DEMOCRACIA, INTELIGENCIA ARTIFICIAL Y DESAFÍOS REGLAMENTARIOS: DERECHOS, DILEMAS Y PODER EN LAS SOCIEDADES DATIFICADAS

Sivaldo Pereira da Silva[1]

Resumen: Este artículo tiene como objetivo identificar cuestiones clave en el debate sobre las políticas de Inteligencia Artificial (IA), teniendo en cuenta la construcción de un panorama regulatorio adecuado para este campo. El análisis se guió por la investigación documental y bibliográfica, tomando principios normativos de la teoría política y las teorías de la democracia. Después de resumir los aspectos más destacados del modus operandi de los sistemas de IA, el trabajo enumeró siete problemas políticos clave que están en la base de esta discusión: (1) personificación e imputabilidad de la máquina; (2) dilemas y juicios morales; (3) autoritarismo estadístico de las métricas; (4) oscurantismo matemático en los procesos; (5) omnipresencia utilitarista de los sistemas autónomos; (6) fronteras de eficiencia y control; (7) diversidad y representatividad en los códigos.

Palabras clave: Inteligencia Artificial; Gobernanza algorítmica; Filosofía de la tecnología; Comunicación digital y regulación; Políticas públicas

1 Introducción

Las innovaciones que ofrecen los sistemas y máquinas basados en la Inteligencia Artificial (IA) constituyen hoy un impulso relevante en la eficiencia de los procesos en los más diversos ámbitos como la comunicación, la política, el transporte, la seguridad pública, la salud, la educación, etc. La tendencia, para las próximas décadas, es que habrá saltos significativos en este campo, con el horizonte cada vez más cotidiano y omnipresente. Estas tecnologías significan mucho más que herramientas puramente instrumentales. Son, de hecho, artefactos técnico-culturales que alteran sustancialmente los procesos de toma de decisiones, con efectos en diversas ramas de la actividad humana, desde los parámetros de consumo hasta las relaciones de poder entre diversos actores (ya sea entre el Estado y los ciudadanos; las empresas y los consumidores; las organizaciones y los individuos).

Para hacer frente a estas transformaciones, los gobiernos y organizaciones de diversas jurisdicciones a nivel local, nacional y multilateral (como la ciudad de Nueva York, la Federación Alemana, la Unión Europea, las Naciones Unidas y la OECD) están desarrollando planes estratégicos, legislación o políticas públicas destinadas a abordar las tensiones derivadas de esta situación, así como acoger dichas innovaciones para garantizar sus posibles efectos

[1] Profesor de la Facultad de Comunicación (FAC) y del Programa de Posgrado en Comunicación de la Universidad de Brasilia (UnB). Doctora en Comunicación y Cultura Contemporánea graduada en la Universidad Federal de Bahía, con una pasantía doctoral en la Universidad de Washington (EE. UU.). Fue investigador visitante en el Instituto de Investigación Económica Aplicada (IPEA); consultor ad hoc en la Unesco para la aplicación de indicadores de desarrollo de medios en Brasil. Es fundador y coordinador del grupo de investigación Centro de Estudios en Comunicación, Tecnología y Política (CTPol) e investigador en el Instituto Nacional de Ciencia y Tecnología en Democracia Digital (INCT-DD).

positivos.

Paralelamente a los beneficios prácticos que proporcionan los sistemas algorítmicos más avanzados, esto también tiende a generar nuevas formas de desigualdad, violaciones de derechos o expansión de la concentración de poder. La proliferación de sistemas autónomos tiene repercusiones en puntos políticamente sensibles como la privacidad; libertades; derechos individuales y colectivos; desinformación; transgresiones éticas; autoritarismo, etc.

En vista de este contexto, este trabajo tiene como objetivo identificar y caracterizar los principales problemas clave a los que cualquier política de Inteligencia Artificial debe responder. En este sentido, el artículo presenta un estudio exploratorio, basado en la investigación documental y bibliográfica, analizada bajo la lente normativa de los principios democráticos. Para ello, el estudio se divide en dos partes: en primer lugar, hace un enfoque conceptual sobre la Inteligencia Artificial (IA), sus orígenes y características fundamentales. La segunda parte resume siete problemas clave importantes de énfasis político que están en la base del actual debate regulatorio sobre la IA en el mundo y que son debates decisivos para entender la compleja relación entre la democracia y los sistemas autónomos digitales.

2 Inteligencia artificial: técnica más allá de la técnica

La expresión "Inteligencia Artificial" nos lleva a imaginar máquinas de pensamiento o artefactos técnicos autoconscientes. Sin embargo, como ocurre en cada metáfora, es una terminología generalista que nos ayuda cognitivamente en una síntesis del fenómeno, pero minimiza aspectos importantes, generando una definición carente de mayor precisión. En cualquier caso, no debemos considerar el uso de esta terminología como un problema ya que ya está ampliamente difundida en el imaginario social, en documentos gubernamentales, noticias, directrices de la empresa, etc. Es posible adoptarla siempre y cuando podamos contextualizar y dimensionar esta metáfora y resaltar, sobre todo, los elementos que esconde la expresión.

Específicamente, Inteligencia Artificial hace referencia a un conjunto de métodos lógicos que tienen como objetivo resolver problemas basados en algoritmos entrenados (a través de **inputs**, entrada de datos) para comprender patrones, aprender de errores y volver a configurar los resultados (**output**) cada vez más cerca de lo esperado. Por lo tanto, es importante tener en cuenta que no estamos hablando de una máquina que piensa pero que resuelve problemas lógicos y está entrenada en este sentido a partir de la experiencia (datos) que recibe.

Desde un punto de vista histórico, los artefactos técnicos que ayudaron en la realización de alguna operación lógica, especialmente matemáticas, no son nuevos. En varias culturas, como Mesopotamia y China, instrumentos como el ábaco han existido desde la Antigüedad. En la Modernidad estos mecanismos han ganado una nueva versión con las primeras calculadoras. Como Gleick se sitúa (2011, p.99):

> Blaise Pascal creó una máquina adicional en 1642, con una fila de discos giratorios, uno para cada dígito decimal. Tres décadas más tarde, Leibniz mejoró el trabajo de Pascal usando un tambor con dientes sobresalientes para "reagrupar" las unidades de un dígito a otro.

Sin embargo, el autor recuerda que los prototipos Pascal y Leibniz permanecieron muy cerca del ábaco, porque hicieron registros pasivos de los estados de memoria de una operación matemática dada.

En el siglo XIX, en el contexto de la Revolución Industrial, Charles Babbage dio un paso más al insertar un elemento importante en las máquinas de cálculo: el automatismo. Esto estableció un precedente en el desarrollo de la computadora que se crearía realmente en el siglo siguiente. Sin embargo, la máquina de Babbage era básicamente mecánica (no utilizaba energía eléctrica) y no asumía una estructura lógica versátil, como la perspectiva binaria (basada en dos dígitos, 0 y 1) o las variables booleano (off/on). El terreno de la Inteligencia Artificial tal como la conocemos hoy en día empieza ser más fructífero en la segunda mitad del siglo XX. Más específicamente, sus orígenes están vinculados al término " machine intelligence" (inteligencia artificial) difundido por Alan Turing, con los primeros registros en manuscritos ya en 1941 (COPELAND, 2004)2. El principio básico de la idea de Turing se refería a la solución de problemas lógicos y matemáticos a través de la automatización en sistemas electrónicos binarios y la posibilidad de construir máquinas capaces de aprender de la experiencia. En un artículo de 1950 titulado " Computing machinery and intelligence " en la revista Psychology and Philosophy, Turing propone pensar en la siguiente pregunta: "¿Pueden pensar las máquinas? (Can machines think?") (Turing, 1950). Aunque el autor hace una larga discusión sobre las objeciones a esta pregunta, su preocupación es en realidad reformular tales cuestionamientos hacia lo que llamó el "Juego de Imitación". Para el autor, la capacidad de imitación universalizada de una máquina (basada en el lenguaje binario) sería uno de los elementos diferenciales que merecían una atención especial:

> This special property of digital computers, that they can mimic any discrete state machine, is described by saying that they are universal machines. The existence of machines with this property has the important consequence that, considerations of speed apart, it is unnecessary to design various new machines to do various computing processes (TURING, 1950, p. 441).

En 1943 Warren McCulloch y Walter Pitts publicaron un artículo titulado "A logical calculus of the ideas immanet in nervous activity" explicando cómo funcionan las neuronas y modeló una red neuronal artificial simple utilizando circuitos eléctricos para demostrar sus

[2] Aunque no habló directamente de inteligencia artificial, en 1945, Vannevar Bush publicó un artículo titulado " As we may think " en la revista The Atlantic Monthly. Propuso una máquina de memoria colectiva que llamó Memex capaz de aglutinar y procesar información convirtiéndola en conocimiento. Una reproducción de este texto está disponible en: https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/ Acceso 04 jul 2019.

hipótesis (MCCULLOCH; PITTS, 1943)[3]. Añadiendo esto a las ideas de Turing, varios investigadores fueron alentados a pensar en redes neuronales artificiales en computadoras y cómo esto podría estar alineado con la concepción de una "máquina inteligente". En 1956, en el estado de New Hampshire (EE.UU.) el término "Inteligencia Artificial" tal como lo conocemos hoy en día apareció por primera vez en una conferencia titulada "The Dartmouth Summer Research Project on Artificial Intelligence", considerado por muchos como la piedra fundadora de este campo de la investigación.

Pero si la noción de Inteligencia Artificial ya existía hace al menos medio siglo, ¿por qué sólo ahora estamos hablando de leyes y regulaciones para este campo como si fuera algo recién descubierto? La explicación es relativamente simple: porque un conjunto de condiciones técnicas que no se dieron anteriormente llegaron a coexistir y converger principalmente desde las primeras décadas de este siglo[4]. En este escenario, se estima que habrá un boom del uso de IA en las próximas dos o tres décadas (CATH, 2017; DAFOE, 2018; GRACE et al, 2018).

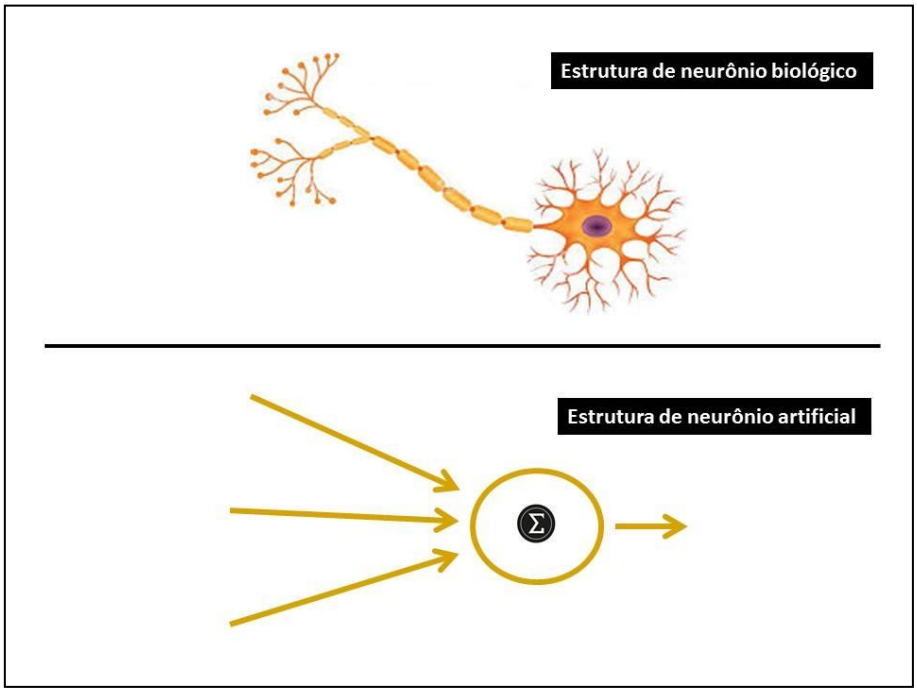
Para vislumbrar mejor la naturaleza de este fenómeno, debemos sintetizar cuatro aspectos que merecen una atención especial porque representan dimensiones que nos ayudan a entender el funcionamiento de los sistemas basados en la Inteligencia Artificial: (a) redes neuronales artificiales; (b) el significado de la noción de imitación; (c) el poder del automatismo y d) los niveles tipológicos de IA.

En primer lugar, para que un sistema se llame "inteligente", esto supone que es capaz de aprender y tomar decisiones basadas en la lógica. Un método innovador en este sentido es la idea de redes neuronales artificiales (también una metáfora vinculada al cerebro) que se ha convertido en una de las concepciones más prometedoras e influyentes de la IA, en la que se basan las técnicas de aprendizaje automático y de *machine learning* y *deep learning*. Más didácticamente, una red neuronal artificial es una composición de algoritmos inspirados en la estructura y el modo de funcionamiento de una neurona biológica, como se ilustra en la Figura 1:

[3] Otras obras también contribuyeron a esta línea: en 1949 el libro " *The organization of behavior: a neuropsychological theory*", de Donald Hebb; y en 1958 el artículo de Frank Rosenblat titulado "*El perceptrón: un modelo probabilístico para el almacenamiento de información y la organización en el cerebro*".

[4] Cómo y creación de infraestructuras digitales avanzadas, especialmente 5G; intensificación de mecanismos de big data que reflejan la capacidad exponencial de recopilar y procesar grandes volúmenes de datos de diversas fuentes con velocidad anteriormente inexistente; desarrollo de algoritmos más sofisticados, etc.

Figura 1 - Ilustración comparativa de estructuras de neuronas biológicas y artificiales.

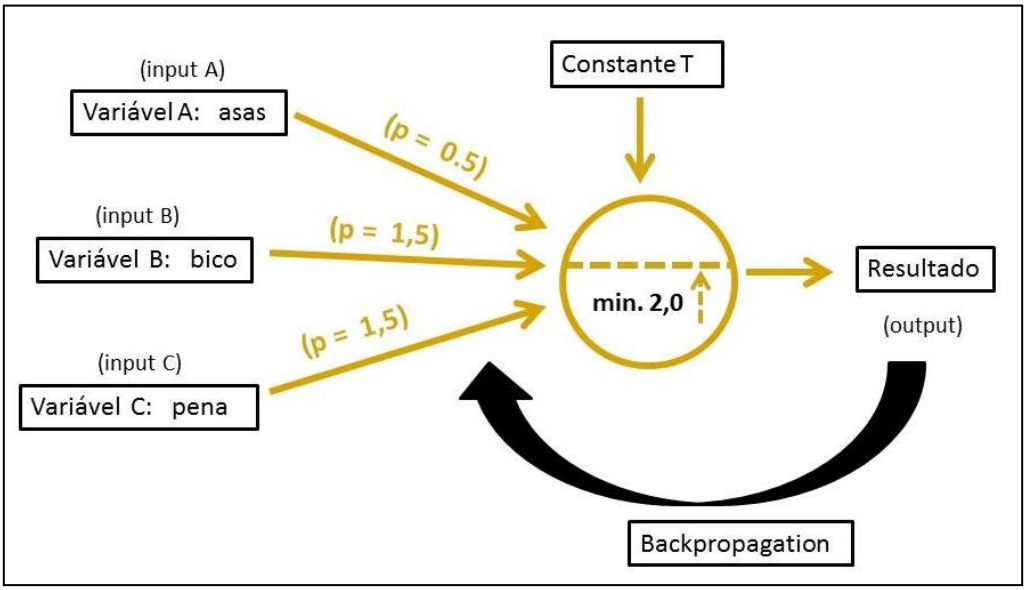


Fuente: elaboración propia de la propuesta de McCulloch y Water Pitts (1943)

Básicamente, una neurona artificial consta de varias entradas (*inputs*) diseñadas para capturar información de las diversas variables que definen lo que se está evaluando (por ejemplo, tomando las variables "pico" o "alas" como indicadores relevantes para definir si una imagen se refiere o no a un pájaro). Para cada entrada (variables capturadas) se da un peso. Los valores cuantitativos de cada variable con sus respectivos pesos se suturan en un único valor que luego se somete a un núcleo en el que se requiere alcanzar una cantidad mínima para que se active la neurona (por ejemplo, un desencadenador que sólo se activará si la suma alcanza un valor determinado considerado significativo)[5]. Por lo tanto, para que la neurona responda positivamente que la imagen en cuestión es de un ave, debe haber una combinación de las variables y sus respectivos pesos con valores que alcancen un cierto nivel capaz de indicar que existe una alta probabilidad de que la imagen sea, de hecho, un ave. La Figura 2 proporciona una ilustración simplificada de este proceso y sus pesos[6], tomando como ejemplo sólo 1 neurona artificial que se basa en 3 variables (alas, plumas y pico) para determinar si una imagen dada es de un pájaro.

[5] Esta es la estructura básica de una neurona artificial. Con el avance de los estudios en este campo, se insertaron otros elementos más complejos para mejorar la agudeza a las respuestas.
[6] Los valores de peso de la Figura 1 son solo estipulaciones ilustrativas. En la práctica, las ponderaciones se estipulan en el proceso de programación teniendo en cuenta la probabilidad de variables.

Figura 1 - Ilustración simplificada de la clave mecanismos operativos de una neurona artificial



Fuente: elaboración propia, síntesis y adaptación, con fines didácticos, la estructura inicial propuesta por McCulloch y Water Pitts (1943)

Si la neurona detecta como positiva sólo la variable A (alas) y las otras son negativas (lo que resultaría en valor cero, porque no están confirmadas) la suma en el núcleo será de sólo 0,5 puntos (valor de peso de la variable "ala"), por lo tanto no alcanzando el nivel mínimo estipulado en esta neurona (2 puntos) para considerar la imagen como la de un ave. Esto se debe a que tener sólo alas no es suficiente para determinar que algo es, de hecho, un ave porque hay otros objetos que tienen alas y que no son aves, como un avión, un murciélago o una mariposa. En otra situación, si la neurona detecta como positiva la variable A (alas) y B (pico), esto ya sumaría 2 puntos (0,5 de la variable A + 1,5 de la variable B), alcanzando así la puntuación mínima requerida. Teniendo en cuenta esto, la neurona se activaría para dar como resultado la respuesta positiva que dicha imagen es un pájaro. Eso es porque si algo tiene alas y pico aumenta estadísticamente la posibilidad de ser un pájaro. Sin embargo, el proceso no termina aquí. Lo que el algoritmo ha logrado hasta ahora es una "apuesta". Lo que realmente hace que la máquina "aprenda" son las correcciones de peso después de comprobar si la apuesta ha demostrado ser "falsa" o "verdadera".

La respuesta de la neurona se compara con la realidad para verificar si hubo un error o se acertó en la estimación. Esta información (error o correcto) genera una oleada de modificaciones o refuerzos de los pesos llamados "*backpropagation*": si una neurona (con ciertos ajustes de peso) se equivoca en su estimación de la imagen porque detectó que en realidad era la imagen de un avión, este error cuando se percibe se convertirá en una retropropagación que disminuirá los pesos de estas variables porque demostraron que no eran eficaces. Lo mismo se hace para las variables que inicialmente fueron juzgadas como no

importantes, pero en los resultados demostraron ser determinantes: esta vez, la retropropagación actúa para calibrar positivamente estas variables, haciendo hincapié en ellas como relevantes. La diferencia en el valor del resultado final se reprocesa, en este caso aumentando los pesos de las variables subestimadas haciendo que la próxima vez que la neurona funcione correctamente con los pesos debidamente calibrados para golpear, basado en experiencias anteriores.

Básicamente, la sofisticación del aprendizaje automático no es más que una corrección de los pesos inicialmente estimados a partir del hallazgo de errores aumentando así la posibilidad de hacerlo bien la próxima vez que se encuentre con las mismas variables. Sin embargo, para que el sistema funcione eficazmente, debe haber muchas neuronas artificiales interconectadas y, principalmente, una gran cantidad de datos para que las neuronas puedan "probar" los pesos de las variables, es decir, para que las redes neuronales sean "entrenadas". La llamada "fase de entrenamiento" de los algoritmos de IA requiere una gran cantidad de información y, como la percibimos, la disponibilidad de datos anteriores y la ocurrencia estadística es un elemento determinante. Por ejemplo, un software de IA sólo será capaz de distinguir entre "pájaros" y "aviones" cuando ha recibido muchas imágenes de entrada hasta que pueda hacer la distinción de forma automatizada. Por supuesto, el ejemplo que se da aquí es sólo didáctico. Las redes neuronales reales incluyen cientos o miles de neuronas con cientos o miles de variables que se están probando y recalibrando.

Un segundo elemento que es fundamental en el funcionamiento de los sistemas de Inteligencia Artificial, como se ha mencionado, es la perspectiva de la imitación. Con técnicas de aprendizaje automático, los algoritmos ahora tienen una inmensa capacidad para identificar (y repetir) patrones, haciendo que el juego de Turing valga la pena (Turing, 1950). No significa que el algoritmo sea consciente de la diferencia entre un pájaro y un avión, pero imita nuestra percepción de las cosas dando pesos a ciertas variables que se prueban y definen como determinantes, al igual que nosotros al mirar los elementos que componen nuestra definición de "pájaro". En este sentido, la imitación se basa en probabilidades estadísticas donde lo que es mayoritaria (después del proceso de retropropagación) es percibido por los algoritmos y reforzado por ellos. Por otro lado, lo que huye del patrón preponderante tiende a ser pasado por alto e ignorado o se convierte en una información que desentona en el sistema.

Por ejemplo, si un pájaro aparece sin pico y sin alas, el algoritmo tendrá dificultad para considerarlo un pájaro, ya que escapa al patrón que la fase de entrenamiento ha sedimentado en el código sobre lo que es un pájaro. Por lo tanto, las aves con discapacidades congénitas o lesionadas tienden a no ser reconocidas como aves. Tenga en cuenta que la fase de entrenamiento es dinámica, pero después de este paso, los sistemas tienden a ser estables (relativamente rígidos) basados en una perspectiva mayoritaria. Aquí podemos notar que la propiedad de imitación se basa en algo que ya se da, es decir, imitar es conservar y reforzar algo preexistente. En este sentido, es posible afirmar que paradójicamente hay una mezcla de

innovación y conservadurismo en la sofisticación de los algoritmos de aprendizaje.

La tercera característica que debemos tener en cuenta en el diseño de la Inteligencia Artificial es el automatismo. Dos siglos más tarde, el sueño de Babbage de buscar la máquina automática se llevó a cabo hasta el extremo. El automatismo es un elemento central en cualquier sistema de IA porque las máquinas solo se consideran inteligentes si son capaces de operar por su cuenta desde un start inicial y encontrar su propio camino. La evolución en la producción y almacenamiento de energía junto con la capacidad de los algoritmos para generar "looping" o ciclos de repetición ad infinitum es una combinación importante que en última instancia implica una nueva forma de potencia. Un ordenador o sistema puede funcionar repetitivamente durante años y siglos, siempre y cuando tenga poder. En un mundo con una vida cotidiana cada vez más datificado, tenemos, en la práctica, el aumento del poder de ciertos actores (como el Estado, las instituciones y las corporaciones) debido a la capacidad de imponer la repetición de procedimientos o formas de comportamiento. El autoritarismo o las acciones injustas pueden repetirse de una manera mucho más ágil, a un bajo costo y mucho más difícil de contrarrestar cuando son ejecutadas por sistemas autónomos. Esto también puede traer una mayor rigidez en la relación entre las partes asimétricas donde el sistema tiende a seguir procedimientos y no observar situaciones de excepción. También coloca a la máquina como una toma de decisiones que, detrás de una decisión aparentemente técnica, es el valor incrustado en las métricas.

Por último, un cuarto aspecto básico importante para entender la Inteligencia Artificial es darse cuenta de que hay diferentes grados de desarrollo de estas tecnologías y esto afecta en varias dimensiones sobre el lugar de los artefactos. La literatura ha señalado tres niveles o tipos de IA, como sintetiza Girasa (2020): Inteligencia Artificial Estrecha, Inteligencia Artificial General y Superinteligencia Artificial. El primero se refiere al rendimiento de una tarea singular. El segundo puede realizar varias tareas al mismo tiempo de una manera similar al cerebro humano. El tercero es la superación de la capacidad humana en varios aspectos. Actualmente, estamos en la primera etapa, pero ya con horizontes prometedores y prototipos de sistemas de segundo nivel. En relación con el tercer grado, esto sólo será posible con la creación de estructuras de procesamiento más sólidas de las que tenemos actualmente, como la computación cuántica, todavía en una etapa muy temprana de desarrollo. Lo que es importante tener en cuenta en estos grados es precisamente la sofisticación, la capacidad de rupturas y el alcance del campo de acción que representan. Cuanto mayor sea el grado de desarrollo de la IA, más intensa será su pervasividad y poder, tendiendo a ser mucho más impactante cultural, social y políticamente mucho más disruptivo.

Todas estas cuestiones que caracterizan el funcionamiento de la Inteligencia Artificial deben ser pensadas a partir de parámetros que pueden ir más allá del horizonte de la eficiencia técnica. Hay cuestiones sociales, culturales, políticas y económicas involucradas. Una buena metáfora de esto es la imagen que el filósofo alemán Martin Heidegger describió cuando analizó

la esencia de la tecnología moderna basada en el concepto Gestell (HEIDEGGER, 2001). Para el autor, en el pasado, construimos puentes que eran dispositivos técnicos instalados en los ríos. Con el uso de la tecnociencia en su búsqueda incesante para extraer, manipular y almacenar energía, la planta hidroeléctrica no es un elemento que se instala en el río, como lo era el puente. Para él, la situación se ha invertido: ahora el río está instalado en la planta (porque lo somete a sus objetivos) y el río se ha convertido así en un dispositivo del sistema tecnológico. Llevando esto a los avances en los sistemas de IA, estamos hablando directamente de problemas de albedrío y autonomía humana.

Dicho esto, dada la expansión de estos sistemas y su creciente centralidad en la vida cotidiana, es necesario desarrollar estrategias para que el Estado pueda promover y estimular todos los beneficios de la IA y, al mismo tiempo, mitigar sus posibles distorsiones, definiendo roles para que los diversos actores involucrados puedan interactuar armoniosamente, garantizando la protección de los derechos; evitando la pérdida de autonomía y libertades.

3 Inteligencia artificial, regulación y democracia: siete problemas clave

Entre 2017 y 2019 varios países, en todos los continentes, lanzaron sus estrategias para la Inteligencia Artificial: Alemania, Canadá, China, Dinamarca, Emiratos Árabes Unidos, Estados Unidos, Finlandia, Francia, India, Italia, Japón, Malasia, México, Nueva Zelanda, Kenia, Singapur, Corea del Sur, Suecia, Taiwán, Reino Unido. Otros países que aún no han lanzado una estrategia oficial (como Australia, España, Polonia y Uruguay[7]), estaban creando comités, consultas públicas o diseñando un presupuesto específico para el desarrollo de este ámbito. Los bloques regionales o las articulaciones panregionales también se han preocupado por el tema, generación de documentos como Declaration on AI in the Nordic-Baltic Region [8] (publicado por un conjunto de países nórdicos y bálticos) o mediante la creación de organismos como High-Level Expert Group on Artificial Intelligence (AI HLEG)[9] de la Unión Europea. Además, las organizaciones multilaterales tradicionales como la ONU[10] y la OECD[11] tienen acciones, directrices o recomendaciones sobre el tema. Organizaciones profesionales como el Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) y articulaciones civiles como la

[7] En el caso brasileño, todavía no hay una estrategia definida. El país tiene una Estrategia Digital (E-digital) lanzada en 2018 que contiene directrices genéricas para la transformación digital, pero no había presentado hasta el primer semestre de 2020 un documento más específico o una directriz oficial para la IA.

[8] https://www.regeringen.se/49a602/globalassets/regeringen/dokument/naringsdepartementet/20180514_nmr_deklaration-slutlig-webb.pdf Consultado 06 jul 2019.

[9] https://ec.europa.eu/digital-single-market/en/artificial-intelligence

[10] La ONU tiene algunas iniciativas sobre IA, una de las cuales es la plataforma "AI for Good https://aiforgood.itu.int/ sobre la base de las reuniones anuales sobre el tema (AI for Good Global Summit), celebradas por la Unión Internacional de Telecomunicaciones (UIT). También hay otras iniciativas como el Centro de Inteligencia Artificial y Robótica http://www.unicri.it/in_focus/on/UNICRI_Centre_Artificial_Robotics (vinculado a UNICRI), además de los documentos sobre el tema: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf Consultado el 08 Ene 2020.

[11] https://www.oecd.org/going-digital/ai/ Consultado 11 jul 2019.

Declaración de Montreal12 son también iniciativas relacionadas con el establecimiento de principios para las buenas prácticas de IA.

Este escenario demuestra que hay una clara efervescencia en el desarrollo de directrices y estrategias para la IA en todo el mundo, pero las acciones regulatorias más concretas, en forma de leyes, todavía están en proceso de ser procesadas o son incipientes, localizadas o parciales (como es el caso de la Ciudad de Nueva York que creó una de las primeras leyes en el mundo sobre el tema y Alemania que aprobó una ley que trata de una parte del problema, regulando específicamente el uso de vehículos autónomos).

Aunque las estrategias y políticas incipientes de la IA tienen sus peculiaridades y énfasis (algunos con más enfoque en los aspectos económicos, otros en cuestiones éticas), al analizar el conjunto de estrategias o proto-regulaciones es posible identificar siete problemas clave de antecedentes políticos que tienen implicaciones directas para las democracias contemporáneas que merecen una atención especial, debido a sus posibles consecuencias e impactos políticos. Podemos sintetizarlos en los siguientes términos: (1) realización e imputabilidad de la máquina; 2) dilemas y juicios morales; 3) autoritarismo estadístico de las métricas; (4) oscurantismo matemático en los procesos; 5) la pervasividad utilitaria de los sistemas autónomos; 6) control y límites de eficiencia; (7) diversidad y representatividad en la codificación.

4 Sobre la realización e imputabilidad de la máquina

En la ciencia ficción principalmente del tipo distópico, el problema de la realización de los sistemas de Inteligencia Artificial es un elemento recurrente y a menudo central en la trama que se desarrolla. En la década de 2000 la serie de televisión *Battlestar Galactica* dramatiza la dificultad de distinguir entre humanos y seres artificiales. En el mundo real, esta indistinción todavía no es tan evidente y tan avanzada como en la ficción, pero esto ya es un problema que está empezando a surgir en el horizonte regulatorio de los sistemas de IA. Sobre todo porque cuantos más algoritmos de IA se entrenan (con datos de los próximos años y décadas), más probable será un interlocutor humano, interactuando con el público y tendiendo a ser cada vez más difícil de percibir como una máquina.

Por lo tanto, una primera pregunta que surge con la realización de estos sistemas es el derecho del individuo a saber si realmente está hablando con otro ser humano o si consiste en un sistema de IA. Esto se debe a que el contrato de comunicación que se establece es bastante diferente cuando hablamos con máquinas y no con subjetividades humanas directamente. Esta diferenciación se vuelve importante en el debate regulatorio, ya que requiere reglas que obliguen a la máquina a declararse como una máquina. Por ejemplo, cuando una empresa o un

[12] https://www.montrealdeclaration-responsibleai.com/the-declaration Consultado 15 jul 2019.

organismo gubernamental adopta un sistema de IA para el servicio público, es necesario que la conversación comience con la identificación del tipo de interlocutor que está al otro lado de la línea o que alguna información en este sentido sea clara y evidente.

La cuestión de la realización de la máquina también tiende a pasar por la misma dinámica del sistema político. Un problema que ahora surge en las campañas electorales son los *chatbots*: algoritmos que simulan la conversación en lenguaje natural, fabricados y encargados (a escala industrial) para actuar masivamente en campañas negativas o en compromisos discursivos para un candidato. Con la evolución de los sistemas de IA, el fenómeno de los *chatbots* tiende a ser cada vez más común y sofisticado, se extiende a varias áreas ya sea en forma de sistemas de servicio al cliente, consultorías legales o incluso servicios de conversación para el apoyo emocional. La pregunta es, ¿cómo regular esta nueva "entidad" que se presenta entre nosotros y cuáles son los límites para su uso sin que esto implique transgresiones de derechos?

Esto también plantea hipótesis sobre la personalidad jurídica de los sistemas de IA, por ejemplo, para algunos analistas, debido a la complejidad de estos sistemas, podría ser apropiado que un robot tenga una personalidad similar a una empresa que está clasificada como una "entidad legal" en comparación con el "individuo", es decir, al individuo natural. Esta perspectiva aparece de diferentes maneras en diversas estrategias gubernamentales. Un buen ejemplo es una resolución del Parlamento Europeo de febrero de 2017 titulada "*Disposições de Direito Civil sobre Robótica*", que trajo una serie de recomendaciones a la Comisión Europea para pensar en el lugar de los robots en una sociedad cada vez más permeada por los autónomos basados en la IA. En el ítem 59, letra "f", los parlamentarios recomiendan:

> Crear un estatus legal específico para los robots a largo plazo para que al menos los robots autónomos más sofisticados puedan ser determinados como titulares de la condición de personas electrónicas encargadas de remediar cualquier daño que puedan causar y, en su caso, aplicar personalidad electrónica a los casos en que los robots tomen decisiones autónomas o de cualquier otra manera interactúen con terceros de forma independiente[13]

La resolución generó controversia y un grupo de expertos incluso publicó una carta abierta a la Comisión Europea criticando la incitación a ignorar dicha proposición[14]. Pero esto sigue siendo un debate abierto, porque debido a la complejidad de los actores involucrados y al creciente grado de realización de los sistemas de IA –que rompen con la noción de autoría una vez que el propio autómata se transforma aprendiendo de su experiencia en el mundo– los analistas explican que:

[13] Transcripción del extracto en portugués, por lo tanto, algunas ortografías están en formatos adoptados en Portugal. Disponible en http://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_PT.html >
[14] Se puede tener acceso a la letra abierta en https://g8fip1kplyr33r3krz5b97d1-wpengine.netdna-ssl.com/wp-content/uploads/2018/04/RoboticsOpenLetter.pdf Consultado 22 Ago 2019.

> [...] surge un conflicto legal, ya que en el marco de la legislación vigente el robot no puede ser considerado responsable de las acciones y (o) la inacción y como resultado la responsabilidad recae en el usuario, desarrollador de software o fabricante. Al mismo tiempo, la resolución de la UE plantea la cuestión de la responsabilidad en caso de que el robot causara daños debido a las decisiones tomadas por el propio robot (basado en los algoritmos integrados) y la definición del tercero responsable del pago de la compensación es imposible (ATABEKOV; YASTREBOV, 2018 p. 779).

El mismo código original puede asumir diferentes "personalidades" y actuar de manera diferente si el algoritmo se entrena a partir de datos de población religiosa o si se entrena con datos de población ateo. El argumento utilizado es que no se puede culpar a los autores de códigos si un sistema de IA se volvió fascista después de ser apropiado y entrenado por grupos fascistas. Sin embargo, es necesario crear mecanismos regulatorios que obliguen a los programadores a desarrollar soluciones técnicas que puedan inhibir que un sistema de IA se convierta en algo dañino en manos de otros. Es necesario establecer grados de responsabilidad de cada agente en la compleja cadena de producción y consumo que recorre este tipo de herramienta.

5 Sobre dilemas y juicios morales

Uno de los temas más importantes en las teorías de la democracia son los dilemas morales. Por otro lado, la dimensión moral también siempre ha estado presente en la concepción más filosófica o ficticia de la Inteligencia Artificial, cuya afirmación más conocida son los principios idealizados por el escritor Isaac Asimov (conocidos como tres leyes de la robótica). Los juicios morales son indicadores pertinentes para analizar cómo las democracias tratan las divergencias, respetando las diferencias y manteniendo un procedimiento racional y justo en los procesos de toma de decisiones; y cuál es el papel de las instituciones (o el Estado) para tratar deliberadamente conflictos complejos que involucran desacuerdos morales básicos (GUTMANN; THOMPSON, 1996).

Los juicios morales reflejan las tensiones sociales y el posible desequilibrio en el equilibrio de poder entre grupos de cosmovisión divergentes. Al mismo tiempo, los artefactos técnicos también nos ponen en la cara de los problemas morales. El dilema del tranvía ("*trolley problem*"), un experimento crítico sobre ética ideado por Philippa Foot, es un buen ejemplo de ello. Brevemente, la pregunta trae una situación hipotética en la que hay un tren en curso de colisión que llegará a 5 personas que están en las vías. Sin embargo, un observador en una posición privilegiada se da cuenta del futuro accidente y tiene en sus manos la posibilidad de tirar de una palanca y desviar el vehículo a otro camino alternativo. Esto salvará la vida de cinco personas, pero la desviación resultará en la muerte de otra persona que está en el camino. Ante esta situación hipotética, la pregunta es: ¿sería ético sacrificar la vida de una persona para salvar la vida de cinco? Generalmente, la mayoría de la gente cree que la respuesta es positiva. Por

otro lado, el dilema se vuelve más complejo con una pequeña variación de la historia: cuando el mismo tren continúa en el curso de colisión de cinco personas, pero en esta segunda versión el observador está en un puente por el que el tren pasará y su lado es un hombre que si fuera lanzado hacia la línea del tren su cuerpo frenaría y detendría el tren antes de llegar a la gente en el carril salvando así las cinco vidas previamente amenazadas. Se hace la misma pregunta para esta segunda versión: ¿sería ético sacrificar la vida de una persona para salvar la vida de cinco? En esta nueva narrativa, el resultado final es el mismo: sacrificar a una persona para salvar a cinco. Sin embargo, las personas tienen mayor dificultad para tomar una posición en este caso porque no hay ningún elemento técnico intermedio como en la primera situación: una palanca que hace "menos personal" el acto de sacrificar una vida por razón de cinco.

En el contexto actual de la expansión de los sistemas digitales ubicuos y predictivos dispersos a lo largo de la vida cotidiana, el artefacto de Inteligencia Artificial ocupa el lugar de observador privilegiado del "problema del carro″ que concibe el futuro y tiene mecanismos técnicos para tomar decisiones sobre vidas. Dado el avance del Internet de las cosas y el uso diario de varios tipos de autónomos (como los coches sin conductor; *drones* que funcionan por piloto automático; robots, etc.) estos objetos tienden a configurarse como nuevos "actores" en la fauna de la dinámica social y, al mismo tiempo, significarán casos cada vez más importantes de toma de decisiones. Por ejemplo, un coche autónomo basado en IA toma decisiones sobre su viaje diario en la vía pública en todo momento (ya sea gira a la izquierda, se detiene en el semáforo en rojo o sigue recto en una avenida). Algunas decisiones son obvias y esperadas: no avance en la luz roja, por ejemplo.

Sin embargo, otras decisiones tendrán un mayor grado de complejidad y serían difíciles incluso para un ser humano porque son decisiones morales. Supongamos que hay un coche autónomo con un par de pasajeros jóvenes que son transportados en una carretera. Inesperadamente, el camión delante de este vehículo se detiene lo suficientemente fuerte como para que el freno del vehículo detrás no pueda responder al punto de evitar la colisión. El accidente sería fatal para el par de jóvenes pasajeros y habría una pérdida total del vehículo. El sistema de IA del coche autónomo podrá anticipar el problema y podrá tomar una decisión sobre lo que hará para evitar el menor daño posible. El problema comienza cuando ponemos en esta ecuación otros elementos (similares al "*trolley problem*") que hacen de la toma de decisiones un posicionamiento moral que expresa una visión del mundo o valores particulares.

Supongamos que si el coche decide desviarse hacia la derecha, golpeará a un niño que viene de la escuela que seguramente morirá en el impacto, pero el daño al vehículo y a sus pasajeros será mínimo y sobrevivirán. Si decides desviarse hacia la izquierda, golpearás el coche con tres ancianos que están en la carretera opuesta y, debido al ángulo de la colisión, los ancianos sufrirán el mayor impacto y sin duda morirán mientras que el coche con la pareja joven tendrá datos más pequeños y estos sobrevivirán. Aquí tenemos una decisión moral que

debe ser tomada por el algoritmo. Si la empresa que construyó el vehículo diseñó el código para que la opción en estos casos sea el menor daño posible a la propiedad (el coche) y a su cliente (la pareja joven), esto podría significar la muerte de un niño o la muerte de tres personas mayores en lugar de la muerte de la joven pareja propietaria del vehículo. Si el criterio es el menor número posible de víctimas, el niño en la acera será sacrificado (y así preservar la vida de los dos jóvenes en el coche o los tres ancianos en sentido contrario).

Si el criterio es otro que considera la preservación de la vida de las personas más jóvenes (que tendrían una vida por delante) como una prioridad en relación con la vida de las personas mayores (que ya han tenido la oportunidad de una mayor experiencia de vida) el resultado será la muerte de los tres ancianos al otro lado del camino. Cualquiera que sea la decisión, será controvertido y caerá en un dilema moral. En teoría, dependiendo de la opinión del reglamento aplicado, la ley puede obligar a la empresa a elegir, en este caso, por los daños al coche y al pasajero, preservando vidas no relacionadas con el accidente. Pero las situaciones pueden ser aún más problemáticas: supongamos que en lugar de un coche con un par de jóvenes, la colisión ocurrió con un vehículo de transporte escolar con 12 niños? ¿Debe el algoritmo elegir la muerte de estos 12 niños que están directamente involucrados en el accidente en la carretera o debe sacrificar a ese niño que está en la acera (no relacionado con el evento en la carretera) o incluso a los tres ancianos en la dirección opuesta?

La única certeza en este caso es que los parámetros de esta decisión no pueden dejarse a la empresa que diseñó los códigos por sí sola. Al menos para 2016, la opción de empresas como Mercedes-benz es priorizar la vida del propietario en caso de dilemas morales[15]. Uno de los mayores retos es precisamente cómo regular estos nuevos órganos de toma de decisiones sin representar la preponderancia de intereses y valores específicos (como los intereses comerciales de la empresa que construyó el coche, o los intereses de un grupo que perjudica los derechos de terceros).

6 Sobre el autoritarismo estadístico de las métricas

La automatización de los sistemas de IA, así como su eficacia se basan en análisis estadísticos impulsados por un gran volumen de datos y guiados por métricas o criterios previamente establecidos por el programador. Las métricas preceden a las ponderaciones, es decir, antes de que el sistema funcione es necesario establecer cuáles son sus objetivos, dónde quiere llegar, qué es realmente importante. Como ilustra Bigonha (2018, p.6):

[15] Ver <https://www.tecmundo.com.br/mercedes/110591-sim-carro-autonomo-mercedes-atropelar-voce-salvar-condutor.htm=. Acceso 20 Oct 2016

> Considere, por ejemplo, un modelo que otorga crédito y predice la puntuación de una persona, dada la probabilidad de que paguen el préstamo. ¿Qué representa el éxito del sistema? ¿Más ganancias? ¿Más gente recibiendo préstamo? ¿Más pagadores?

En este caso, aquellos que tienen el poder de definir la métrica de algoritmo pueden ejercer nuevas formas de potencia codificada en sistemas digitales. Y esto puede ser bastante parcial, hasta el punto de generar discriminación, acentuar las desigualdades y coquetear con una nueva forma de autoritarismo configurado en las máquinas.

Las métricas también pueden ser autorizadas cuando se basan en información que, aunque tienen un valor estadístico relevante y revelan pronósticos realistas, no podrían utilizarse para una toma de decisiones determinada porque extrapolan su función o violan derechos. En los análisis de *big data*, es común conocer las razones de ciertas correlaciones entre variables, incluso si se sabe que existen y utilizar dicho conocimiento en los procesos de toma de decisiones (MAYER-SCHONBERGER; CUKIER, 2013). Esto hace que los datos de varios tipos y fuentes se crucen estadísticamente para que el sistema toma decisiones lógicas-estadísticas pero autorizadas. Por ejemplo:

> [...] es interesante mencionar la controversia en Alemania que involucra a SCHUFA (una empresa alemana que presta servicios de protección del crédito), que, en el contexto de la evaluación del riesgo del consumidor, clasificó como criterio negativo su solicitud de acceso a sus propios datos. Esto se debe a una correlación estadística que se estableció en el sentido de que los consumidores que accedieron más a su puntuación eran más propensos a ser morosos. La empresa sufrió numerosas críticas por esta conducta, que penalizaba a quienes querían contratar un crédito con una puntuación más baja, exclusivamente, debido al ejercicio de un derecho. (DONEDA et al, 2018, p. 6)

En este caso, la regulación debe imponer límites a los cruces estadísticos de los sistemas de IA para el uso de cierta información que no concierne al sujeto en cuestión o que puede subvencionar acciones punitivas por el mero hecho de que alguien ejerza un derecho pero haya generado datos que puedan volverse contra el propio individuo. El desafío es cómo establecer estos límites sin crear un obstáculo genérico con respecto al uso de estadísticas de correlación en el cruce de información y variables diversas.

Es necesario definir en qué situaciones esto es normal y, al mismo tiempo, en qué situaciones esto se convierte en un abuso o violación. Requisito de documentación que detalla claramente cómo se construyeron los códigos y qué métricas se utilizaron; desarrollo de códigos deontológicos que establezcan directrices y límites para empresas y programadores; aplicación regular de auditorías algorítmicas independientes, etc. son algunos elementos que el proceso regulatorio puede utilizar.

7 Sobre el oscurantismo matemático en los procesos

Una de las preguntas más recurrentes en estrategias y documentos sobre políticas o regulación de IA es el problema de la opacidad del algoritmo. Basándonos en la premisa de que los algoritmos son cada vez más omnipresentes en las más diversas áreas de la actividad humana, tener la vida cruzada (y a menudo determinada) por reglas y parámetros que no conocemos se convierte en un problema de autonomía y autodeterminación de los sujetos. Por lo tanto, en general, el requisito de cierto nivel de control, transparencia y *accountability* de las empresas que llevan proyectos de IA ha sido una nota clave presente en los documentos y estrategias para este campo.

En un estudio que abordó este problema de falta de transparencia, Burrell (2016) identificó tres tipos de opacidades que se producen alrededor de los algoritmos. (a) En primer lugar, **opacidad como secreto corporativo o estatal intencional**. Las empresas tienden a entender los códigos como activos comerciales y protegerlos como secretos industriales, manteniéndolos fuera del ojo público bajo la reclamación de *copyright*. Los gobiernos también clasifican ciertos códigos como secretos de Estado o dificultan su publicación. (b) En segundo lugar, **opacidad como analfabetismo técnico**. Los algoritmos son código cuya transparencia y comprensión se limita a aquellos que tienen conocimiento del lenguaje para descifrarlos. Y sólo una pequeña porción de la población tiene esta "literatura". (b) Tercero, la **opacidad como característica estructural de los algoritmos de aprendizaje automático** (*machine learning*). Este último punto trata de un problema estructural directamente relacionado con la Inteligencia Artificial, porque el entrenamiento denso de las redes neuronales y sus calibraciones de peso hace que la reconstitución de todo el proceso de cómo el sistema llegó a una decisión dada algo difícil de visualizar porque cuanto más compleja es la estructura de las redes neuronales, más oscuro se vuelve el proceso (LEESE , 2014; MITTELSTADT et al 2016). Como señala Rendtorff-Smith (2018, p.11):

> Estos sistemas, especialmente cuando hablamos de sistemas de aprendizaje no supervisados, son por su naturaleza oscuros, lo que dificulta mantener los principios fundamentales de transparencia, explicación y rendición de cuentas. Otro desafío ha sido que los gobiernos a menudo adquieren y despliegan tecnología de IA patentada desarrollada por actores privados de la industria.

Para el autor, es importante que los sistemas de IA sean auditables, lo que permite a los investigadores identificar la fuente de un error o consecuencia adversa. Sin embargo, las auditorías tradicionales en el formato que conocemos hoy en día pueden ser ineficaces porque, a simple vista, a los auditores les cuesta descifrar el denso proceso generado por las redes neuronales artificiales.

Desde el punto de vista regulatorio, parte de la solución de este oscurantismo técnico es la inversión en investigación y herramientas de IA capaces de producir auditorías automatizadas adecuadas para hacer frente a las capas profundas de las redes neuronales artificiales. Para ello, es necesario desarrollar tecnologías específicamente orientadas a este fin, que puedan ser posibles por organizaciones u organismos específicos para estimular y fomentar el desarrollo de este campo. La creación de organismos reguladores específicos ha sido un elemento que aparece en este debate como un mecanismo necesario para este y otros problemas de la aplicación y la vigilancia con los sistemas de IA (SCHERER, 2016; TUTT, 2017).

8 Sobre la pervasividad utilitaria de los sistemas autónomos

Un cuarto problema que impregna las estrategias y políticas regulatorias para la Inteligencia Artificial, y merece atención, es el carácter omnipresente y utilitario inherente a estas tecnologías. Generalizado porque los sistemas de IA se insertarán, en un futuro no muy lejano, en todas las áreas de la experiencia humana, actuando sobre los más diversos tipos de servicios; incluyendo la mediación de nuestro uso de otros dispositivos y máquinas (como automóviles, aviones y hogares); tras relaciones sociales, políticas, culturales e incluso afectivas; llegar primero donde los humanos en carne y hueso no estaban (como en las misiones espaciales y la colonización de planetas). Utilitario porque hay una presunción *benthamiano* en el uso generalizado y la difusión de estas tecnologías que son capaces de involucrarnos ontológicamente, porque existe una profunda relación entre la eficiencia técnica, la comodidad y la buena vida. Veamos lo que Dijo Bentham en el siglo XVIII y pensemos en cómo se relaciona con los dispositivos de IA hoy en día:

> By utility is meant that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness, (all this in the present case comes to the same thing) or (what comes again to the same thing) to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered: if that party be the community in general, then the happiness of the community: if a particular individual, then the happiness of that individual (BENTHAM, p. 14-15, [1781], 2000)

Como se puede ver, la moral utilitaria está bastante alineada con las tendencias en el funcionamiento de los sistemas de Inteligencia Artificial. Sobre todo, porque el discurso de exaltación en las estrategias, productos y servicios en este campo tiene que ver con el horizonte de la mejora de la vida en un futuro próximo para el mayor número posible de personas y el mayoritarismo implícito en las métricas de las redes neuronales artificiales, como vimos anteriormente, refuerzan esta perspectiva. Por ejemplo, cuando un sistema de IA captura el deseo mayoritario de una población determinada que genera beneficios específicos adecuados para esta mayoría, aquellos que quedaron fuera de los estándares estadísticos tienden a ser considerados desviados, por lo que la exclusión generada por el artefacto se justificaría sobre la

base de una moralidad utilitarista que fuera capaz de generar felicidad para el mayor número posible de personas.

Al mismo tiempo, como señalan Helbing y sus colegas (2017), estas tecnologías serán cada vez más capaces de conectarse, entrar en la vida y seguir todo el camino de nuestras acciones y la tendencia es que "la programación informática avance a la programación de las personas" que gradualmente se vinculará y dependerá de estos sistemas, en parte por opción (porque consideramos más cómodo y eficiente), en parte porque existe una presión cultural y social, típica de las sociedades datificados:

> Es posible darse cuenta de que en un mundo cada vez más datificado con la abundancia de suministro y el uso masivo de aplicaciones lógicas para ayudar en las rutinas, no renunciar a los datos y decidir no alimentar *dataveillance* (o vigilancia digital) significa optar por una vida de exclusión de las diversas comodidades que ofrece el sistema. Implicará una vida más dura, con menos comodidad, cuya ejecución de rutinas se llevaría a cabo en más tiempo y con más energía para ser liberada. Al final, el gran problema de la privacidad hoy en día está ligado al sesgo práctico-lógico de la cultura digital, que en última instancia significa una mejor vida, es decir, con más comodidad y practicidad, incluso con menos autonomía (SILVA, 2019, p. 164).

Para empeorar las cosas, Nemitz (2018) recuerda que la omnipresencia de los sistemas de IA regirá la mayor parte de las funciones de una sociedad (salud, educación, seguridad, economía, justicia, etc.) que tienden a mantenerse en manos de pocas empresas que realmente concentrarán el *know-how* y la infraestructura necesaria. Y esto requiere un papel de equilibrio del Estado, con la observancia de los principios constitucionales, a través de leyes y mecanismos de equilibrio capaces de evitar el aumento de la concentración de poder y la exacerbación de las desigualdades existentes y el surgimiento de nuevas (NEMITZ, 2018).

Observando cómo las estrategias nacionales de IA han traducido esta relación entre la utilidad y la protección de los derechos, Cath y colegas (2018) advierten que hay un enfoque erróneo cuando se prioriza el aspecto de la utilidad y no hay un énfasis social en estas tecnologías. Los autores analizaron tres informes sobre estrategias y regulaciones de IA publicados en los Estados Unidos, el Reino Unido y la Unión Europea y encontraron que hay una característica común (que también es recurrente en los documentos de otros países): la ausencia de una política efectiva que piense en una "*good AI society*". Los documentos estudiados no son propositivos en este sentido y están más preocupados por cuestiones de aplicabilidad (como la economía, los negocios, etc.) y no con un plan eficaz para preparar a la sociedad para una sinergia positiva con estas innovaciones: "In short, we need a social strategy for AI, not mere tactics" (pág. 3).

9 Sobre el control y los límites de la eficiencia

Como vimos anteriormente, uno de los aspectos fundamentales de los sistemas de IA es la llamada fase de entrenamiento de algoritmos. Con algunas variaciones, por lo general hay dos formas fundamentales de entrenamiento: (a) aprendizaje supervisado y (b) no supervisado. La primera manera, implica dar al sistema datos que apunten exactamente a dónde quiere ir, mostrando el código los parámetros deseados. Tomando nuestro ejemplo inicial, sería como dar al algoritmo el objetivo de identificar si una imagen es de un pájaro, y para eso, el código aprende de un gran volumen de imágenes de aves de todo tipo para que el sistema entienda qué patrones de variables son más propensos a golpear para identificar aves. Es decir, entrenamos el algoritmo para un horizonte específico y ofrecemos esta información supervisada para esto. La forma de aprendizaje no supervisado es cuando no damos al código instrucciones más específicas y dejamos que el sistema consuma aleatoriamente una cantidad de información pudiendo identificar patrones o correlaciones ocultos o apenas perceptibles a simple vista.

En este modelo, el algoritmo puede tomar diferentes rutas y descubrir cosas que no estaban planificadas. Tomando nuestro ejemplo de vuelta, sería como dibujar un código que consume imágenes de todo tipo hasta que se convirtió en capaz de percibir patrones y diferenciar las imágenes de un pájaro, un avión, una mariposa. Especialmente en este último caso, la forma no supervisada, hay una gran preocupación por la pérdida de control del elemento humano en el sistema. Una vez que el código se genera y no se supervisa correctamente dependiendo del área en la que opera el algoritmo puede implicar resultados perjudiciales o acciones no previstas por sus desarrolladores.

Otra dimensión importante que implica ambos casos es el riesgo de pérdida de control, pero esta vez no en el resultado final, sino en el camino que la máquina eligió hacer en pos de su objetivo. Russell (2019) argumenta que uno de los principales problemas de la IA es precisamente el enfoque no medida en la eficiencia que, por lograr, puede significar un camino éticamente cuestionable:

> Sin embargo, a medida que las máquinas diseñadas de acuerdo con el modelo estándar se vuelven más inteligentes, y a medida que su ámbito de acción se vuelve más global, el enfoque se vuelve insostenible. Estas máquinas perseguirán su objetivo, no importa lo equivocado que esté; resistirán los intentos de apagarlos; y adquirirán todos y cada uno de los recursos que contribuyan a alcanzar el objetivo (RUSSELL, 2019, p. 171).

En este sentido, las directrices regulatorias y los principios éticos en el desarrollo del código han discutido la importancia de restringir este tipo de aplicaciones en áreas más sensibles (que afectan a la vida de las personas de manera más directa e impactante) y, al mismo tiempo, crear mecanismos de monitoreo más estrictos en procesos capaces de hacer que los sistemas de IA sean más *accountable*, incluyendo la posibilidad de apagado manual en el sistema en caso de desviación o pérdida de control.

10 Sobre diversidad y representatividad en códigos

Una última cuestión clave que ha cruzado los debates regulatorios y reglamentarios sobre Inteligencia Artificial se refiere a la dimensión inclusiva que requiere el uso masivo de algoritmos para diversas aplicaciones cotidianas. Específicamente, implica pensar en los valores que están incrustados (o se han olvidado) en el código. Esto nos remite a una peculiar forma de representatividad teniendo en cuenta que los algoritmos son representaciones de valores incrustados en la máquina que, aunque parezcan falsamente neutrales, son en realidad el resultado de una cosmovisión o subjetividades objetivadas:

> Los programadores pueden crear algoritmos que tengan suposiciones o limitaciones sesgadas incorporadas en ellos. Inconscientemente, pueden plantear un problema de forma sesgada. Los prejuicios de los programadores individuales pueden tener un efecto amplio y acumulativo porque, en un sistema de software complejo compuesto por subsistemas más pequeños, el sesgo real del sistema puede estar compuesto por reglas especificadas por diferentes programadores (CITRON, 2008, p. 1262).

Esto se aplica tanto a los sistemas de aprendizaje automático supervisados como a los no supervisados, ya que la elección misma de los datos que utilizarán los algoritmos en la fase de entrenamiento puede ser una opción sesgada o puede, los datos en sí o el entorno de entrenamiento, contener una serie de problemas de sesgo. Varios estudios han demostrado el efecto discriminatorio que los algoritmos pueden reforzar (GRAHAM, 2004; CPL, 2017; LEURS; SHEPHERD, 2017; WEST et al 2019) y esto ocurre tanto directa como deliberadamente, así como indirectamente, como reflejo del proceso de producción de *design* de los sistemas, es decir, a nivel de los equipos de los programadores:

> La discriminación y la inequidad en el lugar de trabajo tienen consecuencias materiales significativas, en particular para los grupos subrepresentados que están excluidos de los recursos y las oportunidades. Sólo por esta razón es necesario abordar urgentemente la crisis de diversidad en el sector de la IA. Pero en el caso de la IA, lo que está en juego es mayor: estos patrones de discriminación y exclusión reverberan mucho más allá del lugar de trabajo en el mundo en general. Los sistemas industriales de IA están desempeñando cada vez más un papel en nuestras instituciones sociales y políticas, incluso en la educación, la salud, la contratación y la justicia penal. Por lo tanto, debemos considerar la relación entre la crisis de diversidad del lugar de trabajo y los problemas con el sesgo y la discriminación en los sistemas de IA (WEST et al, 2019, p. 15)
> .

Por lo tanto, varias organizaciones y especialistas han estado prestando atención al problema de la representatividad en el proceso de construcción de los sistemas. Esto implica tanto la existencia de códigos deontológicos que los programadores deben seguir para incorporar una mayor diversidad y visiones del mundo –incluso los ausentes, que van más allá de su círculo social– como para romper con las tendencias de predominio étnico, cultural y racial de aquellos que dibujan código y construyen sistemas de IA que en última instancia serán un reflejo ampliado de los anhelos humanos para bien o para mal.

11 Conclusión

El objetivo principal de este artículo era sintetizar y caracterizar los principales problemas clave que una buena política de inteligencia artificial necesita responder, en el contexto de las democracias contemporáneas. Inicialmente, la preocupación era sintetizar la aparición y evolución de la Inteligencia Artificial como campo de estudios y desarrollo, así como resaltar los principales aspectos que nos ayudan a entender mejor su forma de funcionar. En este sentido, se discutió inicialmente el papel estadístico de las redes neuronales artificiales; la fuerza y los problemas detrás de la idea de "imitación"; el poder de repetición en los mecanismos de automatización y las tipologías o niveles de IA.

Luego, desde la perspectiva del auge de los documentos gubernamentales y no gubernamentales sobre estrategias y políticas para la IA publicadas principalmente entre 2017, se plantearon siete problemas clave importantes de énfasis político que están en la base de este debate: 1) realización y imputabilidad de la máquina; 2) dilemas y juicios morales; 3) autoritarismo estadístico de las métricas; (4) oscurantismo matemático en los procesos; 5) la pervasividad utilitaria de los sistemas autónomos; 6) control y límites de eficiencia; (7) diversidad y representatividad en los códigos.

Lejos de agotar todas las dimensiones que plantea este complejo fenómeno, la propuesta consistía en poner de relieve un conjunto de problemas que tienen implicaciones políticas y que merecen una atención especial porque están configurados como elementos determinantes para la buena relación entre la democracia y los sistemas de inteligencia artificial en pleno aumento. La idea es colaborar con una comprensión más amplia y política de este fenómeno, que es útil para fomentar la profundización del debate sobre estrategias y políticas regulatorias sobre IA en construcción o consolidación, fortaleciendo una perspectiva más humanista que vaya más allá de la eficiencia técnica. Sobre todo, se observaron los dilemas que presenta este escenario, la concentración de poder que se insinúa y el posible aumento de las asimetrías en las sociedades cada vez más datificadas. Todo este escenario requiere que el Estado asuma su papel propio de protector de los derechos individuales y colectivos, poniendo en práctica planes, políticas y regulación apropiada para este campo.

Referencias

ATABEKOV, A.; YASTREBOV, O. Legal Status of Artificial Intelligence Across Countries: Legislation on the Move. **European Research Studies Journal**, v. 21, n. 4, p. 773-782, 2018.

BENTHAM, Jeremy. **An Introduction to the Principles of Morals and Legislation**. Kitchener: Batoche Books, 2000.

BIGONHA, Carolina. Inteligência Artificial em perspectiva. **Panorama Setorial da Internet**, n.2. São Paulo: NIC.Br, 2018, p. 1-9.

BURRELL, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms. **Big Data & Society,** 2016, p. 1-12;

CATH, C. *et al.* Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. **Science and Engineering Ethics**, v. 24, n. 2, 2018, p. 505-528.

CITRON, Danielle Keats. Technological Due Process. **Washington University Law Review**, v. 85, n. 6, p. 1249-1313, 2008.

COPELAND, Jack. Artificial Intelligence. *In:* COPELAND, Jack (org).**The Essential Turing**: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life: Plus The Secrets of Enigma. Oxford: Oxford University Press, 2004, p. 353-261.

CPL. Centre for Public Impact. **Destination unknown:** Exploring the impact of Artificial Intelligence on Government - Working Paper.Londres: CPL. Centre for Public Impact, 2017. Disponível em: https://resources.centreforpublicimpact.org/production/2017/09/Destination-Unknown-AI-and-government.pdf. Acesso em: 15 jul. 2019.

DONEDA et al. Considerações iniciais sobre inteligência artificial, ética e autonomia pessoal. Pensar: **Revista de Ciências Jurídicas**, v. 23, n. 4, p. 1-17, 2018.

GIRASA, Rosario. **Artificial Intelligence as a Disruptive Technology**: Economic Transformation and Government Regulation. Cham: Palgrave Macmillan, 2020

GLEICK, James. **A informação**: Uma história, uma teoria, uma enxurrada. São Paulo: Compahia das Letras, 2011.

GRACE, Katja *et al*. When Will AI Exceed Human Performance? Evidence from AI Experts. **Journal of Artificial Intelligence Research**, n. 62, p. 729-754, 2018.

GRAHAM, Stephen. The woftware-sorted city: rething the "digital divide". *In:* GRAHAM, S. (org.). **The cybercities reader**. Londres: Routledge, 2004, p. 324-332.

GUTMANN, Amy; THOMPSON, Dennis. **Democracy and Disagreement**. London: Cambridge; Massachusetts: Harvard University Press, 1996.

HEBB, Donald O. **The organization of behavior**: a neuropsychological theory. Oxford: Wiley, 1949.

HEIDEGGER, Martin. **Ensaios e conferências**. Petrópolis: Vozes, 2001.

HELBING, Dirk et al. Will Democracy Survive Big Data and Artificial Intelligence? **Scientific American**, n. 98, p.73-98, 2017. Disponível em: www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence. Acesso em: 15 jul. 2019.
LEURS, Koen; SHEPHERD, Tamara. Datafication & Discrimination. *In:* SCHÄFER, Mirko Tobias; ES, Karin van (org). **The Datafied Society**: Studying Culture through Data. Amsterdam: Amsterdam University Press, 2017, p. 211-231.

McCULLOCH, Warren S.; PITTS, Walter H. A logical calculus of the ideas immanet in nervous activity. **Bulletin of Mathematical Biophysics**, n. 5, p. 115-133, 1943.

LEESE, Matthias. The new profiling: Algorithms, black boxes, and the failure of ant)i-discriminatory safeguards in the European Union. **Security Dialogue**, v. 45, n.5, p. 494-511, 2014.

MAYER-SCHONBERGER, Viktor; CUKIER, Kenneth. **Big Data:** como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Rio de Janeiro: Editora Campus, 2013.

MITTELSTADT, B. D. *et al*. The ethics of algorithms: Mapping the debate. **Big Data & Society**, v. 3, n. 2, p. 1-21, 2016

MONTRÉAL DECLARATION, For a **Responsible Development of Artificial Intelligence 2018**. Montreal: Université de Montréal, 2018. Disponível em: https://www.montrealdeclaration-responsibleai.com/the-declaration. Acesso em: 22 jul. 2019.

NEMITZ, Paul. Constitutional democracy and technology in the age of artificial intelligence. **Philosophical Transaction Real Society**, p. 1-14, 2018.

RENDTORFF-SMITH, Sara. Desafios de Governança em Inteligência Artificial. Entrevista. In: **Panorama setorial da Internet**, n.2, São Paulo: NIC.Br, 2018, p. 10-12.

ROSENBLAT, Frank. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychologial Review**, n. 65, p. 386-408, 1958.

RUSSELL, Stuart. **Human Compatible**: artificial intelligence and the problem of control, 2019.

SCHERER, Matthew U. Regulating artificial intelligence systems: reisks, challenges, competencies, and strategie. **Harvard Journal of Law & Technology**, v. 29, n. 2, p. 354-400, 2016.

SILVA, Sivaldo P. da. Comunicação digital, Economia de Dados e a racionalização do tempo: algoritmos, mercado e controle na era dos bits. **Revista Contracampo**, v. 38, n. 1, p. 157-169, 2019.

TURING, A.M. Computing machinery and intelligence**. Mind**: a Quarterly Review of Psychology and Philosophy, v.49, n. 236, p. 433-460, 1950.

TUTT, Andrew. An FDA for Algorithms. **Admin Law Review**, n. 83, p. 83-123, 2017.

WEST, S.M., Whittaker, M. and Crawford, K. **Discriminating Systems:** Gender, Race and Power in AI. Nova York: AI Now Institute. 2019. Disponível em: https://ainowinstitute.org/discriminatingsystems.html. Acesso em: 8 fev. 2020.

Artículo presentado el: 2020-05-02

Artículo aceptado el: 2020-05-31