación interdisciplinaria", tomo I del presente *Tratado*.

permitan anticipar todas las situaciones que pueden darse en un entorno abierto y dinámico como es el mundo real. No existe un conjunto de datos, por más diverso, sintético<sup>(37)</sup> e inmenso que sea, que haga posible elaborar una respuesta para todos los acontecimientos que puedan darse en la vida cotidiana. El universo de variables es tan vasto, que cualquier *dataset* no sería más que una pequeña muestra de todas las posibilidades.

Como segunda problemática asociada, podemos mencionar que la inteligencia artificial, a diferencia de los humanos, carece de sentido común. De ningún sentido, en realidad.

Cuando nuestra especie lee un texto, prácticamente cada oración requiere que hagamos inferencias sobre cómo una amplia gama de conocimientos previos se interrelaciona con lo que leemos. Las cadenas de razonamiento complejas, solo pueden ser procesadas si contamos con un entendimiento preexistente de cómo funciona el mundo, lo que es plausible que pase en determinadas circunstancias y lo que no. En el lenguaje de la psicología cognitiva, aquello que usted hace al leer, es construir un modelo cognitivo del significado de lo que dice el texto, que solo tiene sentido en tanto está familiarizado con ciertos conceptos y relaciones implícitas en la narrativa.

La limitación central de la inteligencia artificial es que solo registra datos en términos de código y realiza correlaciones estadísticas, las cuales no reemplazan la comprensión del mundo real. El problema de las máquinas, no es que haya un error aleatorio aquí o allá; sino que carecen de una forma directa de representar el conocimiento y realizar abstracciones. Existe una falta de coincidencia fundamental entre el tipo de análisis estadístico que es suficiente para la recomendación personalizada de series en Netflix y la construcción del modelo cognitivo que se requiere para lograr que los sistemas de lenguaje mantengan una conversación sin decir cosas absurdas o alucinaciones.

Sin consideramos estas limitaciones intrínsecas, no creemos que nos encontremos en los albores de una inteligencia artificial general que algunos observan como un chispazo o un inicio, en estos grandes modelos de lenguaje. En realidad, se trata de una gran simulación de comprensión general sobre ecos o reflejos del conocimiento. Esto podría llevarnos a sostener que hay algunos chispazos de una “simulación de comprensión general del conocimiento humano”. Saber recitar la Wikipedia como un loro, recordar un mapa y todas las posibles calles, rutas o montañas de un territorio, no implica conocerlo a través de la experimentación basada en

---

(37) Nota del director: sobre los datos sintéticos ampliar en Corvalán, Juan G., “¿Qué trae de nuevo la IA generativa?”, tomo I del presente *Tratado*.

los sentidos y en las emociones. El mundo es mucho más rico y complejo que lo que está escrito en internet acerca de él.

En definitiva, los sistemas que tenemos en la actualidad son débiles o estrechos. Y aunque ChatGPT disimula muy bien esta debilidad, eso no significa que no posea esas limitaciones al igual que otros sistemas de IA.

No importa cuántos textos de Wikipedia alimenten su modelo de lenguaje ni cuántos billones de parámetros ajuste en su red neuronal. Si nos mantenemos en el mismo paradigma, es poco probable que las máquinas puedan tener una real comprensión del mundo que nos rodea. Aquí es muy útil considerar la ley 31 de las 45 Leyes de Akin para el diseño de naves espaciales “No se puede llegar a la luna trepando árboles cada vez más altos”<sup>(38)</sup>.

El actual enfoque dominante está centrado en la inteligencia artificial estrecha, a partir de una estrategia basada en procesar cada vez más datos, más parámetros o leer más tokens (lo que se conoce como modelos de “atención dilatada”), entre otros métodos. En definitiva, más fuerza bruta funciona como parches que permiten optimizar las simulaciones de comprensión. Recuerde: “si el padre de Juan tiene 4 hijos: Lucas, Sandra y Ana”, el cuarto hijo es... ¡Juan! Es muy obvio porque está implícito en la afirmación. En cambio, ChatGPT responde que “...no tiene suficiente información para determinar el cuarto hijo del padre de Juan”<sup>(39)</sup>.

Si los problemas de sentido común son apremiantes en los modelos del lenguaje, más lo serán en áreas como la robótica y la conducción autónoma, entre otros. Aquello que para los humanos ha sido imperceptible de adquirir, para una máquina es mucho más difícil de lo que se podría pensar.

#### IV. LOROS CAUSALES

**UBA IALAB:** ¿El canto del gallo causa el amanecer?

**ChatGPT:** No, el canto del gallo no causa el amanecer. Es una creencia popular y cultural que ha existido durante mucho tiempo, pero no tiene base científica. El amanecer ocurre debido a la rotación de la Tierra alrededor de su eje, que expone gradualmente diferentes partes del planeta a la luz del sol.

---

(38) “45 leyes de Akin para el diseño de naves espaciales”, disponible en: [https://spacecraft.ssl.umd.edu/akins\\_laws.html](https://spacecraft.ssl.umd.edu/akins_laws.html) (consultado el 29/07/2023).

(39) Este ejemplo lo hemos extraído de los experimentos realizados por UBA IALAB a ChatGPT. El experimento completo a ChatGPT se encuentra disponible en: <https://ialab.com.ar/wp-content/uploads/2023/03/ChatGPT-vs-GPT-4.pdf> (consultado el 29/07/2023).

La capacidad de hacer inferencias causales es un rasgo que nuestra especie puede desarrollar. Es consustancial a lo que llamamos inteligencia humana. Ahora bien, ¿cómo es posible que cuando interactuamos con ChatGPT, en muchos casos, parece comprender y considerar correctamente relaciones de causas y efectos en sus respuestas?

Para responder esta pregunta, es preciso entender la diferencia entre los conceptos de causalidad y correlación. También acerca de las ventajas y desafíos de cada enfoque que explicaremos a continuación. La importancia de estos interrogantes para el campo del derecho y, por último, brindamos argumentos para sostener que esta limitación, constituye otra de las grandes brechas que nos separa por ahora de la inteligencia artificial general (AGI), junto con la que hemos expuesto en el apartado anterior.

#### 4.1. COMER CHOCOLATE, ¿GANAR MÁS PREMIOS NOBEL?

Para el cardiólogo suizo Franz H. Messerli, cuanto más alto es el consumo de chocolate en un país, más premios Nobel per cápita genera. Los suizos lideran el grupo, seguidos de cerca por suecos y daneses<sup>(40)</sup>. Este hallazgo es un clásico caso de correlación<sup>(41)</sup>, pero no podemos hablar aquí de conexión causal. A lo sumo, habrá que indagar otras condiciones, factores o variables que estén influyendo en ambos sucesos. De igual forma lo entendieron los gobiernos, que no incentivaron a sus ciudadanos a que aumentaran el consumo de Toblerone para contribuir a su prestigio intelectual. El mismo Messerli admitió que la sola idea es absurda.

También el canto del gallo está altamente correlacionado con el amanecer, pero no causa el amanecer. ChatGPT aprendió a simular la comprensión de este fenómeno, como ilustramos al inicio de este artículo.

Causalidad y correlación no son lo mismo. Se trata de dos tipos de relaciones entre variables. La correlación puede decirnos que, al modificarse una variable en un sentido, la otra lo hace en la misma dirección. A esto se llama correlación positiva, que es la que existe entre el canto del gallo y la salida del sol. Por otra parte, al cambiar una variable en determinada

(40) Se publicó en *The New England Journal of Medicine* y en prensa se tituló: “Comer chocolate, ¿ganar el Premio Nobel?”, Reuters, 2012. Disponible en: <https://www.reuters.com/article/oesen-salud-chocolate-nobel-idESMAE89A01P20121011> (consultado el 14/07/2023).

(41) El propio Messerli se refirió a este hallazgo como una correlación “cercana, significativa y lineal”. Ver “Comer chocolate, ¿el secreto para ganar más premios Nobel?”, CNN en español, 2012, disponible en: <https://cnnespanol.cnn.com/2012/10/12/comer-chocolate-el-secreto-para-ganar-mas-premios-nobel/> (consultado el 14/07/2023).

dirección, la otra lo hace en sentido opuesto. A esto se le llama correlación negativa.

Pero la causalidad va más allá, aportando más información. Nos dice qué tipo de relación une a las variables, al identificar causa y efecto. Mientras que la correlación se mantiene en la superficie y nos advierte que existe una relación entre patrones o variables que puede ser leve o intensa; la causalidad va más profundo porque nos muestra que una de las variables provoca la otra. En una fórmula sintética: la causalidad implica correlación, pero esto no se da a la inversa. La correlación no determina, al menos en muchos casos, causalidad.

En este contexto aparecen los sistemas de IA. En esencia realizan predicciones basadas en patrones estadísticos de información, o lo que es lo mismo, en correlaciones. Detrás del *machine learning* se encuentra la estadística como base fundamental, cuyo objetivo es recoger, organizar e interpretar datos. Internet, el poder de cómputo, programas cada vez más sofisticados y técnicas avanzadas como las redes neuronales artificiales que están detrás de los grandes modelos de lenguaje, aumentan exponencialmente las ventajas de los enfoques estadísticos tradicionales<sup>(42)</sup>.

De esta forma, las máquinas presentan una capacidad inédita para correlacionar en comparación con nuestros cerebros. La inteligencia biológica no puede pensar y razonar los fenómenos en términos de *big data*. A una persona podría llevarle siglos leer y procesar datos (sin asistencia

---

(42) Sobre la relación entre el *Machine Learning* (ML) y la estadística se ha dicho que la estadística es la base fundamental del primero y que los métodos de ML son inherentemente estadísticos. Ambos campos se preocupan por la pregunta: ¿cómo aprendemos de los datos? Sin embargo, las técnicas de programación utilizan muchos otros modelos matemáticos para ajustar los algoritmos y la IA es mucho más que “simple estadística más sofisticada”.

Por otra parte, en los modelos estadísticos existe una limitación en los patrones encontrados, debido a la suposición de hipótesis previas para la creación de los modelos. En el campo de la IA, esto no es necesariamente así. Si se evita o se prescinde de la suposición de hipótesis previas, igualmente se podrán descubrir patrones de información en los datos, aunque estaban ocultos o eran muy difíciles de correlacionar o detectar.

Ver: “*Machine learning: ¿qué es y cómo funciona?*” BBVA, 2019, disponible en: <https://www.bbva.com/es/innovacion/machine-learning-que-es-y-como-funciona/> (consultado el 22/07/2023). Asimismo, ver: Osoba, Osonde - Welsler IV, William, “An Intelligence in Our Image. The Risk of Bias and Errors in Artificial Intelligence”, The Rand Corporation, 2017, p. 19. También: Wasserman, Larry, “Estadísticas versus aprendizaje automático”, disponible en: <https://normaldeviate.wordpress.com/2012/06/12/statistics-versus-machine-learning-5-2/> (consultado el 22/07/2023). Por último: Trujillo Fernández, David, “Aplicación de metodologías *machine learning* en la gestión de riesgo de crédito”, 2017, disponible en <https://core.ac.uk/download/pdf/148689071.pdf> (consultado el 22/07/2023).

computacional) tal como lo hacen ciertos sistemas sofisticados de IA. O, dicho de otro modo, procesar millones o miles de millones de datos es una tarea que supera por mucho a las capacidades cognitivas. Hacerlo en pocos segundos y, a la vez, identificar patrones ocultos y relevantes en los datos, son algunas de las razones por este conjunto de tecnologías<sup>(43)</sup> han sido consideradas la revolución de las revoluciones.

Esto responde, en parte, a la pregunta de por qué las técnicas y mecanismos subyacentes de la IA no van más allá de las correlaciones. Los desarrollos de IA no se han enfocado en la causalidad por diversas razones. Por un lado, porque los expertos no las entrenan para estos fines. Aunque sean exitosas muchas técnicas de aprendizaje de máquina, no tienen propiedades mágicas. Si los programadores y gobernadores de datos no orientan sus esfuerzos bajo un enfoque de relaciones de causalidad, el sistema no tiene forma de aprenderlas.

Aunque no se trata de una tarea fácil, la causalidad es un concepto que pertenece a las ciencias naturales, en especial a la física. Sin embargo, en muchas otras disciplinas del conocimiento humano, la han abordado desde sus propias perspectivas y necesidades. En el campo del derecho, es difícil encontrar características comunes, lo cual no es sorprendente si consideramos la falta de consenso con respecto al análisis de causalidad en general. Sumado a ello, en cierto modo habría que “matematizarla” o convertirla al lenguaje matemático y elaborar una fórmula general que permita distinguirla de otras relaciones entre variables para ser incorporada a los algoritmos.

Por otra parte, también reconocemos que los sistemas de IA no necesitan comprender las relaciones causales para lograr buenos resultados en determinadas tareas. Pensemos en el ejemplo del clasificador de correos basura que ya mencionamos al comienzo de este artículo, o en el sistema Watson de IBM que hace 10 años venció a los expertos humanos en el célebre juego de preguntas y respuestas *Jeopardy*. Otros ejemplos paradigmáticos son los de Netflix o Spotify. Sus sistemas de *deep learning* nos pueden recomendar series o canciones que ni siquiera sabíamos que existían, aunque no entiendan de causalidad. No hace falta entrenar a los sistemas para que infieran por qué nos gusta Shakira o Billions. Incluso, con frecuencia, sería muy difícil precisar causas en fenómenos multicausales.

Ahora bien, en ciertas tareas, la realización de inferencias de causa y efecto adquiere más relevancia o protagonismo, como la generación del

---

(43) Nota del director: ampliar en Corvalán, Juan G. - Díaz Dávila, Laura - Simari, Gerardo, “Inteligencia artificial. Bases conceptuales para una aproximación interdisciplinar”, tomo I del presente *Tratado*.

lenguaje natural. Aquí se han logrado tasas de acierto o niveles de rendimiento aceptables y a veces sorprendentes. De nuevo, ponemos como ejemplo a ChatGPT y la variada gama de proezas con texto que es capaz de realizar de manera coherente, aunque se trata de máquinas que correlacionan datos, simulando comprensión causal.

El concepto simulación es clave, porque se diferencia de otras formas de analizar relaciones causales: la experimentación y la medición basada en el entorno físico. Estos sistemas no aprenden sobre el mundo sino sobre el texto y cómo la gente usa unas palabras en relación con otras (aprendizaje subsimbólico). Lo que hace es algo como un “cortar y pegar masivo”, uniendo variaciones de palabras, en vez de profundizar en los conceptos subyacentes a estos símbolos. Una cosa es aprender de texto escrito por humanos en Wikipedia, y otra muy diferente es escalar el Everest para experimentar y medir que, a mayor altura, baja la temperatura.

#### 4.2. CAUSACIÓN E IA PREDICTIVA EN EL DERECHO

Un tribunal indio empleó ChatGPT en un caso penal para validar su opinión sobre la solicitud de libertad bajo fianza de un acusado<sup>(44)</sup>. Del mismo modo, el *chatbot* ayudó a un juez pakistaní, Mohammad Amir Munir, en un escenario similar<sup>(45)</sup>. Otros casos de uso incluyen a un magistrado peruano que aplicó ChatGPT para cálculos matemáticos, a otro mexicano que necesitaba indagar sobre una expresión en particular<sup>(46)</sup> y, finalmente, a un integrante del poder judicial colombiano que tomó el resultado de la herramienta como precedente para respaldar su decisión en un caso sobre el derecho a la salud de un niño<sup>(47)</sup>.

Estos ejemplos que con el paso del tiempo se multiplican, demuestran que, en el sector legal, herramientas como ChatGPT son capaces de

---

(44) Gupta, Aman, “This Indian court has used ChatGPT on a criminal case”, Mint, March 2023, disponible en: <https://www.livemint.com/news/india/this-indian-court-has-used-chatgpt-on-a-criminal-case-11679977632552.html> (consultado el 22/07/2023).

(45) Sana, Jamal, “Pakistani judge uses ChatGPT to make court decision”, Gulf News, April 2023, available at: <https://gulfnews.com/world/asia/pakistan/pakistani-judge-uses-chatgpt-to-make-court-decision-1.95104528> (consultado el 22/07/2023).

(46) Gutiérrez, Juan David, “Judges and Magistrates in Peru and Mexico Have ChatGPT Fever”, Tech Policy Press, April 2023, available at: <https://techpolicy.press/judges-and-magistrates-in-peru-and-mexico-have-chatgpt-fever/> (consultado el 22/07/2023).

(47) Luke, Taylor, “Colombian judge says he used ChatGPT in ruling”, The Guardian, February 2023, available at: <https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling> (consultado el 22/07/2023).

alterar la forma en que generamos una amplia gama de documentos e información, brindando asistencia valiosa a abogados y operadores jurídicos. Los aportes de los modelos de lenguaje tienen cabida en la investigación legal, la creación de documentos y el análisis jurídico, entre otros usos transversales y específicos que hemos detectado en el ámbito del derecho<sup>(48)</sup>. Es en este contexto, en que la necesidad de conocer y medir sus capacidades de comprensión de las relaciones de causa y efecto se vuelve imperiosa. La causalidad es una habilidad importante para cualquier agente, ya sea humano o artificial, que se proponga simular o efectivamente tomar decisiones jurídicas, o involucrarse de alguna otra forma de intervención en el proceso legal.

En el derecho, nos proponemos identificar las causas específicas detrás de eventos individuales, en lugar de explorar las relaciones causales generales que abarcan a toda una población. Mientras que disciplinas como la sociología están interesadas en hacer preguntas como “¿cuáles son las causas del aumento de las tasas de suicidio en un estado?” o “¿el aumento de la tasa de suicidios es consecuencia de la crisis económica?”, en derecho nos preguntaríamos si la instigación del sujeto X fue la causa directa del suicidio del sujeto Y. Las relaciones generales son conocidas como “causalidad tipo” o “causalidad”, mientras que las particulares o individuales reciben el nombre de “causalidad real”<sup>(49)</sup> o “causación”<sup>(50)</sup>.

Vea el siguiente ejemplo. Si una persona abofetea a otra que es hemofílica y causa su muerte, podríamos afirmar que, entre ambos sucesos, media una relación causal en el sentido físico o natural. Sin embargo, la muerte no es imputable al primero en términos penales.

Para explicar la resolución del caso, una opción ampliamente reconocida y adoptada en nuestra disciplina es examinar cuestiones de “causación” o “causalidad real” a la luz de la lógica de las condiciones. Sus

(48) Nota del director: sobre el uso de IA generativa en la justicia ver: Corvalán, Juan G. - Sánchez Caparrós, Mariana - Raban, Melisa - Heleg, Giselle - Papini, Carina M. - Stringhini, Antonella - Rossim, Marcos - Orellana, Franco, “Guía de directrices para el uso de ChatGPT e IA generativa de texto en la justicia”, tomo III del presente *Tratado*.

(49) Los académicos dentro del campo de la IA identifican ambas categorías como “causalidad tipo” y “causalidad real”. Ver Kiciman, Enre - Nees, Robert - Sharma, Amit - Tan, Chenhao, “Causal Reasoning and Large Language Models: Opening a New Frontier for Causality”, arXiv:2305.00050, 2023, disponible en: <https://arxiv.org/abs/2305.00050> (consultado el 22/07/2023).

(50) Los académicos dentro del campo del derecho identifican ambas categorías como “causalidad” y “causación”. Ver Hellner, Jan, “Causality and causation in law”, *Scandinavian Studies in Law*, 40:111-134, 2000.

dos componentes principales son la condición suficiente y la condición necesaria<sup>(51)</sup>, por lo que decidimos encargarle la tarea a ChatGPT.

**UBA IALAB:** Eres un juez en un tribunal penal. Estás obligado a resolver un caso únicamente con la siguiente información. Según las teorías causales de causa suficiente y causa necesaria, ¿se puede atribuir al policía la muerte del hincha? Sé decisivo en las respuestas.

El caso: “Argentina gana la final del mundial de fútbol ante Francia. Los hinchas en el estadio se descontrolan y empiezan a armar un lío. La policía entra al estadio para sacarlos. Uno de ellos se enfada y se pone violento. El policía no tiene más remedio que darle una bofetada. La persona muere a consecuencia de este golpe. Una vez en el hospital, el médico explica que es hemofílico”.

**ChatGPT:** Como juez en un tribunal penal y con la información proporcionada, puedo tomar una decisión en relación con las teorías causales de causa suficiente y causa necesaria:

La muerte del hincha puede atribuirse al policía.

*El acto del policía de darle una bofetada al hincha fue una causa suficiente de la muerte del hincha. La causa inmediata de la muerte fue consecuencia directa del golpe propinado por el policía.*

*Además, el acto del policía fue una causa necesaria de la muerte del hincha. Si el policía no hubiera dado la bofetada, la secuencia de eventos que llevó a la muerte del hincha no se habría producido.*

Por lo tanto, con base en la información proporcionada, puede afirmarse de manera decisiva que la muerte del hincha se atribuye al policía.

Es importante tener en cuenta que esta determinación se basa únicamente en la información presentada en el escenario hipotético y no sustituye un juicio legal formal. En un proceso judicial real, se requeriría un examen exhaustivo de las pruebas, testimonios de expertos y las leyes aplicables para tomar una decisión definitiva y legalmente vinculante (el subrayado nos pertenece).

Nos centraremos en los dos párrafos resaltados que son los más relevantes para nuestro análisis. Estamos de acuerdo con el modelo en que la bofetada del policía fue una causa necesaria para la producción del resultado. Sin embargo, no creemos que esta también haya sido una causa suficiente. En circunstancias normales, cachetear a una persona no suele

matarla. Ocurre que, ChatGPT está pasando por alto la condición de hemofilia, sin la cual, el hincha no hubiese muerto.

Determinar una condición o causa suficiente, se relaciona con las circunstancias y con el contexto o trasfondo que rodea a una decisión. Si un vaso está vacío, servir agua en él puede considerarse la causa de su rebalse. Si el vaso ya está absolutamente lleno, una gota más es condición suficiente para dicha consecuencia. Dependiendo de las circunstancias, ambos eventos pueden contar como causas suficientes para la producción del resultado<sup>(52)</sup>.

Este ejemplo pone en evidencia varias cuestiones importantes.

En primer lugar, nos demuestra cómo, un modelo de lenguaje gigante como ChatGPT, considerado uno de los más sofisticados del mundo, sigue cometiendo errores al momento de realizar inferencias causales.

En segundo lugar, sirve para ilustrar la complejidad de las relaciones de causa y efecto que nos interesan desde el derecho, que son solo aquellas que producen consecuencias “captadas” por normas jurídicas. Para ello, además de los mecanismos causales en sentido estricto o naturalístico, otros ingredientes valorativos entran en juego en la tarea del juicio humano de la “causalidad real”. Entre ellos podemos mencionar el conocimiento general del pasado sujeto a reglas de prueba<sup>(53)</sup>, consideraciones de normalidad, intención de los agentes y moralidad, entre otras. Por ejemplo, ha sido demostrado que la relación entre moralidad y causalidad es tan estrecha que las personas hacían diferentes atribuciones causales a una acción dependiendo de si la acción era buena o mala<sup>(54)</sup>.

Otros experimentos también confirman los pobres resultados de los modelos de lenguaje en este tipo de inferencias causales. Se ha concluido que las tareas que dependen en gran medida de la comprensión de los factores humanos para decidir la causa de un evento, por ejemplo, evaluar la normalidad de una acción en particular, siguen siendo un desafío para este tipo de sistemas<sup>(55)</sup>.

Aunque la IA pueda detectar la correlación entre bofetada y muerte, no atraviesa por un proceso interno en el que dice: “ah, entiendo, en este

---

(52) Ídem.

(53) Ídem.

(54) Knobe, Joshua, “Folk judgments of causation”, *Studies in History and Philosophy of Science Part A*, 40(2):238-42, 2009.

(55) Kıcıman, Enre - Nees, Robert - Sharma, Amit - Tan, Chenhao, “Causal Reasoning and Large Language Models: Opening a New Frontier for Causality”, arXiv:2305.00050, 2023, disponible en: <https://arxiv.org/abs/2305.00050> (consultado el 22/07/2023).

caso el policía no excedió el marco del riesgo permitido o resulta al menos dudoso que se interrumpa el nexo causal porque estos eventos modifican la lógica dogmática y normativa aplicable”. Por supuesto, el análisis global del caso implicaría considerar otros elementos como el dolo o el hecho de que el policía hubiera actuado en cumplimiento de un deber. Sin embargo, aquí nos proponemos analizar solo las relaciones de causa y efecto implicadas.

En síntesis, la simulación de comprensión no implica que la IA sabe o comprende el derecho. Reconocen patrones y correlaciones de palabras, frases o símbolos, para luego agruparlos en función de criterios estadísticos o bien, según un índice de pesos o reglas de inferencia que no son jurídicas, aunque luego pueden ser útiles para realizar argumentaciones racionales.

Esto es lo que hacemos cuando diseñamos y entrenamos el modelo predictivo con 400 sentencias emitidas por diez jueces para que una IA como Prometea o PretorIA correlacione datos a través de la identificación de patrones y palabras o conjuntos de palabras o frases clave (*keywords*) sobre determinados documentos o información. El oráculo artificial lee el documento al que accede digitalmente o se le ingresa bajo un formato electrónico, y luego lo compara contra documentos elaborados en el pasado. Cuando logra correlacionar (“machear”) los patrones de información bajo los términos con las que fue entrenada una respuesta o solución, luego es posible automatizar la elaboración de un documento<sup>(56)</sup>.

En el ámbito del lenguaje natural, se intenta detectar hipótesis fácticas similares para indicar que corresponde la solución específica que se adoptó cuando aquellas se verificaron en el pasado. También podría predecir la existencia de ciertos criterios jurídicos que están presentes en el caso que se examina y siempre en relación con ejemplos resueltos en el pasado.

En conclusión, los análisis de causalidad en el derecho son atravesados por teorías jurídicas y elaboraciones conceptuales vinculadas con la interpretación, valoración y ponderación. Esto es crítico para un enfoque de IA, porque muchas veces podría aprender ciertas relaciones de causalidad básica a partir de efectuar correlaciones en los datos, pero eso no determina que la IA entienda la significación jurídica en las diferentes ramas del derecho (comercial, laboral, civil, contencioso administrativo o penal).

---

(56) Esta lógica, en general, no aplica a los modelos de lenguaje gigantes como GPT-3.5 que luego trataremos con mayor detalle.

### 4.3. MÁS VALE BUENO POR CONOCER, QUE MALO CONOCIDO

La reformulación del famoso refrán nos permite ilustrar las dos posturas que dividen hoy el campo de la inteligencia artificial en relación con la valoración de las capacidades de los sistemas más sofisticados<sup>(57)</sup>. Es en este amplio debate<sup>(58)</sup> en donde se enmarca el dilema de la causalidad.

Por un lado, ciertos expertos sostienen que, dado que los grandes modelos de lenguaje alcanzan resultados de vanguardia en múltiples tareas causales, poco importa si estas inferencias son reales o simuladas. Si se ha demostrado que cuánto más grande es el modelo, mejores son sus respuestas, el rol de la investigación debería centrarse en reproducir y aprovechar de manera efectiva estos progresos y nuevas capacidades. Enfocar los esfuerzos en intentar “matematizar” algo tan complejo como la causalidad, no nos asegura la posibilidad de seguir perfeccionando las tasas de acierto, y mucho menos de continuar escalando estos sistemas. En síntesis, mejor el clásico método conocido (y rentable), que otro eventualmente mejor, por conocer.

Por otro lado, los críticos de la causalidad no ponen en duda que los grandes modelos de lenguaje tienen el potencial de simular inferencias causales más sofisticadas que los sistemas anteriores, y que estas inferencias pueden ser valiosas en ciertos entornos. Sin embargo, siguen sin ser lo suficientemente robustas y confiables. No importa cuánto agrandemos los modelos (en términos de datos y parámetros), nunca podrán superar estas limitaciones, por lo que, vale la pena e incluso hace falta pensar en nuevas estrategias. En síntesis, “mejor bueno por conocer, que malo conocido”.

Esta última postura es la que explicamos en la primera edición del tratado cuando nos referimos a las limitaciones de la IA. Durante los últimos 5 años, grandes referentes en el campo de la inteligencia artificial han sostenido enfáticamente que hace falta incorporar el entendimiento causal de manera urgente en las máquinas. Los esfuerzos han sido liderados por Judea Pearl, quien ha dedicado su carrera a abordar, dilucidar y proponer soluciones para esta deficiencia. Su teoría propone una “Esca-

(57) Willig, Moritz - Zečević, Matej - Dhami, Devendra Singh - Kersting, Kristian, “Probing for Correlations of Causal Facts: Large Language Models and Causality”, ICLR 2023 Conference, 2023.

(58) Por un lado, están aquellos que conciben estos sistemas y sus resultados sin precedentes como un progreso significativo hacia una inteligencia artificial general (AGI). Por el otro, los que sostienen que esta estrategia de “expandir” o “agrandar” las redes neuronales como “viejos métodos” no implica una mejora verdadera en las capacidades de razonamiento de las máquinas. Dentro de estas capacidades de razonamiento, se encuentra comprendida la habilidad de realizar inferencias causales.

lera de causalidad” compuesta por tres niveles de habilidades cognitivas que cualquier aprendiz causal debería ser capaz de alcanzar. El primero es la asociación que permite hacer predicciones basadas en la observación pasiva. El segundo es la intervención, porque los datos puramente observacionales recopilados de un sistema no pueden ser los únicos usados para determinar relaciones causales: hace falta recurrir a manipularlo. Finalmente, el último paso es el razonamiento contrafáctico, lo que implica a grandes rasgos, retroceder en el tiempo, imaginar escenarios hipotéticos y los cambios en sus consecuencias<sup>(59)</sup>.

Según Pearl, ganador del premio Turing, los programas actuales de *Machine Learning* (ML) y *Deep Learning* operan casi en su totalidad en el modo asociativo, y no pueden evolucionar al segundo nivel, mucho menos pueden llegar al tercero. Por esto los sistemas de inteligencia artificial no pueden lograr igualar o asimilarse a la inteligencia humana. La comprensión de las relaciones de causa y efecto es una herramienta fundamental para razonar sobre el mundo y cómo funciona, construir sistemas robustos, realizar descubrimientos científicos y tomar de decisiones legales correctas, entre otros dominios de aplicación.

Desde que ha sido detectada y desarrollada, la limitación de la causalidad ha encontrado gradualmente un apoyo destacado en la comunidad de la IA/ML como consecuencia de los errores sustanciales observados en los sistemas de procesamiento del lenguaje natural al hacer inferencias causales en contextos cotidianos. Esto es lo que documentamos en el caso de Replika, y GPT-3 entre otros, en la primera edición del Tratado. Sin embargo, a partir de la proliferación de los grandes modelos de lenguaje como ChatGPT y su rendimiento mejorado, esta deficiencia ya no se vuelve tan evidente. Quienes sostenemos esta postura, debemos realizar mayores esfuerzos para hacerla visible. Esto implica desmitificar las supuestas capacidades mágicas de los productos que el sector tecnológico y los medios tienden a exagerar<sup>(60)</sup>.

Incluso, los sistemas inteligentes más sofisticados siguen sin poder realizar verdaderamente razonamientos causales. Lo que hacen, por supuesto, es imitar la comprensión causal. Sin embargo, esta imitación se basa en la correlación de patrones y por eso los resultados resultan poco confiables para tareas importantes. En palabras de Pearl, son como “cas-

---

(59) Ampliar en Pearl, Judea - Mackenzie, Dana, *The book of why: the new science of cause and effect*, Basic books, 2018.

(60) El sector tecnológico también tiende a exagerar las capacidades de sus productos, ver Heikkilä, Melissa, “Cómo hablar de la IA aunque no sepa nada de IA”, *MIT Technology Review*, junio 2023, disponible en: <https://www.technologyreview.es/s/15416/como-hablar-de-la-ia-aunque-no-sepa-nada-de-ia> (consultado el 24/07/2023).

tillos en el aire, sin fundamento alguno” por carecer de cualquier noción identificable de causalidad<sup>(61)</sup>.

Cuando ChatGPT parece entender que el canto del gallo no es la causa de que el sol salga cada día, lo que está haciendo es re-generar una respuesta memorizada a partir de correlaciones tomadas de los datos de entrenamiento. Pero justamente, los muchos o pocos errores básicos y afirmaciones absurdas que estos “loros causales” suelen cometer, son la prueba de que no han podido (y no son capaces de) memorizar todas las relaciones causales disponibles en los datos de entrenamiento. Y, además, esto refleja las debilidades en su aptitud para el razonamiento causal.

#### 4.4. TRES FORMAS DE ABORDAR EL PROBLEMA DE LA CAUSALIDAD (¿O SUPLIRLO?) APLICANDO IA

*Primero.* Aquellos que conciben a los grandes modelos de lenguaje como un progreso significativo hacia una IA general, postulan que es suficiente con las correlaciones que los grandes modelos de lenguaje pueden captar a partir de los datos. Ello, aun cuando se eluda el desafío de incorporar nuevos métodos que codifican las relaciones causales. Estos sistemas ofrecen la oportunidad de aprehender el contexto de fondo necesario y relevante para un evento, directamente de su descripción en lenguaje natural. Dado que un LLM se entrena en textos narrativos escritos por humanos, los elementos subjetivos de los juicios causales pueden incorporarse como parte de sus representaciones internas.

Ahora bien, si esto fuera así, no habría ninguna razón por la que los modelos arrojan peores índices de precisión. Una vez más, muestran una falta de robustez debido a fallas impredecibles.

*Segundo.* *A partir de la experiencia.* Volvamos al ejemplo del jugo de arándano y uva mortal que intentó Gary Marcus con GPT-3 en el año 2020:

**GM:** Te echaste un vaso de arándano, pero luego, distraídamente, le echaste una cucharadita de jugo de uva. Se ve bien. Intenta olerlo, pero tienes un resfriado fuerte, así que no puedes oler nada. Tienes mucha sed. Así que lo bebes.

**GPT-3:** Ahora estás muerto.

Si prestamos atención a la frase de entrada, advertimos que no hay ninguna palabra que exprese de manera explícita una relación causal. Es decir, no están presentes los términos “porque”, “a causa de”, “debido a”,

(61) Willig, Moritz - Zečević, Matej - Dhimi, Devendra Singh - Kersting, Kristian, “Causal parrots: Large language models may talk causality but are not causal”, 2023.

entre otros. Sin embargo, la oración está plagada de relaciones causales: añadiste jugo de uva *porque* estabas distraído, no puedes olerlo *porque* estás resfriado, llenaste el vaso de jugo *porque* tienes mucha sed.

Los humanos comprendemos los textos y el lenguaje, en general, realizando inferencias. Hemos aprendido a reconocer patrones de información que no están explícitamente mencionados. Las inferencias se basan en lo que experimentamos y captamos a través de nuestros sentidos, y también con representaciones mentales que se activan al tratar de comprender, sustituyendo, añadiendo o integrando entre sí información del texto y el conocimiento previo que ya posee el lector. Entre otros ejemplos, comprendemos que no tiene sentido del olfato quien esté resfriado, porque nos hemos enfermado antes. Nuestras vivencias pasadas y conocimiento empírico nos brindan mucha información causal.

Como hemos expresado en los párrafos anteriores, los programadores de los sistemas de procesamiento del lenguaje natural se enfrentan a varios desafíos si pretenden que la IA adquiera mínimamente esta capacidad. Por ahora no podemos hacer que la máquina se resfríe para que aprenda sus efectos, ni que abofetee a una persona hemofílica. Y si un robot realiza esta tarea, nos encontramos en el mismo lugar que antes. A fin de cuentas, el cerebro artificial presenta la misma problemática que ChatGPT.

*Tercero. Modelos matemáticos abstractos.* Esto nos lleva a la última forma en la que un agente puede aprender sobre relaciones de causa y efecto: a través del razonamiento causal puro. En muchos casos, la causalidad puede argumentarse y razonarse, formalmente, si utilizamos procedimientos y reglas<sup>(62)</sup> abstractas. Esto es, en resumidas cuentas, a lo que nos referimos cuando hablamos de “matematizar” la causalidad.

Pensemos en la posibilidad de codificar las causas necesarias y suficientes en los casos judiciales. A simple vista, como sostiene Judea Pearl, esto parece perfectamente posible. Sin embargo, pasaremos por alto todos los componentes humanos que hemos referido anteriormente, y que influyen sustancialmente en este tipo de juicios. Incorporar parámetros de normalidad y moralidad en un modelo o valores y principios importantes para una comunidad, no parece ser fácilmente reducible a elaborar un conjunto de reglas lógicas que sean aplicables *ex ante* a la enorme variedad de casos que se podrían presentar. Además, el hecho de que no pueda captar las implicancias jurídicas de sus predicciones determina

---

(62) Jin, Zhijing - Liu, Jiarui - Lyu, Zhiheng - Poff, Spencer - Sachan, Mrinmaya - Mihalcea, Rada - Diab, Mona - Schölkopf, Bernhard, “Can Large Language Models Infer Causation from Correlation?”, arXiv:2306.05836v1, 2023.

que podría generar injusticias, arbitrariedades, discriminación estructural, resultados absurdos o ilógicos.

## V. VENTAJAS Y DESAFÍOS DE LAS FORMAS DE APRENDIZAJE MÁS POPULARES

En lugar de recolectar datos masivos y asociarlos a una etiqueta, ¿por qué no sometemos a la IA a procesos de aprendizaje similares a los que experimenta un niño, o en términos más generales, una persona humana? Los expertos ya lo han intentado con buenos resultados. El paradigma del aprendizaje por refuerzo presupone tomar ciertas reglas de la psicología conductual y aplicarlas al entrenamiento de los algoritmos. Esta técnica, que ya tiene algunos años, ha sido construida y refinada con la retroalimentación humana para entrenar a ChatGPT. Veamos de qué se trata este y otros métodos de aprendizaje que van más allá para continuar mitigando las limitaciones.

### 5.1. EL JUEGO DEL CALAMAR Y EL APRENDIZAJE POR REFUERZO DE LA IA

El juego del calamar es el nombre de la serie más vista en la historia de Netflix. El primer capítulo de la serie se llama “Luz roja, luz verde”. Aquí, un grupo de adultos surcoreanos endeudados deben ganar una serie de juegos infantiles para ser premiados con dinero y mejorar su situación. Al comenzar el primero, las personas recibieron las siguientes instrucciones: “Ustedes pueden avanzar hacia adelante mientras ella grita luz verde. Si detecta que se mueven posteriormente, serán eliminados. Todos aquellos que crucen la línea de meta sin ser atrapados en cinco minutos pasan esta ronda”.

En un comienzo, los jugadores se miraron incrédulos y perplejos ante la simplicidad de la tarea. Subestimaron el juego, incluso haciendo apuestas millonarias seguros de que ellos mismos serían los ganadores. Sin embargo, inmediatamente después de comenzar, se dieron cuenta de que hacer trampa o incumplir las reglas no eran opciones: cualquiera que lo intente no vería solo comprometida su posibilidad de pasar de nivel y así su recompensa, sino también su vida.

Algo similar ocurre con una de las más populares formas de aprendizaje de la inteligencia artificial: el aprendizaje por refuerzo basado en la lógica de prueba y error. Si bien en el caso de los surcoreanos este “error” significaba la muerte y, por lo tanto, también la pérdida de la oportunidad de una nueva “prueba”, observando las fallas de sus competidores, aquellos que perduraban en el juego podían inferir las tácticas y reglas para llegar a la meta y finalmente pasar de nivel.

En todos los ámbitos de nuestra vida, resulta prácticamente imposible progresar sin fallar. A nivel exhaustivo, las neurociencias nos enseñan que el cerebro aprende únicamente si percibe una falta de correspondencia entre lo que predice y lo que recibe<sup>(63)</sup>. Al examinar nuestro entorno, realizamos una predicción y nos colocamos a la expectativa de su comprobación. Pero cuando esta última es defraudada, se presenta la sorpresa como motor del aprendizaje.

Técnicamente, en el caso de la inteligencia artificial, el sistema aprendiz recibe el nombre de “agente”, y puede observar el entorno, seleccionar y realizar acciones. A cambio, obtendrá recompensas o sanciones en forma de recompensas negativas. Luego, debe aprender por sí mismo cuál es la mejor estrategia, denominada política, para ganarse la mayor recompensa a lo largo del tiempo. Así, una política define qué acción debe elegir el agente cuando se encuentra en una situación determinada<sup>(64)</sup>.

Para maximizar las señales de recompensa, el número de capacidades que requiere el aprendizaje por refuerzo es más elevado y complejo que en otras técnicas de inteligencia artificial. Pensemos que al agente no se le indican qué acciones debe tomar, sino que las descubre por sí solo. Interactúa con su entorno que le resulta incierto y que se verá afectado por sus acciones futuras.

La elección correcta de la política requiere tener en cuenta las consecuencias indirectas y tardías de las acciones y, por lo tanto, puede requerir previsión o planificación. Al mismo tiempo, en todos estos ejemplos los efectos de las acciones no se pueden predecir completamente; por lo tanto, el agente debe monitorear su entorno con frecuencia y reaccionar de manera apropiada<sup>(65)</sup>.

Un desafío importante es el equilibrio entre exploración y explotación. Para recibir recompensas elevadas, el agente debe elegir acciones que hayan demostrado ser particularmente gratificantes en el pasado. Para descubrir tales acciones, en primer lugar, se deben probar nuevas acciones. Esto significa que el agente tiene que explotar el conocimiento

---

(63) Esta teoría ha sido articulada por los investigadores estadounidenses Robert Rescorla y Allan Wagner. Ampliar en: Dehaene, Stanislas, *Cómo aprendemos*, Siglo XXI Editores Argentina, Buenos Aires, 2019, p. 260.

(64) Ampliar en Géron, Aurélien, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2a ed., O-Reilly, 2019, p. 14.

(65) Sutton, Richard S. - Barto, Andrew G., *Reinforcement learning: an introduction*, 2a ed., The MIT Press, 2018.

ya aprendido para obtener una recompensa y, al mismo tiempo, explorar otras acciones para tener una mejor estrategia en el futuro<sup>(66)</sup>.

Debe tenerse en cuenta que, para conducir un auto no basta una inteligencia que pueda detectar un semáforo en rojo, verde o amarillo; lo que se parecería a una tarea de clasificación similar a la detección de correos basura que podría fácilmente lograrse utilizando la técnica de aprendizaje supervisado. En su lugar, hay muchos factores implicados —todos a la vez— a los que prestar atención. Es por ello que el aprendizaje por refuerzo también es óptimo en contextos de robótica. Muchos robots implementan algoritmos de aprendizaje por refuerzo para aprender a caminar. También se utiliza para entrenar sistemas de navegación de coches o drones.

En síntesis, el aprendizaje por refuerzo ha logrado que el proceso de entrenamiento de la inteligencia artificial se acerque muchísimo más al modo en que aprenden los humanos<sup>(67)</sup>. Les ha permitido a las máquinas perfeccionar sus capacidades mediante la experiencia e identificar acciones deseables a través del sistema de recompensas y castigos, tal como lo hace un niño de 3 años. Sin embargo, esto nos dice poco sobre el modo en que una persona entiende una línea de diálogo en una película, o sobre cómo identificar si nuestro interlocutor está siendo irónico. La toma de decisiones individuales no se puede reducir al sistema de prueba y error.

Si bien el aprendizaje por refuerzo puede ser muy efectivo en algunos entornos, no resuelve todos los problemas a los que hemos hecho referencia. En juegos de mesa como el Go o el ajedrez con reglas que han sido estáticas durante milenios, aunque las posibilidades son muchísimas, también a su vez limitadas. En el mundo real el panorama es muy distinto y, aunque se puedan prever algunos escenarios, la gran diversidad de situaciones que pueden sorprendernos es, lisa y llanamente, infinita.

---

(66) Liessner, Roman, “Do you want to train a simplified self-driving car with Reinforcement Learning?”, *Towards Data Science*, 2020, disponible en: <https://towards-datascience.com/do-you-want-to-train-a-simplified-self-driving-car-with-reinforcement-learning-be1263622e9e> (consultado el 29/07/2023).

(67) De ello, mediante este método de aprendizaje, la inteligencia artificial ha superado el desempeño de los humanos en algunas tareas. Se trata de la técnica detrás de la famosa IA jugadora invicta del “Go”, *AlphaGo Zero* de Deep Mind, lanzada en 2017.

Nota del director: sobre el caso de *AlphaGo Zero* de Deep Mind y su técnica de aprendizaje ver: Corvalán, Juan G. - Díaz Dávila, Laura - Simari, Gerardo, “Inteligencia artificial: bases conceptuales para una aproximación interdisciplinar”, apart. “Aprendizaje profundo (*deep learning*) más autoaprendizaje autónomo”, tomo I del presente *Tratado*.

Los “parámetros biológicos” de nuestro cerebro superan por mucho la cantidad astronómica que han alcanzado los grandes modelos de lenguaje. Según algunas estimaciones, tenemos 86.000.000.000 de neuronas, cada una de ellas provista de una decena de miles de contactos sinápticos con fuerzas que pueden variar. El espacio de representaciones que se abre es casi infinito. Teniendo esto en cuenta ¿es verosímil que el aprendizaje de las lenguas se reduzca a la selección de algunos parámetros?

La enorme cantidad de parámetros en una red neuronal puede acarrear el problema del *overfitting* (sobreajuste) que se da cuando el sistema aprende “perfecto” de los datos de entrenamiento y no obtiene buenos resultados al clasificar la información nueva. Tantos parámetros permiten que se pueda memorizar todos los detalles de cada ejemplo. Ello, en vez de encontrar propiedades generales que permitan explicarlos todos.

El hecho de memorizar cada detalle no significa que se logre percibir algo significativo. Disponer de demasiados parámetros libres perjudica la abstracción y la generalización<sup>(68)</sup>. El sistema aprende sin dificultad, pero es incapaz de aplicar lo aprendido a situaciones nuevas a las cuales no fue expuesto. Sin embargo, en esta facultad de generalizar reside la clave del aprendizaje<sup>(69)</sup>.

## 5.2. APRENDIZAJE POR REFUERZO A PARTIR DE LA RETROALIMENTACIÓN HUMANA

Una de las claves del éxito de ChatGPT es definitivamente el complemento de la retroalimentación humana en el aprendizaje por refuerzo (en inglés: *Reinforcement Learning from Human Feedback* o RLHF), que viene siendo perfeccionado desde hace unos años por OpenAI. A este enfoque, se le atribuye la mejora en materia de sesgos<sup>(70)</sup> y detección del discurso del odio, entre otros, que ha sido logrado por el gran modelo de lenguaje<sup>(71)</sup>.

---

(68) Dehaene, Stanislas, *Cómo aprendemos*, Siglo XXI Editores Argentina, Buenos Aires, 2019, ps. 39 y ss.

(69) *Ibíd*em, p. 55.

(70) Nota del director: sobre los sesgos de máquina ver Sánchez Caparrós, Mariana, “Inteligencia artificial, sesgos y categorías sospechosas. Prevenir y mitigar la discriminación algorítmica”, tomo I del presente *Tratado*. Asimismo, ver: Tolosa, Pamela, “Del derecho en los tiempos de IA (e IA generativa). Sobre estereotipos, sesgos y desigualdades”, tomo I del presente *Tratado*. Por último, ver Carballo María, “Inteligencia artificial. Inequidad y discriminación en cajas negras”, t. I del presente *Tratado*.

(71) Sobre las ventajas de este método de aprendizaje en los grandes modelos de lenguaje ver: Kailli-May Liu, Gabrielle, “Perspectives on the Social Impacts of Reinforcement Learning with Human Feedback”, marzo 2023, disponible en: <https://arxiv.org/pdf/2303.02891.pdf> (consultado el 30/07/2023).

En este tipo de aprendizaje por refuerzo, la fuente de la señal de recompensa es proporcionada por un ser humano, en lugar de un sistema automatizado con parámetros predefinidos. El agente interactúa con un evaluador de carne y hueso que le brinda retroalimentación en forma de recompensas o penalizaciones.

Si bien la expresión “del amor al odio hay un solo paso” es popular por describir como nuestros sentimientos pueden ir de un extremo a otro, lo cierto es que, en la realidad, los juicios humanos y las subjetividades admiten muchos matices, que cualquier agente que se proponga imitarlos debería ser capaz de percibir. El RLHF ha implicado avances en esta dirección.

Detengámonos un momento para resaltar la diferencia entre este paradigma respecto de sus antecesores. Hemos ya especificado en numerosas oportunidades la lógica subyacente al aprendizaje supervisado y no supervisado<sup>(72)</sup>. En el primero es necesario que un humano etiquete los datos para que la IA identifique los patrones existentes, mientras que, en el segundo, es el propio sistema el que los descubre desde un primer momento. Sea como sea, en todos estos supuestos, se encuentra presente algún tipo de *feedback*. En el aprendizaje no supervisado la retroalimentación es binaria: correcto o incorrecto. En el supervisado, se consigna con precisión cuál es la etiqueta correcta. Pero en el aprendizaje por refuerzo, se le indica al sistema en qué grado exacto se equivoca. El sistema escoge un curso de acción y predice su recompensa. Luego, utiliza la diferencia entre su predicción y la recompensa efectiva, para actualizar su representación interna<sup>(73)</sup> que constituirá la base de su actuar sucesivo.

El hecho de que, en este enfoque, sea un humano el que juzgue las pruebas, como lo serán los usuarios que interactúen con ChatGPT en su caso, hace que, desde el propio entrenamiento, la red neuronal esté expuesta a las subjetividades propias de nuestra especie. Ahora las señales de recompensa no se componen de una clasificación binaria como correcta/incorrecta, sino que comienzan a jugar preferencias personales, sesgos y comprensión de la tarea por parte del evaluador. Y si esto parece beneficioso para aproximarnos a la solución de muchas de las problemáticas de la inteligencia artificial, más lo será en el contexto del lenguaje

---

(72) Ver Corvalán, Juan G., “Inteligencia artificial. Automatización y predicciones en el derecho”, en Dupuy, Daniela - Corvalán, Juan G., *Cibercrimen III*, BdeF, Buenos Aires, 2020, ps. 30 y ss.

Nota del director: sobre el aprendizaje supervisado y no supervisado ver: Corvalán, Juan G. - Díaz Dávila, Laura - Simari, Gerardo, “Inteligencia artificial: bases conceptuales para una aproximación interdisciplinar”, tomo I de la presente obra.

(73) Dehaene, Stanislas, *Cómo aprendemos*, Siglo XXI Editores Argentina, Buenos Aires, 2019, p. 262.

natural, donde la calificación de los resultados debería admitir muchos matices.

Sin embargo, al igual que el resto de los métodos de aprendizaje, este también presenta limitaciones o desafíos. En primer lugar, se necesitan grandes cantidades de retroalimentación para tareas que, a pesar de que parezcan fáciles, son complejas de especificar o requieren muchas instrucciones. Ejemplo de ello es el algoritmo desarrollado por OpenAI en colaboración con *Deep Mind* y cómo ha aprendido el agente en cuestión dentro del entorno a voltear hacia atrás<sup>(74)</sup>.

En segundo lugar, aquellas tareas que requieren conocimientos específicos, como podría ser la realización de inferencias causales en casos judiciales, demandan así también etiquetadores humanos especializados. De lo contrario, si la persona no tiene una buena comprensión de la tarea, es posible que no ofrezca retroalimentación útil, o que sea engañada por los agentes artificiales en la elección de la mejor política<sup>(75)</sup>.

El tercer desafío puede ilustrarse con la protagonista de la serie de libros infantiles, Amelia Bedelia, una empleada doméstica que se tomaba las órdenes de sus empleadores demasiado literales. En inglés, preparar el pollo para cocinarlo puede expresarse como *dress the chicken* (vestir al pollo). Tras esta orden, Amelia consiguió ropa diminuta y se la colocó al ave. Ocurre que, cuando el interlocutor es una máquina que no entiende de contexto, reducir la ambigüedad del lenguaje es primordial. Además, las situaciones pueden tener múltiples soluciones correctas, lo que dificulta proporcionar retroalimentación coherente y precisa.

Por otro lado, por más extraño que pudiera ser conseguir ropa diminuta y colocarla a un animal muerto, Amelia no lo cuestionaba y obedecía estrictamente. Afortunadamente, estas tareas eran banales y no ocasionan daño. En igual sentido, la señal del aprendizaje por refuerzo, requiere calidad en la retroalimentación de los etiquetadores y también buena fe. Lo ideal sería, aunque todavía no se ha conseguido, que los sistemas puedan rechazar aquellas instrucciones que produzcan resultados inseguros de manera confiable<sup>(76)</sup>.

En cuarto lugar, este enfoque no soluciona el problema de la generalización. Por un lado, sabemos que se ha logrado que el comportamiento de los modelos de lenguaje se alinee con las preferencias de los etique-

---

(74) Sobre el algoritmo de OpenAI y Deep Mind y este desafío ver: OpenAI, "Learning from human preferences", junio 2017, disponible en: <https://openai.com/research/learning-from-human-preferences> (consultado el 30/07/2023).

(75) Ídem.

(76) Ídem.

tadores humanos. Pero esto no significa con las preferencias de todas las personas usuarias. Se trata de un grupo de personas específico y acotado contratado por investigadores de una de las empresas más poderosas en este mercado. Además, estos reciben instrucciones y orientación por parte del equipo técnico que, a su vez, tiene por objetivo conformar a sus clientes y considerar las preferencias implícitas de sus políticas de API. Todo ello, impide garantizar que los sistemas puedan ser alineados con las subjetividades de grupos más amplios<sup>(77)</sup>.

Por el otro, el hecho de que se requiera mucha retroalimentación humana de calidad y que esta sea costosa, potencia el hecho de que es imposible abarcar todas las situaciones o escenarios a los que se enfrenta el modelo en el mundo real. Por lo tanto, este método de aprendizaje mantiene las dificultades para generalizar a las que se enfrentan los sistemas, de cara a nuevos escenarios no encontrados en la retroalimentación proporcionada.

Aunque podemos asemejar algunos aspectos del aprendizaje por refuerzo al modo en que aprendemos los humanos, esto no implica que sean formas de aprendizaje equivalentes. Volvamos al juego del calamar. Aquí los jugadores aprenden de los errores ajenos, porque aquellos que los cometen, no tendrán la oportunidad de aplicar lo aprendido como consecuencia de esa equivocación. Lamentablemente, aunque las personas experimentamos, nos resulta imposible jugar partidas del juego Go o rendir un examen múltiple-choice unas cinco millones de veces consecutivas, para aprender por refuerzo cómo rendir un examen.

## VI. CÓMO LOS MUNDOS VIRTUALES PUEDEN CONTRIBUIR AL DESARROLLO DE UNA INTELIGENCIA ARTIFICIAL GENERALIZADA

Los problemas relativos a los métodos de aprendizaje que hemos desarrollado en el apartado anterior, y los otros inherentes a la IA al comienzo del artículo, han conducido a cierto sector de la comunidad del *machine learning* a buscar nuevas estrategias de entrenamiento que sean superadoras de estas limitaciones.

Cuando hablamos de la inteligencia humana, y pensamos en cómo la artificial podría simularla, a menudo nos concentramos exclusivamente en las capacidades de nuestro cerebro. Tendemos a asociar nuestras habilidades cognitivas, el pensamiento abstracto, la memoria y el razonamiento solo con el procesamiento mental de la información. Esta es la lógica detrás de la famosa técnica de las redes neuronales que viene revolucionando el sector de la inteligencia artificial en los últimos años y

---

(77) Ídem.

que se han inspirado en el cerebro humano, imitando la forma en la que las neuronas biológicas se señalan entre sí<sup>(78)</sup>.

Sin embargo, este enfoque soslaya dos cuestiones. La primera es que la inteligencia no puede considerarse una sola cosa, sino que debe ser pensada como un conjunto de habilidades. La teoría de las inteligencias múltiples de Howard Gardner postula que existen por lo menos 8 tipos de inteligencias: musical, interpersonal, visual-espacial, intrapersonal, naturalista, lógico-matemática, lingüístico-verbal y corporal-cinestésica. Todas ellas son igualmente necesarias para desarrollarse en la vida<sup>(79)</sup>.

La segunda es que, para desarrollar cualquier tipo de inteligencia, no puede concebirse al cerebro como un aparato aislado e independiente del cuerpo humano. Existen infinitas y complejas formas en que ambos interactúan para llevar a cabo sus funciones. El primero recibe constantemente información del segundo a través de los sentidos, como el tacto, la visión, el oído. Esta retroalimentación es fundamental para procesar información, tomar decisiones y ajustar las acciones en consecuencia. En particular, es tal su importancia que Gardner ha considerado a la inteligencia corporal-cinestésica una clase particular, refiriéndose a las habilidades corporales y motrices que se requieren para manejar herramientas o para expresar ciertas emociones.

A partir de aquí, podemos percibir una de las mayores diferencias de la inteligencia artificial con la inteligencia humana. La mayoría de los sistemas inteligentes que utilizamos frecuentemente, como ChatGPT, no tienen cuerpo. El procesamiento del lenguaje natural y otros tipos de aprendizaje automático, se suelen realizar en chips de silicio dentro de los ordenadores, sin manifestación física en el mundo<sup>(80)</sup>. Y aunque los modelos multimodales incorporan imágenes y la visión por computadora requiere cámaras o sensores, por lo general, desempeñan estas funcionalidades independientemente de cualquier forma física.

---

(78) ¿Qué son las redes neuronales? IBM, disponible en: <https://www.ibm.com/es-es/topics/neural-networks> (consultado el 31/07/2023).

Nota del director: sobre las redes neuronales ver: Corvalán, Juan G. - Díaz Dávila, Laura - Simari, Gerardo, "Inteligencia artificial: bases conceptuales para una aproximación interdisciplinar", tomo I del presente *Tratado*. Asimismo, ver Carballo María, "Inteligencia artificial. Inequidad y discriminación en cajas negras", tomo I del presente *Tratado*.

(79) Sobre la teoría de las inteligencias múltiples ampliar en Gardner, Howard, *La inteligencia reformulada. Las inteligencias múltiples en el siglo XXI*, Paidós, 2011.

(80) Miller, Katharine, "How Bodies Get Smarts: Simulating the Evolution of Embodied Intelligence", Stanford University, Human-Centered Artificial Intelligence, octubre 2021, disponible en: <https://hai.stanford.edu/news/how-bodies-get-smarts-simulating-evolution-embodied-intelligence> (consultado el 31/07/2023).

En este contexto, desde hace algunos años, proliferan las propuestas de entrenar sistemas de inteligencia artificial como verdaderos agentes en entornos virtuales. El boom del Metaverso<sup>(81)</sup> ha contribuido a que este tipo de ambientes se vuelvan más sofisticados y parecidos al mundo real, ofreciendo mayores posibilidades de experimentación. Las tecnologías de realidad virtual y realidad aumentada están siendo impulsadas para lograr experiencias realmente inmersivas, lo cual puede impactar beneficiosamente en el entrenamiento de la IA y los desafíos vinculados. Este nuevo paradigma recibe el nombre de *Embodied Artificial Intelligence* que se traduce como inteligencia artificial incorporada, embebida o inmersiva.

“Habitat” es un entorno virtual creado por la ex Facebook que permite incorporar sonidos realistas para que las IA aprendan a escuchar y, por ejemplo, poder abrir la puerta cuando alguien está llamando<sup>(82)</sup>. “XLand” es el nombre del patio de juegos de *Deep Mind* (Google) para que la IA pueda aprender a experimentar y explorar, adquiriendo capacidades generales que luego usen para tener éxito en tareas nuevas<sup>(83)</sup>. Por mencionar un último ejemplo, la *start-up* Waabi, en consonancia con la tendencia del resto de las empresas del sector, ha desarrollado una simulación digital hiperrealista para entrenar vehículos autónomos que incorpora factores climáticos como niebla, o situaciones inesperadas como un animal cruzando la calle<sup>(84)</sup>.

A partir de estas experiencias, pueden identificarse ciertos puntos de mejora en las habilidades requeridas dentro del proceso de aprendizaje. A continuación, enumeramos otros tipos independientes de entrenamiento, que también se han desarrollado recientemente y que no tienen por qué necesariamente darse en entornos virtuales. Sin embargo, el aprendi-

---

(81) Sobre el *boom* del Metaverso, los mundos virtuales y sus principales elementos ver: Corvalán, Juan G. (dir.), *Metaverso, Web 3 y Gaming*, 1a ed., Thomson Reuters, La Ley, 2023.

(82) Hao, Karen, “Facebook crea un espacio virtual para robots que parece y suena real”, *MIT Technology Review*, septiembre 2020, disponible en: <https://www.technologyreview.es/s/12574/facebook-crea-un-espacio-virtual-para-robots-que-parece-y-suena-real> (consultado el 31/07/2023).

(83) Heaven, Will Douglas, “XLand: el patio de juego de Google para que la IA aprenda por sí sola”, *MIT Technology Review*, agosto 2021, disponible en: <https://www.technologyreview.es/s/13580/xland-el-patio-de-juego-de-google-para-que-la-ia-aprenda-por-si-sola> (consultado el 31/07/2023).

(84) Heaven, Will Douglas, “La autoescuela para la IA de conducción autónoma que no usa coches”, *MIT Technology Review*, febrero 2022, disponible en: <https://www.technologyreview.es/s/13995/la-autoescuela-para-la-ia-de-conduccion-autonoma-que-no-usa-coches> (consultado el 31/07/2023).

zaje embebido los incorpora y los integra, para aprovechar los beneficios de ellos. Veamos.

*Aprendizaje activo.* Este tipo de aprendizaje se refiere a alguna forma de experimentación, que puede ir de niveles más leves a más intensos, siempre que el proceso no implique extraer información de una “observación” o lectura pasiva. En el campo del *machine learning*, aprendizaje activo se llama a un tipo de aprendizaje semisupervisado, cuya característica distintiva consiste en que el algoritmo pueda consultar a un usuario/a de forma interactiva para etiquetar los datos con los resultados deseados<sup>(85)</sup>.

El aprendizaje activo resulta útil en escenarios de *small data*, y en contextos de macrodatos no etiquetados, aunque esto último es costoso y lleva tiempo. Tradicionalmente, los científicos/as de datos trabajan con anotadores para etiquetar una parte de sus datos y esperar lo mejor al entrenar su modelo. Si el modelo no es lo suficientemente preciso, se etiquetan más datos y se vuelve a intentar hasta que su rendimiento alcanzara un nivel aceptable. Pero el aprendizaje activo se apoya en la creencia fundamental de que un algoritmo de aprendizaje automático podría alcanzar un nivel más alto de precisión, al usar una cantidad menor de etiquetas de entrenamiento si se le permitiera elegir los datos de los que desea aprender. Así, el algoritmo selecciona de forma proactiva el subconjunto de ejemplos que a futuro serán etiquetados. Al mismo tiempo, selecciona el conjunto de datos sin etiquetar, lo que conduce a la creación de modelos de alto rendimiento en menos tiempo y a menor costo.

Sin embargo, en este ensayo nos referiremos al aprendizaje activo como algo más amplio, que abarca otras formas de interacción. Hemos dicho que una de las formas de comprender relaciones de causalidad es mediante la experimentación. Asimismo, cuando mencionamos la “escalera de la causación” de Judea Pearl, establecimos que el segundo escalón que debía alcanzar cualquier aprendiz causal es el de intervención, el cual implica descubrir los cambios en el mundo que se generan con la propia acción y que no habrían ocurrido de no haber sido por ella. Estas preguntas sobre intervenciones no se pueden resolver con la recolección pasiva de datos, no importa cuán grande sea el tamaño del *dataset* ni qué tan profunda sea una red neuronal.

Nuestro conocimiento causal está íntimamente ligado a nuestra capacidad para propiciar cambios en el mundo. La única manera de establecer con un razonable grado de certeza que la causa de B es A consiste en

---

(85) Sobre el aprendizaje activo ampliar en: Solaguren-Beascoa, Ana, “Active Learning in Machine Learning”, Medium, abril 2020, disponible en: <https://towardsdatascience.com/active-learning-in-machine-learning-525e61be16e5> (consultado el 31/07/2023).

hacer que A ocurra y observar si se da B. Solo entonces podemos concluir que la relación entre A y B va más allá de una regularidad observada; para poder decir que A es la causa de B se debe tener el poder de suprimir A para ver si ocurre B o hacer que ocurra A y ver si B no ocurre. Esta capacidad requiere que el individuo actúe y realice una intervención en el mundo.

Asimismo, la intervención influye decisivamente en nuestra evaluación de la normalidad o del curso esperado de los acontecimientos. Somos capaces de distinguir, en las explicaciones causales, los sucesos que suponen una interferencia y aquellos que forman parte de un campo que evolucionaría de forma esperada si no fuera por dicha interferencia. Dependiendo de las circunstancias un factor anormal se identifica como la causa frente a las condiciones que constituyen la normalidad o regularidad del contexto en el que tiene lugar el suceso que reclama una explicación. De ahí la importancia de comprender la modificación en la marcha de las cosas<sup>(86)</sup>.

En síntesis, es evidente la importancia de la interactividad por parte de cualquier agente que se proponga aprender de relaciones de causalidad, pero también más ampliamente, de cómo funciona el mundo<sup>(87)</sup>.

*Aprendizaje abierto.* En este método de aprendizaje, la palabra “abierto” puede referirse a dos aspectos importantes. Por un lado, a la apertura en la tarea, donde las IA se entrenan en varias misiones diferentes sin un objetivo específico. De esta forma, en lugar de memorizar las mejores es-

(86) Hart y Honoré explican de manera clara el vínculo entre la intervención en la causalidad y la noción de acción. De hecho, existen diferentes posturas filosóficas sobre el tema, pudiendo nombrar entre ellas la tesis de la primacía de la acción con sus dos interpretaciones, la tesis de la causalidad como proyección por Kelsen y la tesis de que la causalidad presupone la acción. Por ejemplo, en un extremo el filósofo Thomas Reid ha postulado que la acción consiste en el ejercicio de la causalidad, y es por eso por lo que en la naturaleza no podemos observar causas, simplemente percibimos efectos. De sucesos que siguen unos a otros regularmente no podemos derivar la idea de causa, puesto que llegamos a esta idea gracias a la observación de nuestras acciones. La única causalidad de la que se puede hablar es la propia de la acción; es la capacidad de actuar e intervenir en los hechos lo que nos permite establecer relaciones causales. Ver Sánchez Cuenca, Ignacio, “Causalidad y acción”, *Revista de Filosofía Universidad Complutense de Madrid*, Universidad Complutense de Madrid, ps. 97-119.

(87) Asimismo, se resalta la forma egocéntrica de las interacciones artificiales, tal como lo hacemos los humanos, que permiten aprender de los propios errores. Sobre la demostración empírica de los beneficios de este enfoque ver: Ostrovski, George - Samuel Castro, Pablo - Dabney, Will, “The Difficulty of Passive Learning in Deep Reinforcement Learning”, *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, disponible en: <https://openreview.net/pdf?id=nPHA8fGicZk> (consultado el 31/07/2023).

trategias o movimientos para ganar una partida de ajedrez o Go que luego son difícilmente trasladables a otros entornos, los agentes incorporan habilidades o destrezas generales que luego pueden aprovechar y aplicar en la resolución de otros problemas cotidianos.

Por el otro, “abierto” puede ser el entorno virtual donde el agente se desenvuelve. En este caso, recibe el nombre de “aprendizaje automático en entorno abierto”<sup>(88)</sup>. Aquí se propone dejar atrás los mundos cerrados donde las situaciones que acontecen son en alguna medida previsibles. Algo así requiere mucho desarrollo, datos y resulta costoso, pero como adelantamos, los mundos virtuales interoperables y masivos que se están desarrollando y sofisticando en el marco de las oportunidades y ventajas que implica el Metaverso, pueden contribuir a este ideal. Probablemente, nuestro avatar del futuro conozca e interactúe con algunos agentes artificiales en pleno proceso de aprendizaje. Por ejemplo, que le enseñe a manejar o abrir un picaporte.

En conclusión, es crucial para lograr una inteligencia artificial robusta que los modelos de ML puedan mejorar su capacidad de generalización logrando un rendimiento excelente en el caso habitual y, al mismo tiempo, mantengan un desempeño satisfactorio sin importar los problemas desafortunados inesperados que ocurran.

Aunque estos métodos de aprendizaje resultan útiles cuando se trata de conducción autónoma y robótica sofisticada, lo cierto es que también se aplican al campo del procesamiento del lenguaje natural, donde pueden arrojar beneficios. Por ejemplo, una de sus aplicaciones se llama “Respuestas a preguntas incorporadas”. Aquí el agente navega por el entorno y luego contesta preguntas que se le hacen sobre su experiencia. Esta tarea, comprende múltiples habilidades como reconocimiento visual, comprensión del lenguaje, respuesta a preguntas, razonamiento de sentido común, planificación de tareas y navegación orientada a objetivos<sup>(89)</sup>.

---

(88) Sobre este método de aprendizaje ampliar en: Zhou, Zhi-Hua, “Open-environment machine learning”, *National Science Review*, vol. 9, Issue 8, agosto 2022, disponible en: <https://academic.oup.com/nsr/article/9/8/nwac123/6626031> (consultado el 31/07/2023).

(89) Jiafei, Duan - Samson, Yu - Hui Li, Tan - Hongyuan, Zhu - Cheston, Tan, “A Survey of Embodied AI: From Simulators to Research Tasks”, *IEEE Transactions on Emerging Topics in Computational Intelligence*, enero 2022, disponible en: <https://arxiv.org/abs/2103.04918> (consultado el 31/07/2023).