CUADERNOS SALMANTINOS DE FILOSOFÍA
Vol. 48, 2021, 187-211, ISSN: 0210-4857
E-ISSN: 2660-9509

GOBERNANZA GLOBAL DE LA IA*

GLOBAL GOVERNANCE OF AI

Aníbal Monasterio Astobiza
Departamento de Filosofía I
Facultad de Filosofía y Letras
Universidad de Granada
Granada/España
amastobiza@ugr.es
ORCID: 0000-0003-1399-5388

Recibido: 19/04/2021
Revisado: 30/07/2021
Aceptado: 6/09/2021

Resumen: El desarrollo de la IA crece cada vez más y se aplica en casi todos los sectores económicos y/o sociales y, por consiguiente, nacen nuevas oportunidades, pero también nuevos retos. En este artículo, pretendo presentar algunos de estos nuevos retos, cuáles son las formas más apropiadas para afrontarlos y qué enfoque ético seguir para una gobernanza global de la IA. Para ello, en primer lugar, comentaré el proceso de difusión y penetración de la tecnología subrayando la inclinación humana de hacer uso de distintas herramientas para manipular y transformar la realidad. En segundo lugar, hablaré de la revolución digital y algunos desafíos que plantea para la sociedad y el individuo. Finalmente, presentaré un esbozo de los esfuerzos colectivos para la gobernanza global de la IA.

Palabras clave: gobernanza, ética, IA, retos, oportunidades.

Abstract: AI is increasingly growing and being applied in almost all economic and/or social sectors and, as a consequence, new opportunities are arising, but also new challenges. In this article, I intend to present some of these new challenges, what are the most appropriate ways to address them and what ethical approach to follow for a global governance of AI. To do so, I will first discuss the process of diffusion and pervasiveness of technology, highlighting the human disposition to make use of different tools to manipulate and transform reality. Secondly, I will discuss the digital revolution and some of the challenges it poses for society and the individual. Finally, I will present an outline of the collective efforts for the global governance of AI.

Keywords: governance, ethics, AI, challenges, opportunities.

* Este trabajo se ha realizado en el marco de los proyectos: EXTEND: Bidirectional Hyper-Connected Neural System" (Horizon2020) y EthAI+3 (PID2019-104943RB-100).

1. INTRODUCCIÓN

Industria, gobiernos, academia y organizaciones civiles buscan nuevas formas de aplicación y desarrollo de la Inteligencia Artificial (IA)[1] para múltiples propósitos. La IA promete a la industria y empresas maximizar beneficios; a los gobiernos automatizar procesos administrativos; a los académicos descubrir nuevas preguntas y respuestas científicas; y la sociedad civil (Organizaciones No-Gubernamentales, por ejemplo) quiere que las tecnologías digitales no incrementen las desigualdades en el acceso al estado de bienestar de los ciudadanos, no discrimine, ni conculque derechos fundamentales.

El uso y desarrollo de la IA crece cada vez más y se aplica en casi todos los sectores económicos y sociales (Mcafee y Brynjolfsson, 2017) y, por consiguiente, esto significa que nacen nuevos retos. En este artículo, pretendo presentar algunos de estos nuevos retos[2], cuáles son las formas más apropiadas para afrontarlos y qué enfoque ético seguir para una gobernanza global de la IA. A medida que el desarrollo tecnológico madura los avances tecnológicos son ubicuos y su impacto e influencia alcanzan a múltiples facetas de nuestras vidas y trabajo. Por todo ello, nos vemos en la necesidad de alinear estas poderosas nuevas tecnologías con nuestros valores fundamentales –integrar dichos valores en el diseño e implementación de la tecnología.

Si queremos que la tecnología –y en concreto la tecnología de la IA– esté al servicio de la gente y esté dirigida al bien común, la gobernanza global de la misma es una potencial solución porque implica una coordinación colectiva para alcanzar un consenso mínimo que permita tener el control sobre dicha tecnología, a saber, la IA y la robótica, que es vista por muchos como una amenaza existencial (Bostrom, 2014).

1 La IA es difícil de definir dado que distintos investigadores tienen una idea diferente. La IA puede verse como la ciencia e ingeniería que aspira a crear inteligencia en sistemas artificiales. La expresión IA se introdujo en los años 50 del siglo XX (McCarthy, Minsky, Rochester y Shannon, 1955), pero su historia profunda se remonta a los trabajos de filósofos como Ramón Llull, Thomas Hobbes o G. W. Liebniz que intentaron axiomatizar y formalizar reglas y principios para describir la razón. Una forma de describir la IA que casi puede servir de definición aproximativa es la que ofreció, en su día, Marvin Minsky: "La IA es la ciencia de hacer que las máquinas hagan cosas que requerirían inteligencia si las hicieran los seres humanos" (Minsky 1968/2003).

2 Los tres retos (más sobre esto en la sección 2) que he seleccionado para dar cuenta de la necesidad de un marco de gobernanza global de la IA son: *Automatización de las noticias y propaganda política computacional; Sesgos algorítmicos y discriminación;* y *El futuro del trabajo: desempleo tecnológico*. Los retos han sido seleccionados de manera subjetiva de la literatura de investigación siguiendo el criterio de conexión de la IA con la gobernanza. Es evidente la elección de estos retos. El uso de la IA en cada uno de estas áreas: Comunicación política, discriminación de usuarios y consumidores y mercado laboral, respectivamente; plantea una transformación radical de actividades esenciales a escala internacional y global.

En primer lugar, comentaré el proceso de difusión y penetración de la tecnología subrayando la propensión humana de hacer uso de distintas herramientas para manipular y transformar la realidad. En segundo lugar, hablaré de la revolución digital y los desafíos que plantea para la sociedad y el individuo. Finalmente, presentaré un esbozo de los esfuerzos colectivos para la gobernanza global de la IA.

Por gobernanza global de la IA –para los propósitos de este artículo– se entiende dos sentidos interdependientes. Por una lado, una gobernanza global de la IA *descriptiva* que hace referencia a diversos procesos por los que se toman decisiones y se implementan dichas decisiones y una gobernanza global de la IA *normativa* que hace referencia al conjunto *correcto* o *bueno* de tales decisiones. Cuando se habla de una gobernanza global de la IA *buena* se quiere decir efectiva, legítima, adaptativa e inclusiva (Renda, 2019; Dafoe, 2020). En algunos lugares de este artículo se utilizará la expresión "gobernanza global ética de la IA" para referirse a esta segunda acepción o sentido normativo de gobernanza y en otros, simplemente, "gobernanza global de la IA" para referirse a la primera acepción o sentido descriptivo de gobernanza. Sin embargo, nótese la interdependencia de ambos sentidos, aunque se puedan distinguir lógicamente. También el lector podrá encontrar que "regulación" y "gobernanza" se utilizan como términos coextensivos. Esto es así porque la gobernanza como proceso de toma de decisiones incluye, normas, instituciones, reglas y, por supuesto, leyes.

La situación de la pandemia de la Covid-19 muestra claramente la necesidad de una coordinación global para la resiliencia ante catástrofes, ya sean naturales o provocadas por la acción del ser humano. Prepararnos para una gobernanza global de la IA u otras tecnologías de uso-dual[3] es uno de los temas de gran importancia si queremos que la tecnología se dirija hacia un uso para el bien común. Países desarrollados y en vías de desarrollo han de ser conscientes de la necesidad de una gobernanza global de la IA, pero tampoco debemos olvidar la capacidad regulativa y sus contextos. El Sur Global[4] tiene bajos niveles de capacidad regulativa e institucional y quizá una menor garantía de la gestión de riesgos y esto es algo que puede generar retos adicionales.

3 Uso-dual hace referencia al uso de la tecnología con propósitos militares o comerciales. Más sobre esto en la sección 1.2.

4 A efectos de este artículo, el "sur global" se refiere a los países en desarrollo de África, América Latina y Asia incluyendo Oriente Medio. El sur global es un término que sustituye a "tercer mundo" y "países en desarrollo" en muchos debates académicos, aunque no está exento de controversia. El término sur global trasciende las fronteras y abarca países que comparten un pasado colonial o mantienen poblaciones oprimidas y privadas de derechos, incluso en Occidente o países desarrollados (véase, de Sousa Santos, 2016).

Europa es un promotor activo de la gobernanza global de la IA, pero organizaciones internacionales, múltiples países y regiones también entienden la necesidad de operativizar e implementar una gobernanza global de la IA.

En la siguiente sección, 1.1., introduciré y contextualizaré brevemente la historia del desarrollo tecnológico y su relación con el progreso civilizatorio.

1.1. Difusión y penetración de la tecnología

Desde la aparición del linaje homínido –en concreto a partir del género *Homo Habilis* (Oakley, 1959)– y hasta llegar al ser humano, nuestra especie no ha parado de instrumentalizar la realidad física. Gracias al desarrollo de la tecnología el ser humano ha podido ocupar todos los nichos ecológicos de este planeta. Pero hasta llegar a la hegemonía del ser humano en el control de los recursos naturales, si uno mira a la historia de nuestra especie nos llevó mucho tiempo hasta alcanzar una plena prosperidad.

Uno puede defender la idea de que en la historia global de la humanidad solo ha habido dos grandes eventos transformacionales: 1) la revolución agrícola y 2) la revolución industrial. Hasta hace 15.000 o 10.000 años los seres humanos sobrevivían cazando, recolectando o buscando carroña. Sin embargo, una serie de transformaciones culturales permitió el desarrollo de distintas técnicas para cultivar la tierra, hacer acopio del excedente, domesticar animales y un rápido proceso de urbanización y estratificación social (Herrera y García-Bertrand, 2018).

El segundo gran evento, la revolución industrial –que comenzó en el siglo XIX– es transformacional en muchos sentidos pero, principalmente, porque por primera vez en la historia de la humanidad el crecimiento económico permitió la explosión demográfica. Hasta la revolución industrial el crecimiento económico, la invención y desarrollo de nuevas tecnologías, únicamente permitía la mera, y difícil, subsistencia de la población. Estábamos en una condición *malthusiana* o en palabras del filósofo inglés Thomas Hobbes, la vida era "solitaria, pobre, sucia, bruta y corta"[5]. Pero todo esto cambió con la revolución industrial. Para hacernos una idea, con una imagen bastará.

5 '… and which is worst of all, continual fear, and danger of violent death; and the life of man, solitary, poor, nasty, brutish, and short' (Leviathan, i. xiii. 9).

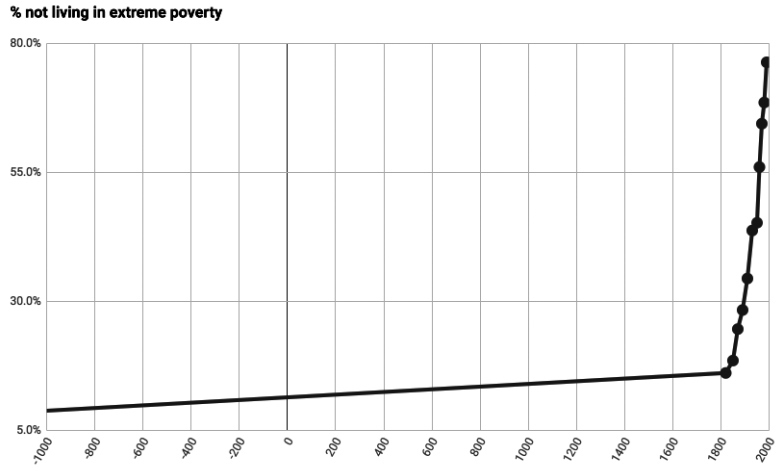

Figura 1. Imagen tomada Luke Muehlhauser Creative Commons Attribution-Noncommercial-ShareAlike 3.0.

La imagen de arriba muestra de manera clara y nítida cómo hasta la llegada de la revolución industrial el bienestar subjetivo era negativo y se mantuvo constante en la mayor parte de la población. Solo con la llegada de la revolución industrial el número de personas viviendo en pobreza extrema decreció.

De acuerdo con historiadores y demógrafos las condiciones de vida antes del año 1800 eran penosas. Solo el 43% de los recién nacidos llegaban al quinto cumpleaños y la esperanza de vida de quienes llegaban a la edad adulta rondaba los 30-35 años (Roser y Ritchie, 2019).

En términos económicos, renta per cápita, las cosas antes de la llegada de la revolución industrial no eran mejores. Otra imagen, bastará para mostrarlo.

Esta otra imagen de abajo muestra como la renta per capita mundial en dólares de 1990 era casi nula la mayor parte de la historia de la humanidad, es decir, las personas casi vivían en un absoluto estado de indigencia y/o pobreza, hasta que de nuevo el factor de la productividad a través de la revolución industrial hizo su aparición.

La revolución industrial representó un ciclo o periodo de gran innovación, difusión y penetración de la tecnología ayudado por factores como el capital humano y físico (Mankiw et al. 2992) e institucional (Acemoglu y Dell, 2010).

La revolución digital, en la que incluyo el desarrollo y avances de la IA y robótica, supone un salto exponencial en los usos y potenciales aplicaciones de la tecnología, pero también un incremento de los potenciales riesgos. Veámoslo, brevemente, en la siguiente sección 1.2.

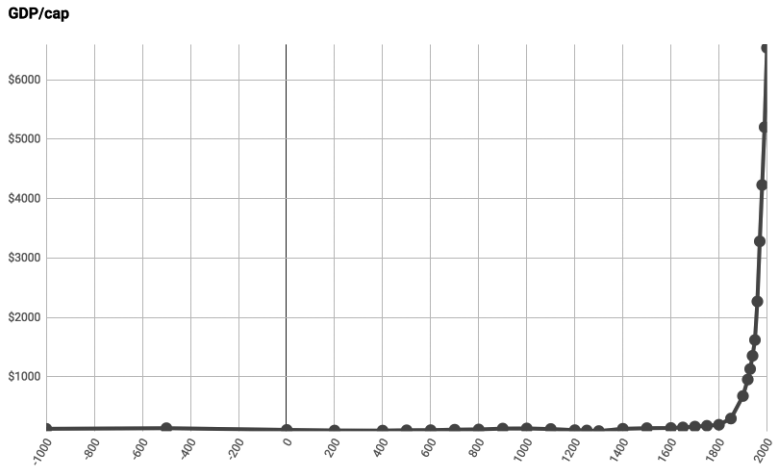

Figura 2. Imagen tomada de Luke Muehlhauser Creative Commons Attribution-Noncommercial-ShareAlike 3.0. con datos de De Long (1998).

1.2. Revolución digital y uso-dual de la tecnología

Como vemos, las nuevas tecnologías desarrolladas durante la revolución industrial mejoraron los niveles de bienestar individual y social. La mecanización de las fábricas hizo despegar el nivel de producción de bienes y productos de consumo. Llegada la década de los 1920s la población mundial llegó a 1.000 millones de personas, 4.000 millones durante la década de los 70s y 7.000 millones a principios del siglo XXI. Los niveles de vida, medidos por renta per cápita, para estos 7.000 millones de personas en todo el mundo, de acuerdo con datos del Banco Mundial (véase, https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=1W), creció de estar por debajo de los 500 $ en 1960 a 11.000 $ en 2018.

Desde hace unas décadas el mundo ha mejorado, incluyendo una mejora en la esperanza de vida, declive de la pobreza extrema, reducción de la mortalidad infantil, mayor acceso a la educación, agua potable y atenciones sanitarias (Pinker, 2018).

Muchos autores consideran que el desarrollo de la IA y tecnologías digitales es otra gran transformación y/o revolución para la historia de la humanidad. De hecho, la IA tiene el potencial de incrementar el producto interior bruto mundial de ahora a 2030 entre 13 y 15 $ trillones (Mckinsey Global Institute, 2018). La

IA es como dicen los economistas una "tecnología de propósito general" porque puede cambiar muchos aspectos de nuestras vidas, propiciar la innovación y hacernos más prósperos y productivos.

Pero como toda tecnología emergente con gran potencial, la IA tiene su lado oscuro. De acuerdo con el Instituto Nacional de Salud de los EE.UU. determinadas tecnologías emergentes en campos como las ciencias biomédicas, física, energía nuclear, nanomateriales, biología etc. generan una preocupación por su potencial "uso-dual" de la tecnología derivada (DURC Policy, 2012).

Esta agencia sanitaria de los EE.UU. define el "uso dual" de la tecnología de la siguiente manera:

> "…investigación que, con base en el entendimiento actual, se puede anticipar razonablemente que proporcionará conocimiento, información, productos o tecnologías que podrían ser directamente aplicados incorrectamente para plantear una amenaza significativa con amplias consecuencias potenciales para la salud y la seguridad pública, los cultivos agrícolas y otras plantas, animales, el medio ambiente, los materiales o la seguridad nacional".

Hasta hace poco la expresión "uso-dual" se refería primariamente a tecnologías que podían tener aplicaciones civiles y militares. Actualmente la expresión tiene múltiples sentidos. El desarrollo de la ciencia y la tecnología crea demandas hasta hace poco impensables por los propios científicos e ingenieros que las desarrollan. Considerar la amplia gama de potenciales usos indebidos de la tecnología junto a sus implicaciones de seguridad, el uso dual de la investigación, es altamente complejo, pero necesario.

Muchas tecnologías emergentes presentan importantes desafíos éticos, legales, sociales y políticos. Pero la IA, investigación genómica, nanotecnología, robótica, ciencias de la computación en general y neurociencia aplicada tienen la capacidad de ser aditivas y multiplicativas. Es decir, aplicadas conjuntamente la tecnología derivada de estas áreas de investigación pueden manipular los átomos de la materia física, la información digital, biológica (ADN) y las células nerviosas.

Si nos centramos en la IA, la IA tiene como objetivo crear máquinas inteligentes y como corolario entender mejor la inteligencia biológica. En su propósito de crear máquinas inteligentes pronto los investigadores en IA se dieron cuenta que la mejor estrategia era crear modelos de inteligencia que aprendan en lugar de tener que ser programados por un ingeniero o matemático humano desde el principio.

Así nació el "aprendizaje automático" (machine learning) y una técnica dentro de este subcampo llamada "aprendizaje profundo" (deep learning) en la que se desarrollan "redes neuronales" que imitan la arquitectura y propiedades de las neuronas del cerebro humano. En otras palabras, nodos computacionales

interconectados entre sí que imitan la capacidad del cerebro humano de percibir, procesar y representar información.

El aprendizaje automático mejora cada día gracias a las grandes cantidades de datos e información que generamos y podemos anotar/etiquetar, así como analizar (Big Data). Y si bien es cierto que gracias a estos desarrollos la IA puede ofrecer diagnósticos fiables en medicina sobre la base de los síntomas y recomendar determinados tratamientos y detectar anomalías en imágenes con mayor fiabilidad que radiólogos; o en el derecho procesar miles de sentencias en cuestión de milisegundos; o en la movilidad ofrecernos los vehículos sin conductor mejorando la seguridad y reduciendo la siniestralidad y los accidentes; la IA es también capaz de conducirnos a un futuro distópico.

Con IA se pueden crear *robots kamikazes*, vehículos sin conductor diseñados para colisionar intencionalmente, drones convertidos en misiles o con la capacidad de portar armas, videos (e imágenes) sintéticos generados por IA con la intención de propagar bulos o falsas noticias (fake news), armas autónomas letales, virus para inutilizar sistemas informáticos que controlan nuestras infraestructuras básicas, sistemas de reconocimiento facial que vigilen y persigan a grupos determinados… y la lista puede seguir.

Por estos "usos-duales" con consecuencias negativas, la IA tiene un lado oscuro que no debemos obviar. En la siguiente sección hablaré con más detalle de tres retos a los que nos enfrentamos si usamos la IA con propósitos espurios y la necesidad de superarlos para alcanzar una verdadera gobernanza global ética de la IA.

2. TRES GRANDES RETOS DE LA IA

Uno de los retos más importantes que plantea el desarrollo de la IA es el uso presente y futuro de las tecnologías computacionales para manipular el comportamiento del público y sus actitudes con propósitos propagandísticos y políticos, así como comerciales (marketing). Otro segundo reto importante son los sesgos algorítmicos. Cada vez más automatizamos nuestra toma de decisiones y delegamos nuestras decisiones a sistemas algorítmicos para tomar decisiones con consecuencias en múltiples campos económicos y sociales. Sin embargo, estos algoritmos pueden estar sesgados y discriminar a ciertos colectivos o grupos específicos por su condición, raza, sexo etc; y al mismo tiempo acceder a información personal sensible. Por último, el tercer reto de gran importancia es el futuro del trabajo. El desempleo tecnológico o el desplazamiento de las personas de sus lugares de trabajo por la implantación de máquinas, es quizá el reto más acuciante. Si la gente

no tiene posibilidad de trabajar porque las máquinas pueden hacer su trabajo se puede quebrar la paz social.

Veamos cada uno de estos retos con más detalle para tener una idea de la importancia que tiene superarlos para alcanzar una gobernanza global ética de la IA.

Automatización de las noticias y propaganda política computacional

Bajo este epígrafe quiero introducir el reto que se deriva de la aplicación presente y futura de las tecnologías computacionales para la manipulación de la conducta de la gente con propósitos políticos y comerciales (marketing).

La interacción entre automatización, algoritmos y política puede ser explosiva. Se pueden crear chat-bots (agentes de software autónomos) que interactúen en las redes sociales virtuales con los usuarios y manipulen la opinión pública sobre ciertos asuntos cruciales para la política. Por ejemplo, durante las elecciones presidenciales de los EE.UU. acabó siendo elegido Donald Trump como 45º presidente. Numerosos estudios achacan este hecho a la acción de operativos rusos (hackers) que compraron miles de anuncios en plataformas como Facebook para explotar las divisiones raciales, religiosas e ideológicas fomentando la polarización de la sociedad norteamericana[6].

Se puede fabricar la polarización o el consenso con el uso de sofisticados algoritmos que crean anuncios micro-dirigidos a estratos sociales específicos generando "filtros burbuja" o cámaras de eco, es decir, entornos de información donde la gente solo recibe aquella información que confirma lo que ya sabe o, en el peor de los casos, confirma sus prejuicios (Pariser 2012). Los filtros burbuja o cámaras de eco son búsquedas en Internet que algoritmos sofisticados personalizan para ofrecernos resultados que nos gustaría ver. Si somos conservadores ideológicamente, únicamente tendremos resultados en los motores de búsqueda de noticias que confirmen nuestras inclinaciones ideológicas conservadoras y si somos progresistas ideológicamente, únicamente tendremos resultados que confirmen nuestras inclinaciones progresistas.

También con el uso estratégico de chat-bots se puede retroalimentar y reforzar ideas. Los bots programados para interactuar con ciertos usuarios pueden

6 Ciertamente, en teoría, es posible presentar anuncios personalizados a los usuarios de ciertas redes sociales como, por ejemplo, Facebook. Lo que los expertos ponen en duda es la relación causal entre estar expuesto a contenido personalizado y la modificación del comportamiento político, es decir, que dicho contenido tengo algún efecto a la hora de cambiar la preferencia de voto. Esto es algo que todavía no está demostrado (Baldwin-Philippi, 2017).

amplificar ideas cuyo contenido puede ser negativo, explotarlo y, en última instancia, manipular la conducta.

La consecuencia de todo esto es una sociedad polarizada, dividida, sin capacidad de puntos de encuentro, gente aislada intelectualmente por la acción de algoritmos con un propósito político. El escándalo de Cambridge Analytica (The Guardian 2018) es un caso de filtración de datos y erosión de la confianza y de la privacidad por parte de compañías tecnológicas como Facebook, pero también es un caso donde se pone de manifiesto la posibilidad de manipulación política a través de las redes sociales (Véase, nota al pie de página número 3).

La desinformación es otra consecuencia nefasta del uso de herramientas computacionales o IA con el propósito de desestabilizar la sociedad. Es cierto que las mentiras, las falsas noticias, han existido siempre. Pero con cada tecnología nueva el público potencial y la audiencia a la que podían llegar se multiplicaba. Las mentiras y el engaño evolucionaron gracias al lenguaje. Con la escritura, la información falsa multiplicó su campo de acción, pero es que con las TIC (tecnologías de la información y comunicación) se pueden difundir mentiras a todos los rincones del globo y en todos los soportes imaginables. Por otra parte, características de captura de atención (economía de la atención) e incluso mecanismos de refuerzo condicionado conducentes a hábitos pseudo-adictivos en el diseño de la tecnología (Williams 2018), permiten una propagación y viralización de contenidos mucho mayor (Aral, 2020).

Sesgos algorítmicos y discriminación

La IA se puede aplicar a los gobiernos y los asuntos administrativos (Ramió, 2019). La IA puede realizar tareas predictivas, asignación de colegios para niños y niñas, distribución de recursos sociales y sanitarios. En la empresa privada, la IA puede utilizarse para la selección de personal y contratación, facilitar créditos financieros, selección de perfiles para publicidad etc. Todas estas decisiones, sumamente importantes, se pueden automatizar con la aplicación de sistemas algorítmicos de toma de decisiones.

El problema es que muchas veces los algoritmos responsables de la toma de decisiones están sesgados o son opacos y/o poco transparentes, no protegen la información personal etc., es decir, no se sabe cómo han tomado la decisión que han tomado, ni porqué (Pasquale, 2015).

Por ejemplo, en un distrito de California, EE.UU., jueces y fiscales usan una herramienta algorítmica para decidir si una persona investigada por un delito debe salir de la cárcel antes de un juicio ante un tribunal. Una investigación del grupo de

periodistas ProPublica determinó que el algoritmo usado discriminaba negativamente a personas investigadas de raza negra. La herramienta algorítmica llamada COMPAS generaba una puntuación para estimar el grado de reincidencia en un periodo de tiempo de dos años tras salir de la cárcel.

La empresa que había creado el algoritmo llamada Northpointe (ahora se llama Equivant) sacó un comunicado de prensa defendiendo que los algoritmos no estaban sesgados. Pero el estudio de ProPublica fue la suficientemente riguroso como para mostrar los sesgos y desequilibrios estadísticos existentes en el algoritmo que es usado para automatizar decisiones muy importantes como, por ejemplo, a quién se le priva de libertad (Véase, Monasterio Astobiza 2017, para un análisis en detalle sobre la ética de los algoritmos).

El problema, muchas veces, consiste en qué definición de equitativo o justo (fair) se está usando a la hora de diseñar y construir un algoritmo con aplicaciones sociales. Habitualmente, los científicos computacionales no colaboran con filósofos para obtener una definición de equidad o justicia de la ingente literatura en filosofía política con una historia de más de 2.500 años de teorización. Y aún así, se aventuran a ofrecer una caracterización matemática de la idea filosófica de justo o equitativo (fair) que puede implicar variadas y diferentes implementaciones y eso solo si tenemos en consideración la literatura de investigación en ciencias de la computación que además diferencia entre sesgos sociales de la aplicación del aprendizaje máquina (machine learning) y sesgos estadísticos.

Estos últimos, los sesgos estadísticos, son simplemente la diferencia entre el valor esperado de una estimación y el verdadero valor. Por otro lado, los sesgos sociales, que reproducen los prejuicios de un sistema social injusto y que pueden estar presentes en los algoritmos usados en modelos de aprendizaje automático para automatizar decisiones, son más complejos y con graves repercusiones que pueden amenazar y poner en peligro la democracia (O'Neil, 2016).

Algunas definiciones de equidad en algoritmos para evitar que expresen sesgos sociales son:

— *Equidad de grupo*: trata grupos diferentes de manera igual (Dwork et al. 2012).

— *Equidad individual*: individuos similares deben ser considerados similarmente (Pleiss et al. 2017).

— *Equidad de proceso*: equilibrio entre precisión, equidad de proceso y equidad de resultados (Grig-Hlaca et al. 2016).

— *Probabilidades igualadas*: (equalized odds): la probabilidad de una persona perteneciente a una clase que se le asigna un atributo positivo y la probabilidad de una persona perteneciente a una clase que se le asigna

un atributo negativo debe ser igual que para miembros protegidos y no-protegidos del grupo general (Hardt et al. 2016).

— *Igualdad de oportunidad*: la igualdad de oportunidades para grupos protegidos o no-protegidos debe tener ratios positivos iguales (Hardt et al. 2016).

— *Equidad a través de la conciencia*: (fairness through awareness): dos individuos cualquiera que sean similares con respecto a una métrica deben recibir el mismo tratamiento (Dwork et al. 2012).

Una cuestión básica a tener en cuenta sobre el impacto social de los sesgos algorítmicos es que los algoritmos, procedimientos y/o protocolos computacionales, no son neutros ni objetivos. Los algoritmos se enmarcan en un contexto tecnológico, económico, ético, temporal y espacial concreto. Y, además, están imbuidos de los valores conscientes e inconscientes de las personas que los diseñan. Responden a los intereses del diseñador que de manera, muchas veces, inconsciente, en forma de prejuicios, sesgos, estereotipos; crea algoritmos que retroalimentan el sistema de prejuicios social injusto existente.

Los sesgos sociales no son los únicos efectos perversos de los algoritmos que afectan a la vida de las personas creando discriminación etc. Los algoritmos también explotan vulnerabilidades y pueden extraer información sensible de las personas. Por consiguiente, la privacidad se ha convertido en uno de los valores centrales en nuestro tiempo. En un primer momento, uno podría pensar que los sesgos algorítmicos y los temas de privacidad son fenómenos autónomos e independientes, pero no es así. Un algoritmo pude tomar decisiones automatizadas que causen un perjuicio o discriminen a un individuo que pertenece a un colectivo vulnerable, porque accede a una variable identificativa de ese individuo que no está suficientemente protegida como información sensible y personal. Por eso en esta sección hablo sobre discriminación, sesgos y privacidad de manera continua e indistintamente.

La privacidad es difícil de definir. Quizá la definición de privacidad más famosa es la de Warren y Brandeis (1890, p. 20): "el derecho a estar solo". A partir de una revisión de las distintas teorías filosóficas y legales de privacidad, Tavani (2007, p. 2) agrupa en cuatro categorías las teorías de privacidad:

a) No-intrusión: privacidad definida como derecho a ser libre de toda intromisión.

b) Aislamiento: privacidad entendida como estar físicamente inaccesible a otros.

c) Limitación: privacidad entendida como acceso a nuestra información restringida en ciertos contextos.

d) Control: privacidad entendida como el control sobre nuestra información.

A partir de estas definiciones de privacidad obtenidas de la filosofía y el derecho y tras décadas de promover distintos enfoques y metodologías se ha llegado a una definición matemática significativa de privacidad, llamada: privacidad diferencial (differential privacy). La privacidad diferencial cuantifica la semántica de la privacidad introduciendo una serie de constreñimientos matemáticos a un algoritmo (Dwork y Roth 2014). La privacidad diferencial ofrece una definición matemática rigurosa de privacidad. Para entender la idea de la privacidad diferencial imagina un algoritmo que analiza una base de datos y realiza operaciones computacionales sobre ella (por ejemplo, la media, mediana, varianza de los datos, etc.) Dicho algoritmo se dice que es diferencialmente privado si al mirar el resultado no se puede saber si los datos de algún individuo se incluyeron en el conjunto de datos original o no.

La privacidad diferencial como técnica protege a los individuos de una base de datos de su posible identificación introduciendo "ruido" en los datos. Esto produce una distribución probabilística que hace prácticamente imposible determinar el registro de alguien en una base de datos. Dicho de otra manera, alguien que quiera identificar a una persona a partir de la información contenida en la base de datos no podrá hacerlo, porque no se puede vincular la información contenida en la base de datos con ningún individuo.

Para conseguir que no haya sesgos sociales, ni discriminación algorítmica, porque previamente se ha accedido a datos sensibles y personales sin consentimiento, es decir, se ha violado la privacidad de alguien; se debe diseñar algoritmos que sean socialmente sensibles. No solo se deben establecer leyes, regulaciones y comités de supervisión, también se debe crear y diseñar tecnología ética (ethics by design). De esta forma se construirán algoritmos con mejores definiciones de equidad, precisión, transparencia, privacidad y con una ética integrada en su diseño (Kearns y Roth 2019). La "ética por diseño" es un enfoque que busca la inclusión sistemática de valores y principios éticos así como otros requerimientos y procedimientos en el diseño y procesos de desarrollo de la tecnología y sistemas de IA. La "ética por diseño" es un ejemplo de Enfoque de Diseño Sensible al Valor (Friedman, Kahn y Borning, 2002), que se está convirtiendo en el enfoque globalmente recomendado para el desarrollo de la IA. En conclusión, la "ética por diseño" es un enfoque que se asegura que los sistemas de IA y robótica cumplan ciertos valores éticos fundamentales para que su implementación respete los derechos y dignidad de las personas.

El futuro del trabajo: desempleo tecnológico

La IA va a transformar el futuro del trabajo. De hecho ya lo está haciendo. La IA tiene y va a tener un impacto en la productividad, en los salarios y en el empleo. En general, la IA y la automatización del trabajo seguirá el famoso adagio del filósofo Sombart, popularizado por el economista Shumpeter: destrucción creativa. Es decir, la IA destruirá empleos, pero, al mismo tiempo, creará nuevos empleos.

No obstante, a pesar de la ingente cantidad de inversiones, talento, productos y servicios (como por ejemplo, capacidad de visión computacional a un nivel de rendimiento mayor que el humano o reconocimiento de voz también a un nivel de rendimiento mayor que el humano gracias a técnicas como las redes neuronales artificiales) el nivel de productividad se ha reducido (Brynjolfsson, Rock y Syverson 2017). Esto se conoce como *la paradoja de la productividad*.

Brynjolfsson, Rock y Syverson (2017) dan cuatro razones para explicar la paradoja de la productividad. La primera es quizá que pecamos de una ilusión de optimismo. Puede ser que seamos muy optimistas y la tecnología de la IA no está a la altura de las expectativas. Otra razón es que puede que no estemos midiendo bien los beneficios de la tecnología en términos de productividad. Una tercera razón es que puede que la tecnología solo esté beneficiando a unas pocas organizaciones o industrias y todavía no ha llegado al público en general. Finalmente, la razón que defienden los autores es que la mejora de la tecnología en productividad es real, pero como las organizaciones tardan en reestructurarse los beneficios de la tecnología tardan en hacerse visibles.

Brynjolfsson –profesor en el Instituto de Stanford para la IA centrada en el ser humano (HAI) y director del Laboratorio de Economía Digital de Stanford– y otros economistas no dudan en describir a la tecnología de la IA como una tecnología de propósito general como lo es la electricidad. Porque la IA es una tecnología de propósito general que necesita innovación complementaria e inversión para poder hacer realidad sus promesas. Pero para ello quizá tenemos que reinventar nuestras organizaciones, instituciones y métricas.

A pesar de que la tecnología desplazará y dejará sin trabajo a muchas personas –y no tienen por qué ser solo personas de sectores que realicen tareas mecánicas, rutinarias y con bajo cualificación, también los trabajadores de profesiones liberales como abogados, ciertas especialidades médicas, etc.; pueden ser desplazados por máquinas (Frey y Osborne 2013)– la mayoría de los economistas creen que la IA y la robótica no dejará sin empleo a grandes cantidades de personas y que el desempleo tecnológico es un miedo infundado.

Sin lugar a dudas, la tecnología de la IA, la robótica y la tecnología digital transformarán nuestros trabajos y vidas. Pero estamos todavía muy lejos de un mundo sin trabajo (Danaher, 2019). Para bien o para mal, todavía no estamos en un mundo donde reine el ocio y el tiempo libre, ni tampoco en un mundo en caos y sin paz social porque la gente no tenga, ni encuentre, trabajo por el desempleo tecnológico.

En la siguiente sección 3, hablaré sobre el tema central de este artículo y que da título al mismo: La gobernanza global de la IA. Como hemos visto los sistemas de IA son el resultado de esa querencia humana por el uso del conocimiento para la creación de artefactos y herramientas que permiten resolver problemas y extraer un mayor rendimiento del entorno para mejorar el bienestar. Pero con la llegada de los avances y desarrollos de la IA y la revolución digital los potenciales riesgos también han aumentado. He comentado tres retos de escala global que necesitan una cooperación internacional para una estabilidad geopolítica. Veamos ahora la necesidad de una respuesta colectiva para mitigar los efectos no deseados de la transformación digital del mundo.

3. GOBERNANZA GLOBAL DE LA IA

Existe un creciente debate sobre cómo regular y hacer un uso responsable de las tecnologías digitales, en particular de la IA. Los países desarrollados, con una larga tradición de instituciones y mecanismos de contrapeso al poder, son capaces de implementar procesos de deliberación pública entre diferentes actores y/o partes interesadas (por ejemplo, la sociedad civil, grupos de afectados, organizaciones no gubernamentales etc.) y, por lo tanto, empezar a construir una arquitectura de gobernanza ética para la IA (y tecnología digital conexa) supuestamente basada en el imperio de la ley. Sin embargo, en el llamado sur global (por ejemplo, países con una historia poscolonial, también llamados países no desarrollados) y, principalmente, en países con regímenes autoritarios donde su situación de vulnerabilidad y dependencia económica del liderazgo del norte global les lleva a importar tecnología digital, capital y modos de organización a pesar de sus diferencias políticas, legales y culturales; es realmente muy difícil una regulación de la IA (Arun, 2020).

Esto puede tener un impacto muy negativo en sus poblaciones ya de por sí excluidas, oprimidas y discriminadas. La cuestión central es hasta qué punto los países no desarrollados con culturas políticas e institucionales diferentes que importan tecnología digital de los países del norte global pueden verse afectados si no se tiene en cuenta la necesidad de una gobernanza global ética de la IA y a varios niveles desde una perspectiva de derechos humanos/democrática. En el

debate sobre la gobernanza ética, hay una posición que sugiere una moratoria sobre cualquier intento de regular la IA. Según los defensores de esta postura, la regulación mata la innovación (Ahmed, 2015). Creo que los argumentos de esta posición son erróneos, porque la IA ya tiene un enorme impacto en la vida de millones de personas en todo el mundo (por ejemplo, en la economía de plataformas ya controlada por algoritmos, y/o la IA que se utiliza para contratar a personas, conceder préstamos o decidir la libertad condicional de delincuentes).

Por ello, una distinción útil es: "gobernanza ética blanda" y "gobernanza ética dura (basada en la ley)". Esta distinción puede ser aceptada por los escépticos de la regulación de la IA. Por ejemplo, la "gobernanza ética blanda" describe normas como los marcos ISO o IEEE que pueden utilizarse en las primeras fases del desarrollo de la tecnología digital. Por otro lado, la "gobernanza ética dura (basada en la ley)" se refiere directamente a las prohibiciones e impide el uso de una tecnología cuando el riesgo supera los beneficios[7]. Quizá la mejor recomendación para una gobernanza global de la IA es partir de la experiencia con otras tecnologías, como la energía atómica, con la intención de crear una gobernanza global ética de la IA desde un enfoque de derechos humanos/democrático. También es útil plantear el caso de la gobernanza global de la IA como un riesgo existencial para la humanidad similar a como resulta ser el cambio climático.

El problema del cambio climático desborda las fronteras de un único país y se convierte en un problema de coordinación colectiva entre países. Si se quiere atajar los riesgos que supone un incremento de las temperaturas por la acción antrópica de emisión de gases de efecto invernadero a la atmósfera, se necesita una respuesta global coordinada. De la misma manera, las tecnologías disruptivas, como la biotecnología, pero especialmente la IA; también requiere de una

7 A fecha de escritura de este artículo la UE ha publicado el documento: Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas sobre inteligencia artificial-Ley sobre Inteligencia Artificial. La propuesta indica que la UE sostendrá una postura firme respecto a determinadas aplicaciones de la IA, diferenciándose de Estados Unidos y China. Este nuevo reglamento de ser adoptado –que implica que será de aplicación en toda la UE– quiere que los europeos puedan confiar en lo que la IA puede ofrecer. El proyecto de reglamento incluye la prohibición de la IA para la "vigilancia indiscriminada", incluidos los sistemas que rastrean directamente a las personas en entornos físicos o agregan datos de otras fuentes. Prohibición de los sistemas de IA que crean puntuaciones de crédito social, es decir, que juzgan la fiabilidad de una persona en función de su comportamiento social o de los rasgos de personalidad previstos. Autorización especial para el uso de "sistemas de identificación biométrica remota", como el reconocimiento facial, en espacios públicos. Notificaciones necesarias cuando las personas interactúan con un sistema de IA, a menos que sea "obvio por las circunstancias y el contexto de uso".Nueva supervisión de los sistemas de IA de "alto riesgo". Evaluación de los sistemas de alto riesgo antes de su puesta en servicio, lo que incluye garantizar que estos sistemas sean explicables para los supervisores humanos y que se entrenen con conjuntos de datos de "alta calidad" sometidos a pruebas de sesgo y la creación de un "Consejo Europeo de Inteligencia Artificial".

gobernanza global porque los riesgos son compartidos y no exclusivos de un solo país.

Mucho se habla de la "carrera armamentística" por el liderazgo en la revolución digital (Lee 2018) entre dos potencias hegemónicas: USA y China. Ninguna de estas dos potencias quiere quedarse atrás y las razones son claras. El primer país que domine la tecnología de la IA obtendrá una ventaja competitiva muy grande que será reforzada porque el país ganador obtendrá mayores recursos que a su vez mejorará sus sistemas. Esto refleja un escenario fuertemente competitivo donde la seguridad y el control de la tecnología pude pasar a un segundo plano, porque lo que verdaderamente importa, la prioridad, es ganar a los competidores.

Este duopolio (USA-China) en la carrera armamentística por liderar la tecnología de la IA es quizá una simplificación. En primer lugar, porque bajo el paraguas abstracto de IA hay distintas aplicaciones, técnicas etc. que no todas se solapan (el "aprendizaje profundo" no es lo mismo que un "sistema experto" y no se aplican a los mismos campos, p. ej.), pero, principalmente, porque no tiene en cuenta la interdependencia de los complejos militares-industriales de los países y ni siquiera el hecho de que se comparten internacionalmente infraestructuras, patentes y talento.

Sin embargo, aun siendo una simplificación esta caracterización de duopolio entre USA y China en el desarrollo de IA, hay que enfatizar que la motivación subyacente de la competición nacional por el dominio en la revolución digital tiende a centrarse en adquirir nuevas capacidades y mucho menos en la gobernanza, control y regulación del impacto social y ético de las tecnologías emergentes de la IA. Dicha competición nacional entre dos superpotencias y el resto de países por el liderazgo en la tecnología de IA, no solo es una simplificación, sino que es un sinsentido desde un punto de vista del interés general y el bien común. La tecnología de la IA no es un juego suma-cero porque nada se pierde por compartir nuevas capacidades, algoritmos, infraestructuras, talento etc. De hecho, los investigadores en IA de universidades forman parte de una comunidad internacional que comparten intereses, conferencias, acuerdos, patentes y sociedades profesionales como la AAAI (siglas en inglés de la Asociación para el Avance de la Inteligencia Artificial) y IEEE (siglas en inglés del Instituto de Ingenieros Eléctricos y Electrónicos).

El riesgo está en los desarrollos de IA en corporaciones y empresas grandes y/o pequeñas que en lugar de regirse por un *etos* colaborativo y cooperativo, pueden guiarse por la competencia y la maximización del beneficio. Pero aun así, en las corporaciones tecnológicas suelen existir comités éticos que supervisan la I+D+I (investigación, desarrollo e innovación). También existen consorcios de industria, como el Parnership on AI (https://www.partnershiponai.org/) y organizaciones supranacionales como la ONU que tienen el poder de convocar eventos

y pueden redactar documentos no vinculantes, pero con fuerza simbólica y moral que denotan acuerdo y consenso en principios básicos. De hecho, de ser posible una potencial gobernanza global de la IA está debe inspirarse en la experiencia de éxito de instancias supranacionales como la ONU. La cooperación internacional para la regulación de la IA es fundamental.

Del mismo modo, numerosos países están dotándose de grupos de expertos y consejeros que ayudan en el proceso de regulación. Un ejemplo notorio en este sentido es el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Unión Europea (Véase para más información este vínculo: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

Este Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Unión Europea ha elaborado una guía de principios y propone siete requerimientos clave que todo sistema de IA debe cumplir para ser considerado confiable (AI HLEG 2019):

- *Agencia y supervisión humana*: Los sistemas de IA deben empoderar a los seres humanos, permitiéndoles tomar decisiones informadas y fomentando sus derechos fundamentales. Al mismo tiempo, es necesario garantizar unos mecanismos de supervisión adecuados, lo que puede lograrse mediante enfoques basados en el principio de "humano en el bucle", "humano en el bucle" y "humano en el mando".
- *Robustez y seguridad técnica*: Los sistemas de IA deben ser resistentes y seguros. Necesitan ser seguros, asegurando un plan alternativo en caso de que algo salga mal, además de ser precisos, fiables y reproducibles. Esa es la única manera de asegurar que también se pueda minimizar y prevenir el daño no intencional.
- *Privacidad y gobernanza de los datos*: además de garantizar el pleno respeto de la privacidad y la protección de los datos, también deben garantizarse mecanismos adecuados de gobernanza de los datos, teniendo en cuenta la calidad y la integridad de los datos, y garantizando un acceso legítimo a los mismos.
- *Transparencia*: los datos, el sistema y los modelos de negocio de la IA deben ser transparentes. Los mecanismos de trazabilidad pueden ayudar a conseguirlo. Además, los sistemas de IA y sus decisiones deben explicarse de manera adaptada a las partes interesadas. Los seres humanos deben ser conscientes de que están interactuando con un sistema de IA y deben estar informados de las capacidades y limitaciones del sistema.
- *Diversidad, no discriminación y equidad*: Debe evitarse el sesgo injusto, ya que puede tener múltiples consecuencias negativas, desde la marginación de los grupos vulnerables hasta la exacerbación de los prejuicios y

la discriminación. Al fomentar la diversidad, los sistemas de IA deben ser accesibles para todos, independientemente de cualquier discapacidad, e involucrar a los actores relevantes a lo largo de todo su ciclo de vida

- *Bienestar social y ambiental*: Los sistemas de IA deben beneficiar a todos los seres humanos, incluidas las generaciones futuras. Por lo tanto, hay que garantizar que sean sostenibles y respetuosos con el medio ambiente. Además, deben tener en cuenta el medio ambiente, incluidos otros seres vivos, y su impacto social debe ser cuidadosamente considerado.
- *Rendición de cuentas*: Se deben establecer mecanismos para garantizar la responsabilidad y la rendición de cuentas de los sistemas de inteligencia artificial y sus resultados. La auditabilidad, que permite la evaluación de algoritmos, datos y procesos de diseño; juega un papel clave en ello, especialmente en aplicaciones críticas. Además, debe garantizarse una reparación adecuada y accesible.

Gobiernos, empresas, organizaciones civiles y activistas de todo el mundo han desarrollado principios de gobernanza ética de la IA. Entre los principios de gobernanza global ética de la IA que preponderan entre los distintos actores destacan la justicia, seguridad, transparencia, protección de la privacidad, tecnología alineada con valores humanos, tecnología humano-céntrica, transparencia, explicabilidad, y finalmente, responsabilidad (Jobin, Ienca y Vayena 2019).

Miremos de nuevo al problema del cambio climático como ejemplo de problema de coordinación colectiva del que extraer lecciones para la gobernanza global de la IA. El problema del cambio climático se asemeja a la tragedia de los comunes (Hardin 1968). Imagina un pasto abierto. Es evidente que cada pastor intentará que sus animales se alimenten en el pasto. Como seres racionales cada pastor piensa que sí mete más animales mayor será el beneficio, pero está lógica o manera de pensar, al llevarla todos acabo, al final crea un drama o tragedia visible. El pasto es un recurso natural, finito, y si cada pastor va metiendo más animales al pasto, el recurso común, el pasto, se acaba.

Por su parte, el problema del desarrollo tecnológico en IA es un problema similar que requiere una estrategia global para evitar escenarios suma cero, es decir uno gana, pero el resto pierde; y crear escenarios de ganancia para todos (e.g. win-win). Por ejemplo, si un país desarrolla una tecnología de IA que resulta ser beneficiosa para la economía y prosperidad de los ciudadanos de un país, el resto de países la copiará porque no se querrán quedar atrás. Sin embargo, esto conduce a una carrera armamentística de todos contra todos que puede llevar a una guerra comercial y de patentes que finalmente puede provocar una reducción en el estado de bienestar de los ciudadanos de todos los países. En la medida en que ningún país se querrá quedar atrás en el desarrollo tecnológico, muchos

pueden saltarse las precauciones de desarrollar IA segura que no produzca riesgos existenciales (Bostrom 2014).

En definitiva, el cambio climático, así como el desarrollo tecnológico y potenciales riesgos de la IA, es un problema de coordinación y gobernanza global porque soluciones en el ámbito de un solo país no pueden luchar contra el cambio climático, ni tampoco contra los riesgos y retos del desarrollo tecnológico en IA, respectivamente.

Uno de los objetivos para una plena gobernanza global de la IA y tecnologías adyacentes es crear un comité ético que incluya a gobiernos, industria y sociedad civil similar a como se coordinan los gobiernos y estados-nación en torno al cambio climático con el Panel Intergubernamental sobre el Cambio Climático. De hecho, una iniciativa similar se está llevando a cabo por funcionarios de alto nivel de Canadá y Francia para establecer el Panel Internacional sobre IA (IPAI, siglas en inglés).

Esfuerzos académicos también están sumándose para crear una gobernanza global ética de la IA. Por ejemplo, Wendell Wallach, investigador de la Universidad de Yale en el Centro Interdisciplinar de Bioética y Gary Marchant, director del Centro del Derecho e Innovación de la Universidad de Arizona, han propuesto lo que llaman Comités Coordinadores de Gobernabilidad (CCG).

Los CCG coordinarían las actividades de las distintas partes interesadas, supervisarían la evolución de la situación, tomarían nota de las mejores prácticas y determinarían qué instituciones están asumiendo la responsabilidad de las diversas preocupaciones, así como las lagunas existentes.

Las aplicaciones de técnicas computacionales provenientes del aprendizaje máquina o IA en general tienen un impacto en la sociedad (definen cómo interactuamos a través de plataformas como Facebook, Twitter, Whatsapp…), economía (Airbnb, Uber…), educación (MOOCS…) y hasta determinan nuestro entretenimiento (You Tube, Spotify…). Este poder transformador que encierra la IA puede generar grandes beneficios y mejorar nuestras vidas, pero al mismo tiempo tiene el potencial de ser discriminador, reforzar los prejuicios sociales y polarizar a la sociedad.

Las ramificaciones de la toma de decisiones por algoritmos o máquinas se dejan ver en el derecho, en la medicina, ejército y fuerzas armadas, ciencia… Para que realmente tengan un efecto transformador positivo deben integrar una supervisión humana basada en principios de ética práctica y lo que es más importante, el derecho internacional y las regulaciones estatales deben a través de distintos mecanismos, como códigos de buenas prácticas (soft law), poner en marcha una plétora de prácticas de monitorización internacional del impacto de la IA.

También existen las regulaciones y las legislaciones (hard law) que instituyen obligaciones y deberes, además de prohibir ciertas conductas. A través de la ley y agencias regulatorias se puede establecer lo que se puede hacer y lo que no se puede hacer. Sin embargo, entre los códigos de buenas prácticas (soft law) y las leyes regulatorias (hard law), muchos autores prefieren los primeros porque no coartan el avance y progreso de la innovación e investigación, en este caso de la tecnología en IA. De ahí la distinción presentada más arriba entre "gobernanza ética blanda" y "gobernanza ética dura" Pero siempre se ha de seguir un enfoque ético o marco normativo que minimice los riesgos.

Sin embargo, al ser gobernanza de la IA basada en *soft law* y no gobierno o burocracia basado en *hard law* es posible que agentes estratégicos en el desarrollo de la IA tengan incentivos para no cooperar y coordinarse. Por otra parte, muchos investigadores consideran que dejar que la propia industria se autogobierne es un error. Porque esto solo llevaría a un blanqueamiento ético (ethics washing) de ciertas compañías tecnológicas para proyectar una responsabilidad social corporativa falsa. Por consiguiente, el trabajo para implementar una verdadera gobernanza global de la IA es crear un marco normativo efectivo que contemple los riesgos a largo plazo y desarrolle aplicaciones concretas para su implementación por todos los grupos de interés.

La gobernanza global de la IA pretende establecer una seguridad internacional en materia de desarrollo tecnológico basada en el multilateralismo y no caer presa de los distintos modelos de feudalismo digital. Tampoco simpatiza con el modelo del capitalismo de la vigilancia abanderado por los EE.UU., ni el imperialismo digital defendido por China. La gobernanza global ética de la IA es quizá una tercera vía mucho más cercana a los valores humanistas de Europa donde los avances e innovación de la IA no deben comprometer los derechos y privacidad de los ciudadanos. La neutralidad tecnológica es el valor central donde el regulador establece los principios y el mercado los aplica teniendo en cuento los estándares del producto o servicio.

La confianza en la IA ha de ser un pilar central sobre el que sustentar la gobernanza global ética de la IA. Así se podrá garantizar la co-existencia en armonía entre los sistemas de IA y las personas, sin miedo a la explotación u opresión. Una IA centrada en el ser humano donde la libertad de elección y decisión no se vea comprometida y las máquinas, sistemas de IA etc., no supongan un riesgo o una amenaza existencial para la humanidad. Por supuesto, las buenas intenciones deben estar acompañadas de acciones. Y en los últimos tiempos estamos viendo nacer un panorama de guiás de buenas prácticas, reglamentos, planes, e incluso legislación y regulación que permiten hablar de un ecosistema de derechos y garantías en el uso de la tecnología de la IA. En concreto, Europa ha formado el Grupo de expertos de Alto Nivel sobre IA –un consejo independiente formado

por expertos–, un plan de Coordinación sobre IA, donde los estados miembro de la UE han acordado coordinarse y colaborar en políticas públicas, el Libro Blanco sobre IA y, más recientemente –como comentaba en la nota a pie de página número 7–, un reglamento.

Como he manifestado más arriba la cooperación internacional es fundamental si queremos que la humanidad afronte el problema común de la gobernanza de la IA que no puede ser abordado por ningún país por sí solo. Como prioridad para la gobernanza de la IA no solo se encuentran los ámbitos que he mencionado como retos más arriba, también se encuentra el uso militar de la IA que puede crear riesgos para la seguridad internacional. Los Sistemas de Armas Autónomas Letales –armamento dirigido por IA sin necesidad de supervisión y control humano– deben prohibirse para evitar la escalada de conflicto en el futuro. La seguridad de los datos a nivel internacional también debe ser otra prioridad para la gobernanza global de la IA. De la minería de datos y recolección, a la clasificación y asignación de etiquetas, todo el proceso de curación de datos debe ser regulado para prevenir la formación de modelos equivocados que hagan que los agentes decisorios (e.g. grupos de interés relevantes desde consumidores y particulares, hasta organizaciones, empresas y gobierno) formen juicios erróneos.

Ya sabemos que el desarrollo acelerado de la IA y sus aplicaciones demandan urgentemente ética y gobernanza, seguridad, privacidad, transparencia, explicabilidad y equidad para el uso e implementación de sistemas de IA. Pero para ello es clave un desarrollo estandarizado a nivel global propiciado por un nuevo diseño institucional, nuevas leyes, y reglas de gobernanza. Todo marco de gobernanza global ética de la IA no puede ser efectivo si no tiene en cuenta la voz de los usuarios. Por ello es necesario tener en cuenta las experiencias, preocupaciones, intereses y necesidades de los destinatarios finales de la tecnología.

No me gustaría terminar esta sección dedicada al tema central del articulo sin mencionar la gobernanza de una Inteligencia General Artificial (IGA) sin control. Cuando se habla de gobernanza global ética de la IA se sobreentiende cualquier iniciativa para hacer de la IA beneficiosa para la humanidad a través de los gobiernos mundiales, organizaciones internacionales, multinacionales y corporaciones etc., que colaboran conjuntamente para establecer un marco regulatorio y estándares industriales. Esta gobernanza que se sobreentiende no es directa y trata de influir en los productos y servicios que los creadores y desarrolladores tienen permitido hacer, sería inútil si los creadores y desarrolladores fueran capaces de dar lugar a una IGA.

Con una IGA como agente independiente todo intento de regulación por parte de los gobiernos sería en vano. La gobernanza global ética de la IA pasaría a convertirse en predictibilidad, explicabilidad y controlabilidad de una IGA. Este es un escenario poco probable, pero no por ello imposible. Una IGA altamente

capaz, creativa e incontrolable podría ser ingobernable, por ello es perentorio tener cuanto antes el consenso internacional y el marco de regulación y gobernanza global ética de la IA en el estado actual de desarrollo de la tecnología.

4. CONCLUSIONES

La gobernanza global de la IA tiene que ser una de las prioridades de la agenda internacional muy similar a como lo es la acción climática. Los agentes estratégicos (e.g. gobierno, industria, academia y sociedad civil) de la comunidad internacional necesitan trabajar juntos para conseguir obtener lo mejor que puede ofrecer la tecnología de la IA y evitar consecuencias negativas como las que he intentado subrayar en este artículo. En este artículo, he comentado tres grandes retos que considero se han de atajar para establecer una gobernanza global de la IA que beneficie a la sociedad y las personas. Una gobernanza global de la IA puede hacer que los sistemas artificiales y las personas colaboren de manera más fructífera. Una gobernanza global ética de la IA es una exhortación clara y directa para la cooperación internacional ante el desarrollo de tecnologías emergentes y potencialmente disruptivas que permita evitar consecuencias desastrosas que socaven gravemente las perspectivas de las generaciones actuales y futuras.

REFERENCIAS BIBLIOGRÁFICAS

ACEMOGLU, D. y DELL, M., "Productivity differences between and within countries". *American Economic Journal: Macroeconomics*, 2, 2010, 169–88

AHMED, K. "Google's Demis Hassabis –misuse of artificial intelligence 'could do harm'," BBC News. http://www.bbc.com/news/business-34266425 2015

AI HLEG, 2019. *ETHICS GUIDELINES FOR TRUSTWORTHY AI*. [ebook] Brussels: European Commission. Disponible en: <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> [Accedido 28 Julio 2021].

ARAL, S., *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health--And How We Must Adapt*. New York. Currency. 2020

ARUN, C., "AI and the Global South: Designing for Other Worlds" En: DUBBER Marcus, PASQUALE Frank y SUNIT Das (ed.) *The Oxford Handbook of Ethics of AI*. Oxford. Oxford University Press. 2020, 589-607

BALDWIN-PHILIPPI, J., The Myths of Data-Driven Campaigning. *Political Communication* 34, 2017, 627-633

BOSTROM, N., *Superintelligence: Paths, Dangers, Strategies*. Oxford. Oxford University Press. 2014.

BRYNJOLFSSON, E.; ROCK, D. y SYVERSON, C. (2017), "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics". NBER Working Paper 24001 [https://www.nber.org/papers/w24001]

DAFOE, A., "AI Governance: Opportunity and Theory of Impact". Effective Altruism Forum. 2020. Recuperado el 29/07/2021 https://forum.effectivealtruism.org/posts/42reWndoTEhFqu6T8/ai-governance-opportunity-and-theory-of-impact

DANAHER, J., *Automation and Utopia: Human Flourishing in a World without Work*. Mass. Cam. Harvard University Press. 2019.

DE LONG, J., "Estimating World GDP, One Million B.C. – Present"., 1998. Recuperado el 06/04/2021 https://delong.typepad.com/print/20061012_LRWGDP.pdf

DE SOUSA SANTOS, B., "Epistemologies of the south and the future". From the European South, 1, 2016, 17-29.

DURC Policy. (2012), "United States Government Policy for Oversight of Life Sciences Dual Use Research of Concern". Recuperado Abril 2, 2021 de http://www.phe.gov/s3/dualuse/Documents/us-policy-durc-032812.pdf

DWORK C. et al., "Fairness through awareness". *Proceedings of the 3rd Innovations in Theoretical Computer Science Conferences ACM*, 2012, 214-226

DWORK, C. y ROTH, A., *The Algorithmic Foundations of Differential Privacy*. Delft. Now Publishers. 2014.

FREY, C. y OOSBORNE, M., "The future of employment: How susceptible are jobs to computerization?" University of Oxford, 2013. [https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

FRIEDMAN, B. et al. "Value Sensitive Design: Theory and Methods", University of Washington Technical Report, No. 2–12, 2002.

GRICK-HLACA, N. et al., "The case for process fairness in learning: Feature selection for a fair decision making". *NIPS Symposium on Machine Learning and the Law*. Vol. 1, 2016, 2

HARDIN, G., "The tragedy of the commons". *Science*, 162, 1968, 1243-1248

HARDT, M. et al., "Equality of opportunity in supervised learning". *Advances in Neural Information Processing Systems*. 2016, 3315-3323

HERRERA, R. y GARCIA-BERTRANDT, R., "The Agricultural Revolutions". En HERRERA, Rene y GARCIA-BERTRAND, Ralph (eds.), *Ancestral DNA, Human Origins, and Migrations*. Oxford. Academic Press, 2018, 475-509.

JOBIN, A.; IENCA, M. y VAYENA, E., "The global landscape of AI ethics guidelines". *Nature Machine Intelligence* 1, 2019, 2522-5839.

KEARNS, M. y ROTH, A., *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford. Oxford University Press. 2019.

LEE, K., *AI Superpowers: China, Silicon Valley, and the New World Order*. Boston. Houghton Mifflin Harcourt. 2018.

OAKLEY, KP., *Man the Tool-Maker*. Chicago, IL: University of Chicago Press. 1959.

O'NEIL, C., *Weapons of Math Destruction*. New York. Random House. 2016.

MANKIW, N. et al., "A contribution of the empirics of economic growth". *The Quaterly Journal of Economics*, 107, 1992, 407-437.

MCAFEE A. y BRYNJOLFSSON E., *Machine, Platform, Crowd: Harnessing our Digital Future*. New York. WW. Norton & Company. 2017.

MCKINSEY GLOBAL INSTITUTE, "Notes from the AI frontier: Modeling the impact of AI on the world economy. Recuperado el 15/04/2021 de https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

MINSKY, M., *Semantic Information Processing*. Cam. Mass. MIT Press. 1968/2003

MONASTERIO ASTOBIZA, A., "Ética algorítmica: Implicaciones éticas de una sociedad cada vez más gobernada por algoritmos". *Dilemata: Revista Internacional de Éticas Aplicadas*. 24, 2017,185-217.

PARISER, E., *The Filter Bubble: What The Internet Is Hiding From You*. London. Viking. 2012.

PASQUALE, F., *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cam. Mass. Harvard University Press. 2015.

PINKER S., *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. Viking. New York. 2018.

PLEISS, G. et al., "On fairness and calibration". *Advances in Neural Information Processing Systems* 52017, 680-5689.

RAMIÓ, C. *Inteligencia Artificial y Administración Pública: Robots y Humanos Compartiendo el Servicio Público*. Madrid. Catarata 2019.

RENDA, A. "Artificial Intelligence: Ethics, governance and policy challenges". *CEPS Task Force Report*, 2019, ISBN 978-94-6138-716-5.

ROSER, M. y RITCHIE, H., "Child & Infant Mortality". Publicado online en OurWorldInData.org. Recuperado de 'https://ourworldindata.org/child-mortality' [Fuente online]

TAVANI H., "Philosophical theories of privacy: Implications for adecuate online privacy policy". *Metaphilosophy*. 38, 2007, 1-22.

GUARDIAN, The, "Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach". Accedido Abril 3, 2021 https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

WARREN, S. y BRANDEIS, L., "The right to privacy". *Harvard Law Review*, 4, 1890, 193- 220.

WILLIAMS, J., *Stand out of our Light: Freedom and Resistance in the Attention Economy*. Cambridge. Cambridge University Press. 2018.