Ética algorítmica: Implicaciones éticas de una sociedad cada vez más gobernada por algoritmos

Aníbal Monasterio Astobiza
UPV/EHU- IFS/CSIC
anibal.monasterio@ehu.eus

Algorithmic Ethics: Ethical Implications of a Society Increasingly Governed by Algorithms

Resumen: Necesitamos monitorizar y cuantificar el impacto moral de los algoritmos en la vida de las personas. Los algoritmos toman decisiones en nuestro nombre discriminando sobre la base de nuestros ingresos, raza, o sexo. En esta era donde la sociedad cada vez más está gobernada por los algoritmos, un análisis ético -y control de las múltiples formas con las que un algoritmo puede causar daño- es un deber. En este escrito, presentamos una definición operativa de algoritmo, algunos casos paradigmáticos de daño causado por algoritmos y diferenciamos las distintas áreas éticas dedicadas al análisis de la transformación tecnológica y digital del mundo. Finalmente, terminamos con algunas recomendaciones para crear algoritmos éticos.

Palabras clave: ética, algoritmos, ética algorítmica, discriminación, Big Data

Abstract: We need to monitor, track and quantify the moral impact of algorithms on people. Algorithms can take decisions on our behalf discriminating us on the basis of our income, ethnic affinity, or sex. In this era where society is increasingly being governed by algorithms, an ethical analysis -and control over the many ways an algorithm can harm people- should be of paramount importance. In this paper, we present an operative definition of the notion of algorithm, some paradigmatic cases of harm caused by algorithms and the difference among the diverse ethical branches dedicated to the analysis of risks and opportunities coming from the technological and digital transformation of the world. We conclude with some recommendations to build ethical algorithms.

Keywords: ethics, algorithms, algorithmic ethics, discrimination, Big Data

1. ¿Qué es un algoritmo?

La clásica definición de algoritmo dice que un algoritmo es una lista de instrucciones que lleva directamente a un usuario a una respuesta o resultado particular dada la información disponible (Steiner 2012). Por ejemplo, un algoritmo puede ser un árbol decisorio que dada la información sobre la temperatura, el viento, el día del año, si llueve o no, si hace sol o no etc., me diga qué chaqueta escoger de mi armario. Sin embargo, el concepto "algoritmo" significa muchas cosas para mucha gente y si además tenemos en cuenta que ha entrado en el lenguaje y cultura popular, el término y concepto se ha vuelto más ambiguo y vago si cabe. Una definición más técnica es de Robin Hil cuando dice que un algoritmo "es un constructo matemático con una estructura de control finita, abstracta y efectiva de acción imperativa para cumplir un propósito dada una serie de criterios" (Hil 2015, 39). Aún así, una definición más comprehensiva y operativa que permita entender y englobar las distintas instanciaciones de un algoritmo que puede estar

Agradezco el patrocinio del Gobierno Vasco para desarrollar una beca posdoctoral de investigación en el Uehiro Centre for Practical Ethics de la Universidad de Oxford y a esta última institución su cálida acogida. Este trabajo se ha realizado en el marco del proyecto de investigación KONTUZ!: "Responsabilidad causal de la comisión por omisión: Una dilucidación ético-jurídica de los problemas de la inacción indebida" (MINECO FFI2014-53926-R); el proyecto de investigación: "La constitución del sujeto en la interacción social: identidad, normas y sentido de la acción desde la perspectiva de la filosofía de la acción, la epistemología y la filosofía experimental" (FFI2015-67569-C2-2-P)"; y el proyecto de investigación "Artificial Intelligence and Biotechnology of Moral Enhancement. Ethical Aspects" (FFI2016-79000-P)"

Received: 20/04/2017
Accepted: 12/05/2017

ISSN 1989-7022

ISSN 1989-7022
DILEMATA, año 9 (2017), nº 24, 185-217

detrás de la conducción de un vehículo autónomo, plataforma online, agentes o sistemas de Inteligencia Artificial (IA de ahora en adelante) etc., es la siguiente: *un código software que procesa un conjunto limitado de instrucciones*.

La palabra algoritmo viene de Abu Abdulah Mihamad ibn Musa AlKhwarismi, un matemático persa del siglo IX que escribió el que es considerado por los historiadores de la ciencia el primer libro de algebra: Al-Kitab al-Mukhtasar fi Hisab al-jabr wa l-Muqabala (*Compendio de Cálculo por Compeción y Comparación*). El mismo nombre de algebra viene directamente de una palabra del título del libro: al-jabr. Tan pronto los escolásticos y filósofos medievales empezaron a diseminar la obra de Al-Khwarismi la traducción de su nombre por "algorismo" pronto empezó a describir cualquier método sistemático o automático de cálculo. Los algoritmos forman parte esencial de las ciencias de la computación, informática, ingeniería o IA. Tanto es así que en el primer curso de ingeniería en las universidades norteamericanas es costumbre que se les pida a los nuevos alumnos que creen o diseñen un algoritmo para el juego "piedra, papel o tijera" (Steiner 2012).

La revolución algorítmica o lo que la autora Sherry Turkle (2012) ha llamado "el horizonte robótico" (la progresiva introducción de la tecnología y las maquinas en todas las facetas de la vida que hace que esperemos y confiemos más en ellas que en las propias personas) tiene uno de los grandes desafíos y amenazas en el hecho de que los algoritmos detrás de las máquinas y la tecnología se han vuelto cada vez más complejos. Las personas estamos perdiendo la capacidad de entender cómo funcionan y cómo anticipar comportamientos inesperados o brechas en su seguridad. Como ejemplo, señalar que las líneas de código programado que se necesitaron para poner al ser humano en la Luna en 1969 fueron 145.000 y en el año 2015 las líneas de código programado para gobernar Google fueron de 2 billones. Los sistemas algorítmicos son laberintos incomprensibles, en muchos casos hasta para los ingenieros, matemáticos y físicos que las escribieron.

Para contextualizar y entender la dependencia algorítmica de nuestras sociedades (el gobierno de los algoritmos o algoritmocracia) y la visión algorítmico-céntrica de la vida y del trabajo hay que saber que la historia reciente de los algoritmos, la computación y automatización de procesos, tiene su comienzo en los mercados de acciones de Wall Street, New York. Hasta bien entrado la década de los años 70 del siglo pasado la forma de comprar y vender acciones era prehistórica a cómo

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

se viene haciendo en la actualidad. Brokers pasados de adrenalina, gritando y gesticulando en el parquet, con los teléfonos en una mano y en la otra papeles, eran la estampa habitual. Pioneros como Thomas Peterffy empezaron a introducir algoritmos para decidir qué acciones tenían las mejores ventajas y menores riesgos y en el parquet empezó a verse menos gente. Como la velocidad de la información es crucial para obtener una ventaja competitiva frente a tus rivales, por la misma época de forma secreta Daniel Spivey y David Barksdale fundaron Spread Networks para construir un cable que cruzará valles, montañas y otros accidentes geográficos para conectar, con un cable de fibra óptica, Chicago con Nueva York. Se empezó a fraguar la toma de Wall Street por los algoritmos y ciencias de la computación y apareció un nuevo tipo o figura en el mundo financiero: los *Quants* (gente con mente analítica, frecuentemente licenciados en matemáticas, física o ingeniería) que sustituyeron de un plumazo a licenciados en economía con un MBA, y las decisiones ya no las tomaba ningún broker, o ser humano, sino algoritmos. La gran recesión del 2008 o la crisis crediticia en buena parte es causa de los instrumentos financieros matemáticamente complejos introducidos por esta nueva estirpe de financieros donde el algoritmo era la nueva herramienta de trabajo para predecir el comportamiento de los mercados. Pero muy pocos conocen que el día 6 de mayo del 2010 se produjo una caída de casi 1000 puntos en el mercado de valores (índice Dow Jones) cuyas consecuencias de no corregirse hubieran llevado al abismo al sistema financiero internacional. Por la tarde del 6 de mayo del 2010 la caída era de 4% debido a las preocupaciones de la crisis de deuda europea. A las 14:32 una gran vendedor, un complejo de fondos mutuos, inicio un algoritmo para disponer de un gran número de E-Mini S&p 500, contratos a futuros para ser vendidos a un ratio de venta vinculado a la medida minuto a minuto con la liquidez de la bolsa. Estos contratos fueron comprados por algoritmos de negociación de alta frecuencia que están programados para rápidamente eliminar sus posiciones vendiendo estos contratos a otros agentes.

Los monitores del *Dow Jones Industrial Average*, quizá el índice bursátil más seguido del mundo, mostraba en pantalla una caída de 998 puntos. Parecía como si el índice hubiera sido *hackeado*. Cerca de 1 trillón de dólares se volatilizaron en el éter electrónico (CFT&SEC 2010). A las 14:45 un sistema de control interrumpió o, mejor dicho, detuvo las operaciones del algoritmo de negociación de alta frecuencia. Cuando de nuevo el algoritmo se inició los precios se estabilizaron, pero se había

ISSN 1989-7022
DILEMATA, año 9 (2017), nº 24, 185-217

perdido un trillón de dólares con efectos cascada en todas las bolsas con precios absurdos (un centavo la acción o 100.000 dólares). Cuando el mercado cerró los representantes de las bolsas se reunieron con los reguladores y decidieron que todos los intercambios que se habían ejecutado a precios mayores de 60% de los precios pre-crisis, fueran cancelados. Todavía no hay un consenso sobre las causas reales de lo que se bautizó como el "Flash Crash". Michael Lewis (2014) narra estupendamente esta historia fatídica y de suspense en el seno de Wall Street sobre cómo los algoritmos programados e implicados en el Flash Crash tomaron el control de las finanzas globales. Los algoritmos utilizados en la bolsa para comprar y vender acciones cumplen un buen servicio incrementando la liquidez y la eficiencia del mercado. En el incidente de Wall Street de mayo del 2010, conocido como Flash Crash, huelga decir que aunque los algoritmos contribuyeron a la crisis, también a su resolución. Los algoritmos que se han convertido en el estándar del mundo de las finanzas y los negocios y han tomado Wall Street, tienen otras aplicaciones beneficiosas para la gente y el mundo. Mejorar la emisión de los programas de radio, mejores servicios de atención al cliente, mejor servicio de inteligencia de los países para luchar contra las amenazas de terrorismo global y mejores herramientas para detectar el cáncer y otras enfermedades.

Pero los algoritmos también son una caja de Pandora. Los algoritmos y su aplicación en sistemas de IA y/o robótica para la automatización de procesos desplazan a miles de trabajadores actualmente. Esta creciente automatización del trabajo y de la vida cuyo responsable directo son los algoritmos y software desarrollados por la investigación en IA, creará una fractura sin precedentes en el mercado laboral convirtiendo en inempleables a millones de personas a medio y largo plazo (Frey y Osborne 2013; Ford 2015). Los robots o sistemas automatizados controlados por algoritmos son más rápidos que nosotros, más baratos y además no se cansan. Las consecuencias de las actuaciones de los algoritmos pueden ser catastróficas como el episodio del Flash Crash nos ha enseñado. Por ello, se hace condición *sine qua non* la necesidad de una ética algorítmica en la era del Big Data, la IA, los coches sin conductor y la automatización del trabajo y la vida. Eliezer Yudkowsky (2008) propone diseñar una "IA *amigable*". Él entiende esto como la idea de que "valores de tolerancia y respeto por el bienestar de los seres humanos sean incorporados como elementos nucleares en la programación de agentes de IA". Es decir, agentes artificiales de todo tipo que tengan implementados valores humanos que conviertan

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

a la IA en un factor positivo en lugar de uno negativo a la hora de contemplar riesgos globales. Nick Bostrom (2014), por su parte, nos habla de técnicas de control y seguridad en IA avanzada que prevenga un uso instrumental de los seres humanos por parte de una Super-Inteligencia.

Pero, ¿cómo se consigue programar moralmente sistemas de IA o robots?, ¿cómo se consigue introducir la ética en el silicio? Una cosa debemos tener clara. La pregunta sobre si la tecnología es buena o es mala -que inunda las páginas de opinión de los periódicos, la blogosfera y otros medios- es demasiado simple y está mal enfocada. No es el *quid* de la cuestión. La tecnología ya está aquí y además está por todas partes. La verdadera pregunta, como sugiere el teórico de nuevos medios Douglas Rushkoff, es: "¿queremos dirigir la tecnología o queremos que ella nos dirija a nosotros?: Programar o ser programados". Wendel Wallach en su último libro (2015) recogiendo una famosa cita de Christian L. Lange lo expone de una forma más dramática: "La tecnología es una buena sirvienta, pero un amo peligroso". Tradicionalmente, los filósofos de la tecnología y los éticos han actuado como críticos y analistas pasivos ante el avance y desarrollo tecnológico. Sin embargo, una nueva generación de filósofos está militando en el activismo global poniendo el énfasis, y ayudando, en la introducción de la sensibilidad hacia los valores humanos a la hora de diseñar sistemas de IA, (ro)bots o cualquier software que tenga un potencial impacto en las personas. En palabras de Yuval Noah Harari (2016) la regulación ética de la transformación tecnológica y digital, e incluso de los avances en biotecnología como la ingeniería genética, en un mundo interconectado como el que vivimos debe ser global. Porque podemos encontrarnos un país donde el control ético previene, por principio de precaución, no investigar ni aplicar ciertas tecnologías, pero encontrarnos otro país donde el control ético es más laxo. Esta situación genera una posición de vulnerabilidad (lose-lose game). Si un país puede contar con una tecnología que le proporciona una ventaja competitiva con respecto a otros, nadie querrá quedarse atrás y se invertirá en dicha tecnología también, conduciendo a todos a una carrera de desarrollo tecnológico cuyas disrupciones en las personas y sociedad pueden ser devastadoras y, por supuesto, poco éticas (zero-sum game). El cambio de paradigma de autoridad delegando cada vez más decisiones de los seres humanos a las máquinas, incluidos los algoritmos, sobre las vidas personales, asuntos económicos o políticos, es posible que sea un síntoma de la complejidad del entorno en el que vivimos. Nuestros cerebros evolucionaron

ISSN 1989-7022
DILEMATA, año 9 (2017), nº 24, 185-217

en un entorno físico concreto de la sabana africana hace más de 500.000 años con una cantidad determinada de información y datos. En el entorno del siglo XXI, un espacio de información abundante, nuestra capacidad natural de procesamiento de la información es limitada. Necesitamos los algoritmos. Pero esto no es óbice para que aceptemos cualquier implementación de algoritmos poco ética. Para ver la necesidad de la auditoria de los algoritmos por su potencial impacto en la sociedad y la vida de las personas en el siguiente apartado presentamos casos donde los algoritmos amenazan valores básicos de las personas e, incluso, amenazan la misma organización democrática de nuestras sociedades. Sin embargo, sirva de limitación de responsabilidad que no queremos dar ninguna impresión apocalíptica, tecnófoba o ludista. No nos damos por aludidos. Al contrario. Abrazamos y damos la bienvenida a la tecnología. Pero a una tecnología ética (win-win game). Porque como veremos a continuación en varios ejemplos y casos la tecnología puede actuar no solo en contra de la legalidad, sino actuar de manera inmoral.

2. Las implicaciones éticas de una sociedad cada vez más gobernada por algoritmos

La ética de la IA, ética de datos y ética algorítmica son diversos campos de estudio -sus diferencias y similitudes las veremos en el siguiente apartado- que nacen como consecuencia de la necesidad de afrontar los problemas, implicaciones y desafíos que plantean los avances en IA, TICs, economía y mundo digital; para las personas, comunidades y sociedades a nivel global. El interés por mantener un análisis, control y diseño ético de la tecnología digital surge de la omnipresencia de Internet y tecnologías afines. Hace unos diez años se empezaba hablar de IoT (acrónimo inglés para *Internet de las Cosas*), la visión de un mundo donde distintos dispositivos estarían conectados entre sí y simultáneamente a Internet. Esa visión se ha quedado desfasada, porque ahora el desarrollo tecnológico permite el "*Internet de Todo*", (IoE acrónimo en inglés), donde esencialmente no hay nada que no esté afectado por la presencia de agentes de IA: coches, casas, dispositivos electrónicos, electrodomésticos, ropa, accesorios, entorno urbano, alimentos etc. Esta nueva conectividad "máquina a máquina" (M2M) permitirá la mejora de las aplicaciones electrónicas y revolucionará las comunicaciones permitiendo la medición y lectura inteligente de los contadores de luz y agua de nuestras residencias, de la calidad

ISSN 1989-7022 DILEMATA, año 9 (2017), nº 24, 185-217

del aire en nuestras ciudades, interoperabilidad de dispositivos, móviles etc. Pero al mismo tiempo pueden generar nuevas vulnerabilidades y amenazas si estas tecnologías caen en las manos equivocadas y son "*hackeadas*" por organizaciones criminales (Goodman 2016). Este nuevo espacio digital que propicia el "*Internet de todo*" puede ser utilizado para espiar cualquier movimiento de las personas, teniendo en cuenta que se espera que en los próximos años más de 75 billones de objetos estén conectados entre si y a Internet. La empresa Cisco en un reciente informe ya incluso estima que hay más objetos conectados entre sí y a Internet que personas con dispositivos conectados a Internet (Evans 2011).

Casi seguro, esto transformará nuestra forma de comunicarnos, pero también incrementará las vulnerabilidades convirtiendo la calidad e integridad de la información en algo fundamental para la seguridad y otros valores sociales. Y por calidad e integridad de la información se entiende no solo que la información no sea degradada o corrompida (este pasado año 2016 hemos asistido al uso de la palabra "posverdad" -mejor decir simplemente *bulo* y no utilizar el neologismo que no es ni siquiera una palabra en su idioma original- donde el aspecto fáctico de la información ya no importa y el crecimiento de "burbujas de filtro" o "cámaras de eco" amplifican noticias con sesgos de auto-interés o simplemente se difunden falsedades interesadas con fines oscuros sin poder verificarse), sino también que se proteja y asegure esa información, no se acceda sin consentimiento (privacidad) y que tenga un propietario bien identificado. Ya no podemos vivir sin IA. Vivimos dentro de la IA en todas las esferas de nuestra vida. La cuestión es que estos agentes o software de IA están en todas partes, son ubicuos, pero están de forma invisible y hace que demos las cosas por sentadas o seguras sin saber que detrás de ellas está operando la IA. Cuando vas a pedir un crédito, un software de IA hace una valoración de tu aval financiero, cuando compras billetes de avión, un algoritmo te campara precios, un vehículo sin-conductor autónomo está gobernado por un agente de IA, cuando Spotify, Itunes u otras plataformas de música te recomienda un nuevo disco o artista, lo hace un algoritmo y software de IA etc.

Hay que entender y saber quién decide, quién gobierna y quién distribuye la información; esencial para las libertades individuales y nuestras democracias. Las implicaciones éticas de una sociedad cada vez más gobernada por algoritmos obligan a la industria, academia e instituciones públicas a buscar alianzas para crear una gobernanza transparente,

ISSN 1989-7022
DILEMATA, año 9 (2017), nº 24, 185-217

ética y justa de la caja de Pandora que puede ser la IA. En los últimos tiempos y con la misión de llevar a cabo este propósito varios gigantes tecnológicos de Silicon Valley, instituciones académicas y fundaciones han firmado varios acuerdos.

Uno de estos acuerdos es la Alianza para beneficiar a la gente y la sociedad que los gigantes tecnológicos Google, Amazon, Microsoft, IBM, Facebook y Deep Mind (solo se echa en falta a Apple) crearon el pasado septiembre del 2016 para tomar ventaja de las promesas de la IA (https://www.partnershiponai.org/). Su misión no es la de ser otro lobby más u organización que sirva para presionar las políticas y favorecer la regulación del mercado hacia sus intereses. Más bien tiene como objetivo apoyar buenas prácticas en la investigación en IA, el entendimiento público de la IA y crear un debate público sobre las implicaciones éticas de la IA con expertos, industria, gobiernos y público en general. Nada más conocerse la noticia de esta alianza el crítico y analista tecnológico Evgeny Morozov puso un tuit caracterizado por su cinismo e ironía fina que decía literalmente: @evgenymorozov --- "Solo una alianza ética entre Goldman Sachs, Deutsche Bank, and JP Morgan sería más irrisoria". A Morozov no le falta razón, por otra parte. Otra iniciativa es *OpenAI* una organización sin ánimo de lucro apoyada por Elon Musk, flamante hombre de negocios y empresario del sector tecnológico, dedicada al análisis ético de la IA. Como bien dan cuenta Julia Powles y Carissa Veliz (2016) en un artículo para *The Guardian* donde exponen las maniobras de colusión y/o monopolio de las empresas tecnológicas de Sillicon Valley para ordeñarnos como "vacas digitales" y así extraernos todos los datos posibles que solo ellos contralaran como señores feudales; no solo es necesario una legislación fuerte como el Reglamento Europeo de Protección de Datos (2016)[1], sino la ética y la cultura.

Con la intención de tener en cuenta a la ética y la cultura otro acuerdo reciente a fecha de escritura de este artículo, es el que han firmado en enero de 2017 un grupo de fundaciones, inversores e instituciones académicas llamado "*Ethics and Governance of Artificial Intelligence Group*". Su objetivo, fomentar desde distintas perspectivas la ética de la IA. En los últimos tiempos estamos asistiendo un renacimiento de la IA con importantes hitos. Las fases históricas de la investigación en IA han estado sucedidas de altos y bajos, y frecuentemente a los estadios más bajos se les ha venido en llamar "inviernos". La euforia que atravesamos actualmente con declaraciones de los más importantes científicos e incluso hombres de empresa (por ejemplo, Stephen Hawkins o Bill Gates) en torno a la amenaza que supondría una IA descontrolada debido a los importantes avances alcanzados -hasta el punto que muchos de ellos hablan de que

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

una IA descontrolada o *singularidad* tecnológica (punto temporal donde las máquinas puedan pensar de un modo más eficiente que los seres humanos) es la mayor amenaza o problemática para la humanidad- nos dice muy claramente que estamos en una nueva "primavera" de la IA, un nuevo estadio "alto". Muchos autores argumentan que la automatización de la vida y el trabajo, la robotización y la aplicación de la IA causarán importantes disrupciones en la economía.

No debemos ser catastrofistas, la IA puede dotarnos de herramientas maravillosas para mejorar el día a día de las personas en todo el mundo. Pero su desarrollo continuado también se presenta con grandes retos. Para que el desarrollo de la IA se haga siguiendo las mejores buenas prácticas y principios el *Future of Life Institute* organizó la segunda conferencia sobre IA *beneficiosa* en Asilomar, California. Allí se reunieron investigadores en IA de la academia y de la industria y diversos autores relevantes en el mundo de la economía, el derecho, la ética y la filosofía durante cinco días, del 3 de enero al 9 de enero de 2017. Durante estos cinco días el programa consistió en sesiones de comunicaciones, charlas y presentaciones. Los allí presentes (entre ellos Nick Bostrom de Oxford, Yann Lecunn de Facebook, Stuart Rusell autor del manual en IA más usado, Eric Schmidt de Google, Daniel Kahneman premio Nobel de economía… muy larga la lista como para mencionarlos a todos aquí, pero podemos decir que no faltó nadie que actualmente tenga algo que decir en el campo de la reflexión ética de la IA, así como de la investigación científica académica o industrial) establecieron 23 principios (hay un total de 847 firmantes (investigadores en robótica e IA) y 1222 que apoyan dichos principios) englobados en tres grandes áreas: temas de investigación, ética y valores, y temas de largo-plazo.

Los 23 principios son los siguientes:

Temas de investigación

1) Objetivo de investigación: el objetivo de la investigación en IA es crear inteligencia, pero inteligencia beneficiosa.

2) Financiación de la investigación: la inversión en IA debe estar acompañada de inversión en su uso beneficioso, incluida la financiación para cuestiones problemáticas relacionadas con la ciencia de la computación, economía, derecho, ética y estudios sociales.

3) Un vínculo entre la ciencia y la política: debe haber un dialogo constructivo entre investigadores en IA y políticos.

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

4) Cultura de investigación: una cultura de cooperación, confianza y transparencia debe fomentarse entre los investigadores en IA.

5) Evitar la competición: los equipos que desarrollen sistemas de IA deben cooperar activamente.

Ética y valores

6) Seguridad: los sistemas de IA deben ser seguros durante su ciclo vital y verificables en su seguridad.

7) Transparencia de funcionamiento: si un sistema de IA causa un daño debe ser posible su identificación y corrección.

8) Transparencia judicial: cualquier sistema de IA implicado en una toma de decisiones judicial debe proveer de una explicación satisfactoria para ser auditada por una autoridad humana competente.

9) Responsabilidad: diseñadores y casas manufactureras de sistemas avanzados de IA son grupos de interés en las implicaciones morales de su uso.

10) Alineación de valores: sistemas de IA autónomos deben ser diseñados de tal forma que sus objetivos y comportamientos se alineen con valores humanos.

11) Valores humanos: los sistemas de IA deben estar diseñados para ser compatibles con los ideales de la dignidad humana, derechos, libertades y diversidad cultural.

12) Privacidad personal: la gente tiene el derecho de acceder, tratar y controlar los datos que generan, dada la gran capacidad de los sistemas de IA de analizar y utilizar esos datos.

13) Libertad y privacidad: la aplicación de la IA a los datos personales no puede coartar la libertad percibida o real de la gente.

14) Beneficio compartido: las tecnologías de IA deben beneficiar y empoderar a cuanta más gente mejor.

15) Prosperidad compartida: la prosperidad económica que pueda traer la IA debe ser compartida con toda la humanidad.

16) Control humano: los seres humanos elijen cómo y si delegar decisiones a sistemas de IA para realizar objetivos humanos.

17) No subversión: el poder conferido a un sistema de IA no debe utilizarse para subvertir el orden civil o social del que la sociedad depende.

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

18) Carrera de IA: una carrera armamentística en sistemas de armas autónomas letales debe evitarse.

Temas de largo-plazo

19) Precaución en las capacidades: de no haber consenso debemos evitar cualquier presunción sobre los límites de las capacidades de sistemas de IA futuros.

20) Importancia: IA avanzada puede representar un profundo cambio en la historia de la humanidad por ello debe ser planeada y gestionada con la mayor precaución posible.

21) Riesgos: cualquier riesgo contemplado de un sistema de IA, tanto catastrófico como existencial, debe ser acompañado de esfuerzos para mitigar y gestionar su impacto.

22) Auto-mejora recursiva: cualquier sistema diseñado para replicarse o auto-mejorarse debe estar sujeto a estrictas normas de control y seguridad.

23) Bien común: la superinteligencia solo debe ser desarrollada al servicio de amplios ideales éticos y para el beneficio de toda la humanidad en lugar de un estado u organización.

Estos principios salidos de Asilomar son muy importantes para el desarrollo de una IA segura. Son análogos a la guía de principios que ha dominado el debate en biotecnología, y todavía se siguen discutiendo, desde que los consorcios público y privado anunciaron la secuenciación del genoma humano.

Junto a estos principios en torno a la creación de IA segura y las alianzas entre fundaciones, instituciones académicas y empresas podemos sumar iniciativas como el Comité Internacional para el Control de Armas Robóticas (ICRAC acrónimo en inglés) fundado por académicos (Altmann, Asaro, Sharkey y Sparrow) cuya misión, entre otros objetivos, es la prohibición del desarrollo y uso de estos sistemas para uso militar; las máquinas no deben tomar la decisión de matar a gente. Recientemente, también se ha creado la fundación *Responsible Robotics* (http://responsiblerobotics.org/) que busca la rendición de cuentas de la innovación humana detrás de los robots. La ética algorítmica en una sociedad cada vez más gobernada por algoritmos también debe poner énfasis en los usos que los algoritmos tienen en la guerra en el siglo XXI. Tras la Segunda Guerra Mundial podemos asegurar que las tensiones geopolíticas ya no solo están causadas por estados-nación, sino por

ISSN 1989-7022
DILEMATA, año 9 (2017), nº 24, 185-217

empresas multinacionales, actores no estatales y hasta incluso la tecnología. En este nuevo orden mundial caracterizado por el multilateralismo, ya no son solo los combatientes, sino algoritmos y los (ro)bots, los que deciden la diferencia entre la guerra y la paz.

Es muy difícil subrayar todas las diferencias entre las diversas técnicas y recursos de la IA, sistemas de IA, agentes de IA, Aprendizaje Máquina, redes neuronales, algoritmos, robots, sistemas autónomos… pero podemos agruparlos bajo el mismo paraguas: la transformación tecnológica y digital del mundo. A no ser que se especifique cuando hablamos de la transformación tecnológica y digital del mundo nos estaremos refiriendo a la anticipación e identificación de los riesgos y oportunidades de todos estos procesos y avances. En el caso concreto de los algoritmos existen muchas clases de algoritmos y distintas aplicaciones de los mismos para distintos campos y tareas, pero de manera general son tres las propiedades que los algoritmos poseen:

a) Universalidad
b) Opacidad
c) Impacto en la vida de las personas.

La primera característica, la *universalidad*, hace que los algoritmos se vuelvan indispensables en esta era donde la tecnología está presente en todas las esferas de la vida. Un algoritmo puede servir para controlar el tráfico aéreo, las señales de tráfico de una ciudad, seleccionar y recomendar qué película ver o programar el encendido de la calefacción centralizada de tu casa. La segunda característica es la *opacidad*. A pesar de que los algoritmos están en todas partes y gobiernan múltiples esferas de nuestra vida y trabajo; estos están ocultos, son invisibles. No solo son invisibles porque son el software que hay que "arrancar" en el hardware de computadoras, sistemas o artefactos. Son invisibles porque resultan ser inescrutables entre las capas y capas de programación informática. Opacos en el sentido de que son casi herméticos a la interpretación, corrección, y mejora como decíamos más arriba en la sección 1. A día de hoy la complejidad tecnológica y la gran cantidad de líneas de código que han de programarse hacen que ningún ingeniero de ninguna gran compañía tecnológica sepa decirte cuántas líneas de código son básicas para el funcionamiento de sus servicios o productos y ni siquiera descubrir un posible "*bug*". Finalmente, la tercera y última característica es que afectan a la vida de las personas. Esta es, quizá, la propiedad más importante: el impacto en la vida de

ISSN 1989-7022 DILEMATA, año 9 (2017), nº 24, 185-217

las personas. Si combinas todas y cada una de estas tres características, entonces tenemos lo que la matemática y científica de datos Cathy O´Neil llama: "*Armas de destrucción matemática*". Expresión que da título a su reciente libro (O´Neil 2016).

En esta nueva era de los vehículos autónomos o coches sin conductor, el Big Data y la automatización de la vida y el trabajo, junto con el empleo masivo de algoritmos para la toma de decisiones individuales, gubernamentales y sociales; nos dirige a nuevos terrenos de exploración para la ética. Existen múltiples implicaciones éticas de esta nueva sociedad algorítmica en la que vivimos. Para mostrar los riesgos (aunque por supuesto recordemos que existen muchas oportunidades) a continuación se describen una serie de casos donde el uso de "algoritmos amorales" (o por lo menos sin sensibilidad moral, ni social) y las consecuencias que generan, puede no solo resultar en un daño ético; sino ir en contra de principios y valores básicos democráticos. Se divide el impacto de los algoritmos en la vida de las personas en cuatro grandes categorías o dimensiones de discriminación:

(1) Discriminación Social
(2) Discriminación Económica
(3) Discriminación de acceso libre a la información y privación de Libertad
(4) Discriminación y abuso de Control

Cuando procedimientos o protocolos automatizados (algoritmos) deciden por los seres humanos y encima lo hacen de manera sesgada y en contra de derechos y libertades civiles que las personas poseen se produce un fenómeno ético particular: el daño causado tiene difícil identificación para rendición de cuentas y/o responsabilidad, la complejidad de la programación de los algoritmos impide corregir o enmendar, y/o dada la ubicuidad e invisibilidad de los algoritmos uno cree que cualesquiera efectos que produzcan (por muy negativos que resulten ser) hemos de aceptarlos porque así es como son las cosas y nada puede hacerse para evitarlo (conformidad y resignación). La falta de *transparencia/opacidad*, la *complejidad/ubicuidad/invisibilidad* y la *conformidad/resignación* ante los efectos de los algoritmos hace imposible aplicar reglas éticas particulares.

Pero se puede intentar aplicar reglas éticas específicas si se eliminan los mitos de la objetividad y eficiencia total de los algoritmos. Los algoritmos no son NEUTRALES, OBJETIVOS o PRE-ANALÍTICOS. Los algoritmos se enmarcan en un contexto tecnológico, económico, ético, temporal y espacial. El actual sistema económico

ISSN 1989-7022
DILEMATA, año 9 (2017), nº 24, 185-217

neoliberal, pero también la tecnocracia, el capitalismo de la vigilancia (surveillance capitalism) el gobierno de "*los señores del aire*" (Echeverria 1999), el *tecno-solucionismo* (Morozov 2014); constituyen el estado actual de la gestión sociopolítica y sociotecnológica donde los intereses corporativos y privados son reforzados por los algoritmos y la tecnología imponiendo su propia agenda e intereses sin contar con la deliberación, discusión y participación pública democrática de los ciudadanos. Jonathan Taplin (2017) describe muy bien este ethos tecnolibertario presente en las compañías tecnológicas y que se va extendiendo por todas las esferas de la sociedad. Porque los algoritmos no existen independientemente de IDEAS, PRÁCTICAS, INSTRUMENTOS, CONTEXTOS. *Ideas* que los profesionales de la ingeniería informática y ciencias de la computación tienen, muchas veces, incluso, de manera inconsciente en forma de prejuicios, sesgos, estereotipos que de manera flagrante se ven reflejadas en los mismos algoritmos que programan o la tecnología que diseñan. *Prácticas* que institucionalizan ciertos comportamientos retroalimentados en la dinámica que crean las ideas y las mismas prácticas. El estado del arte en ciencia e ingeniería que crea *instrumentos* que, a veces, pueden tener una carga axiológica muy fuerte. Finalmente, los *contextos* generales en los que todos estamos implicados son muy importantes. Contextos sociales que muchas veces nos impiden tomar decisiones totalmente libres a las personas. En cierta medida, como hemos podido ver hasta ahora, los algoritmos controlan y gobiernan en secreto nuestras vidas. En una *algoritmocracia*, se puede decir que vivimos. Cuando *surfeas* por la red para ver si compras unas zapatillas deportivas, cuando elegimos una película en Netflix, una canción en spotify o solicitamos una hipoteca; un algoritmo está detrás de todo el proceso y tendrá la última palabra. Por ello, una ética algorítmica para garantizar un diseño e innovación responsable es normativamente necesaria e imprescindible dado que los algoritmos no son inherentemente justos, ni inteligentes, sino que responden a los intereses del diseñador conduciendo muchas veces a aumentar el impacto negativo y daño en la vida de las personas.

Ahora veamos las cuatro grandes categorías o dimensiones de discriminación algorítmica con casos concretos de daño causado por algoritmos.

2.1. Discriminación Social

Este tipo de discriminación es muy común, pero no por ello deja de tener un gran impacto en la vida de las personas y se refleja en las acciones más cotidianas. El

ISSN 1989-7022 DILEMATA, año 9 (2017), nº 24, 185-217

algoritmo de la página web francesa *Ton prenom* (http://tonprenom.com/bebe) discrimina en contra de ciertos nombres personales y a favor de otros. Los padres primerizos que quieren elegir el nombre para sus hijos pueden recurrir a sus servicios siempre y cuando acepten que dicho algoritmo va a asumir por defecto que deseas evitar un nombre de origen árabe. El algoritmo deja marcada por defecto la opción de *favorecer* un nombre de origen francés, marca por defecto la opción *indiferente* para los nombres de origen inglés o judío, pero marca la opción *evitar* para nombres de origen árabe. En la imagen de abajo se puede ver una captura de la página y las distintas opciones a la hora de elegir un nombre.

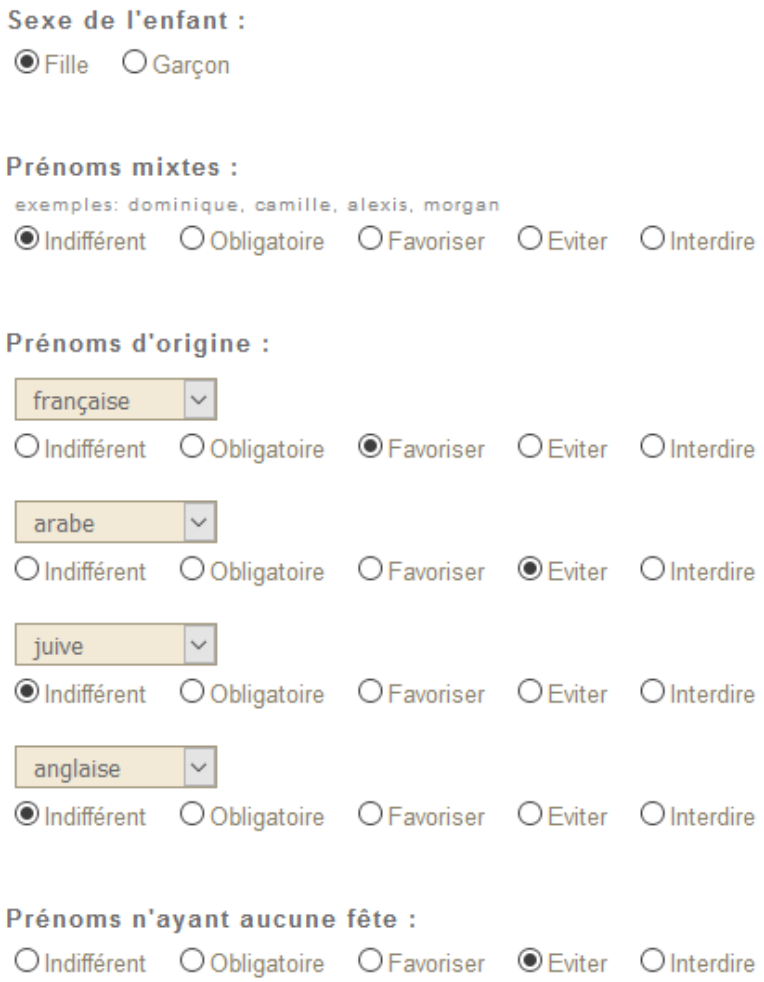

Figura1. Captura de pantalla de la página web (18/05/2017)

Este algoritmo es un caso típico de discriminación social que tiene amplias repercusiones al conducir, a largo plazo, a una menor diversidad cultural y eliminación de la alteridad o inclusión del "otro" en nuestras sociedades. Puede

ISSN 1989-7022
DILEMATA, año 9 (2017), nº 24, 185-217

parecer trivial la elección del nombre que tienen las personas, pero entre todas las cosas que refleja el nombre de una persona, muestra la apertura, inclusividad o chovinismo de una comunidad y, por supuesto, puede ser una expresión de *racialismo* o racismo patente o el indicio de auto-segregación basada en la raza, las rentas, el tipo de trabajo etc. Esta auto-segregación expresa la tendencia individual de las personas a juntarse con aquellos que comparten las mismas, pero que a nivel macro y colectivo lleva a sociedades desiguales, descohesionadas y separadas.

Figura 2. Captura de pantalla de Google translate (18/05/2017)

En la imagen de arriba tenemos el algoritmo de Google Translate, servicio de traducción de la compañía tecnológica Alphabet, que introduce una discriminación o sesgo basado en el género. La traducción es del turco al inglés. "O bir" es un pronombre o artículo de género neutro, pero como podemos observar en la captura de pantalla al traducir del turco "O bir doktor" que tendría un significado aproximado de, "ello doctor", el algoritmo de Google translate asume que "doctor" o médico, solo es una profesión reservada a los hombres. En cambio, cuando se traduce del turco al inglés, "O bir hemsire" que significaría algo así como "ello enfermero/a" el algoritmo de Google translate directamente traduce que la profesión de enfermería solo la ejercen mujeres. He aquí un caso de algoritmo que discrimina sobre la base del género, en otras palabras, es un algoritmo sexista.

ISSN 1989-7022 DILEMATA, año 9 (2017), nº 24, 185-217

2.2. Discriminación Económica

Sobre discriminación económica hay una ingente literatura (Arrow 1972). Los derechos civiles principalmente se tipifican para evitar la discriminación, pero aún así hay evidencias de discriminación en los mercados económicos. Las minorías son discriminadas por el mero hecho de serlo frente a la mayoría. A las minorías (que pueden ser discriminadas por su condición, raza, etnia, sexo…) se les ofrece, o esperan obtener, mucho menos que lo que obtienen las mayorías. La discriminación cuando la realizan los algoritmos es mucho más cruel porque, como decíamos más arriba, las características generales que tienen los algoritmos hacen que el daño que causan sea difícil de corregir, de identificar y/o asignar responsabilidades. Además generan conformismo o aceptación del daño por parte de quienes lo sufren y son víctimas.

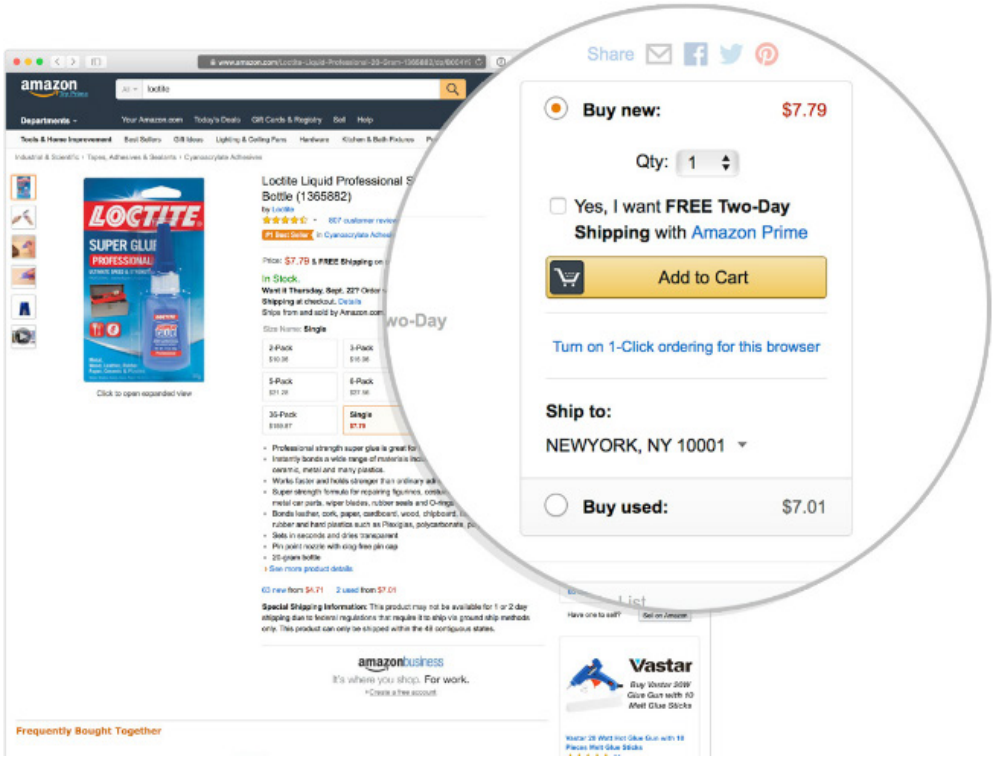

Figura3. Imagen tomada de Propublica.

En la captura de pantalla o imagen de arriba vemos el algoritmo de asignación de precios de la plataforma Amazon que está discriminando económicamente, afectando al bolsillo de las personas. En la plataforma Amazon distintos vendedores ponen distintos ítems a la venta para su compra y consumo. Normalmente, este algoritmo

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

busca los mejores precios de los productos o ítems, es decir, filtra o selecciona en función de varios criterios, uno de ellos el precio más barato. Para el caso de pegamentos industriales, como vemos en la imagen, el algoritmo ha seleccionado el producto Loctite™ con un botón de opción de compra y un importe total visible. Lo que el algoritmo no te indica es que no se incluyen gastos extra de coste de envío. Por lo que el importe total, finalmente, se incrementa. Aquí estamos siendo testigos de un algoritmo que discrimina económicamente.

2.3. Discriminación de acceso libre a la información y privación de Libertad

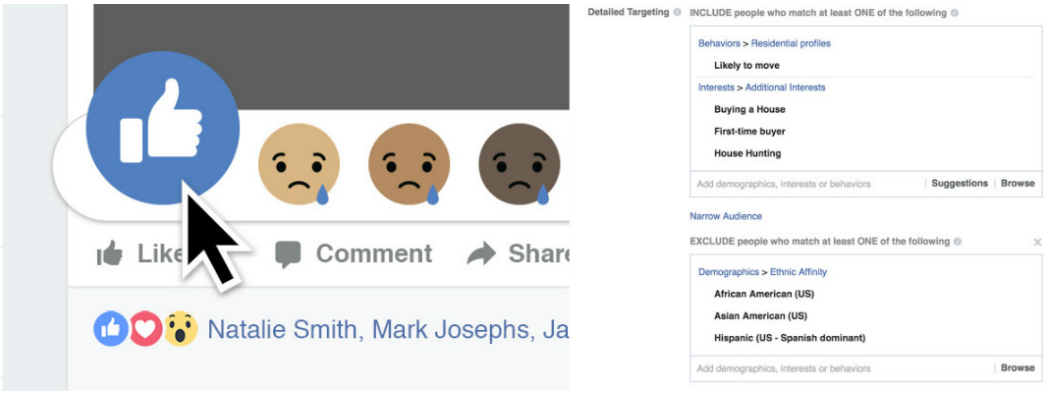

Figura 4. Imagen tomada de Propublica.

En la imagen de arriba tenemos una captura de pantalla que muestra el algoritmo de la red social de Facebook que una vez entras en el proceso de querer poner o crear un anuncio de alquiler y/o compra de vivienda en Facebook, selecciona de entre la población usuaria de Facebook a quién quieres que se dirija, o en definitiva, quién quieres que vea el anuncio. Pero en un paso de creación del anuncio inmobiliario el algoritmo de Facebook contempla una opción de selección de la población objeto al que vas a dirigir el anuncio en función de la raza. Lo que los ingenieros de la compañía llaman: "afinidad étnica". Sin lugar a dudas, la expresión "afinidad étnica" es un eufemismo para *raza*. De acuerdo con el grupo de investigación periodístico americano Propublica, cuya misión es la defensa de las libertades y derechos civiles, este algoritmo de Facebook está violando varias leyes de acceso a la información libre por parte de todo ciudadano, incluyendo leyes de derecho a la vivienda y no discriminación sobre la base de condición, credo, sexo o raza.

Pero quizá de entre todos los posibles casos de discriminación algorítmica el más grave es el que se comete con ciertos grupos de ciudadanos por parte de empresas

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

que utilizan protocolos automatizados (algoritmos) para determinar el riesgo de reincidencia en la comisión de delitos. Más abajo se muestra una gráfica de barras que representa los distintos ítems o indicadores de puntuación de riesgo contemplados para el grupo de acusados de raza blanca y el grupo de acusados de raza negra.

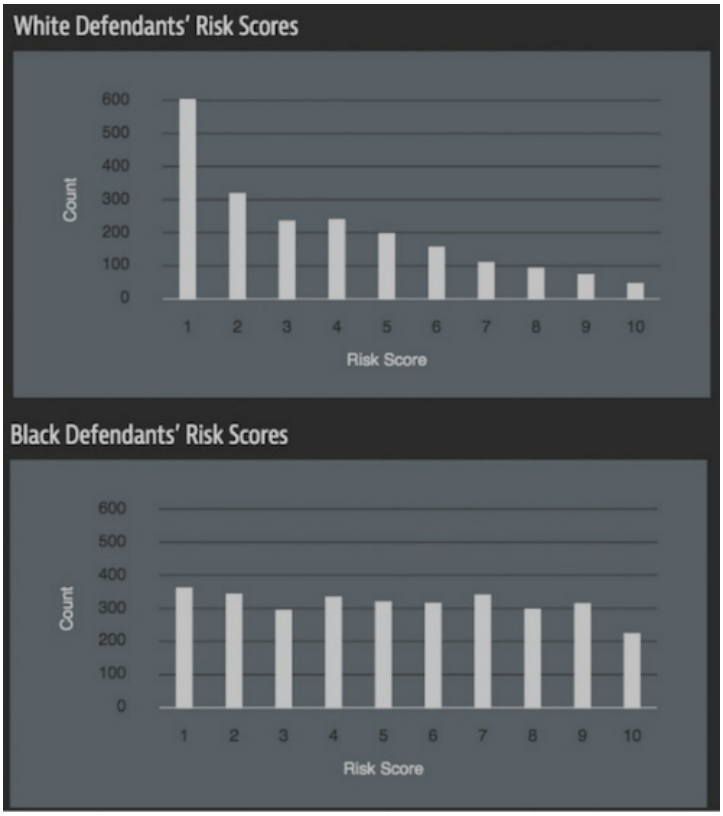

Figura 5. Imagen tomada de Propublica.

Este algoritmo para medir el riesgo de reincidencia, algoritmo propiedad de Northpointe Inc., está sesgado contra ciertos grupos. Los acusados negros tienen mayor probabilidad que los acusados blancos de ser juzgados con mayor riesgo de reincidencia. La imagen que sigue describe muy bien la discriminación en función del grupo al que perteneces y la consiguiente privación de libertad que conlleva.

ISSN 1989-7022
DILEMATA, año 9 (2017), nº 24, 185-217

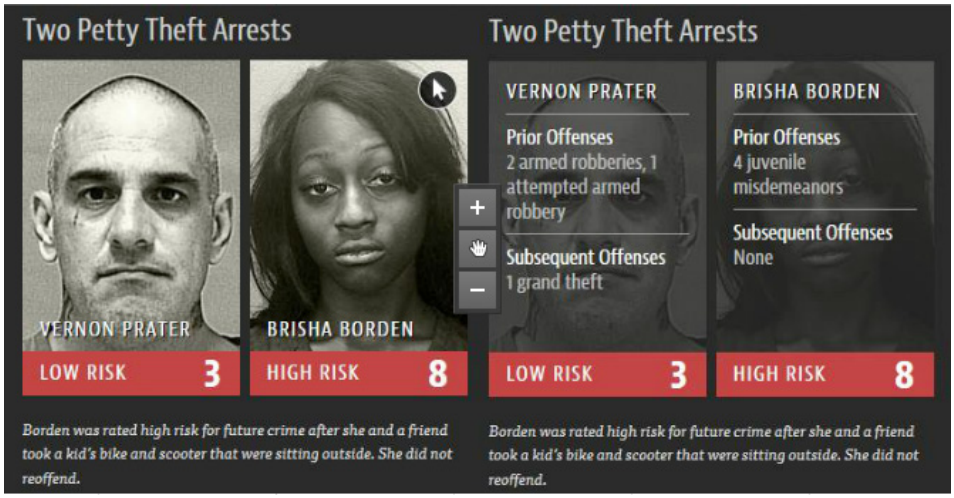

Figura 6. Imagen tomada de Propublica.

La imagen de arriba muestra la ausencia de imparcialidad y tratamiento justo por parte de este algoritmo en función del grupo al que perteneces. El detenido varón de raza blanca, Vernon Prater, por la comisión de un robo menor en tienda tuvo una puntuación de riesgo de reincidencia de 3 (sobre una escala de 0, muy bajo riesgo, 10, riesgo alto.) A pesar de que su historial delictivo era extenso: dos robos con arma de fuego y un intento de robo con arma de fuego. Si lo comparamos con la mujer de raza negra que cometió el mismo delito es objetivamente más grave. Brisha Borden solo cometió cuatro ofensas en forma de comportamiento antisocial (e.g. pintar las paredes de la vía pública) durante su adolescencia. El varón blanco a pesar de llevarse una pena más leve, al cabo del tiempo volvió a cometer un delito: un gran robo. En cambio, la mujer de raza negra, a pesar de llevarse una pena más dura; no volvió a cometer ningún delito en contra de la predicción de reincidencia del algoritmo.

2.4. Discriminación y abuso de Control

Esta es quizá la expresión más extrema de discriminación que pueden ejercer los algoritmos. Es el equivalente a un panóptico digital. El que se conoce como "sistema chino de crédito social" de acuerdo con informaciones publicadas por diversos periódicos, The Indepent o Wall Street Journal. Este sistema pretende dar a todos los ciudadanos chinos una puntuación o clasificación. Una forma de marcar, etiquetar, clasificar etc. a los ciudadanos sobre la base de información personal disponible "on-line". Los chinos que puntúen bajo en este sistema de clasificación de crédito social

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

se les impedirá viajar por el país y fuera de él, acceder a créditos bancarios u otros servicios y ayudas públicas. Este sistema crea una base de datos de toda la información disponible de empresas y ciudadanos con el objetivo de controlar. El gobierno chino con la implementación de este sistema de acuerdo con la literatura de investigación en autoritarismo político podría ser considerado un régimen autoritario. De hecho, ya existen pruebas empíricas de cómo el régimen chino, gobierno chino, con el uso del llamado Big Data fabrica contenido en plataformas sociales de Internet como una estrategia intencional de distracción para impedir el debate y discurso público necesario para la movilización colectiva y por tanto el cambio social en un sistema democrático (King, Pan y Roberts, 2017, en prensa). No es posible pensar en otro ejemplo de un gobierno que haga uso de los algoritmos y de la transformación tecnológica y digitalización del mundo con propósitos coercitivos y autoritarios.

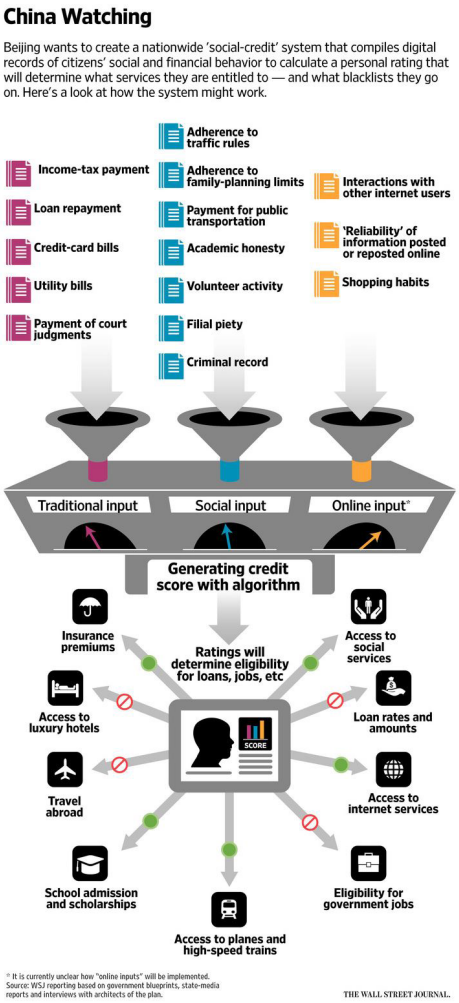

Figura 7. China Watching (Wall Street Journal)

ISSN 1989-7022
DILEMATA, año 9 (2017), nº 24, 185-217

Estos son casos paradigmáticos todos ellos de algoritmos que discriminan, o por decirlo de otra manera, ejemplos de algoritmos *inmorales*. En el siguiente apartado hablaremos un poco de las diferencias en el análisis ético de la IA, datos y algoritmos. Pero antes de acabar este apartado sobre las implicaciones éticas de una sociedad cada vez más gobernada por algoritmos y habiendo mencionado casos paradigmáticos de discriminación o daños causados por algoritmos, es interesante mencionar como el *Internet de Todo* o la comunicación máquina a máquina (M2M), puede de manera exponencial incrementar el daño que puede causar un único algoritmo. Es como el famoso adagio gestáltico: "el *todo es más que la suma de las partes*"; solo que aquí el *todo* puede llegar a ser una gran amenaza para las personas. "Alexa" es el asistente virtual del producto de Amazon, Echo ™. Echo™ es un sistema de monitorización domótica que fue el producto estrella de las pasadas Navidades del 2016. En casa de millones de usuarios Echo™ puede hacer la compra en el supermercado, responder preguntas, poner música, informar del tiempo y encender o controlar la calefacción centralizada. El problema es que este dispositivo único con un software algorítmico que permite comunicarse y controlar otros muchos dispositivos puede tener fallos. Y los fallos pueden producir daños a gran escala. Uno de los más curiosos casos de fallo es el que se produjo en Dallas (link de la noticia aquí: http://dfw.cbslocal.com/2017/01/06/amazon-alexa-orders-dollhouses-for-owners-after-hearing-tv-report/) cuando un niña de seis años pidió al recién comprado Alexa (Echo™) que cantará con ella una canción llamada "dollhouse" y que además le trajera una "casa de muñecas" (dollhouse es la palabra inglesa para "casa de muñecas"). Alexa inmediatamente ordena la compra de una gran mansión de muñecas y no sabemos por qué razones "cuatro libras de galletas de azúcar". Niños pidiendo cosas a través de ordenadores o dispositivos de manera accidental o premeditada no es nada nuevo. Pero esta historia no se quedó ahí. Este caso anecdótico fue recogido en las noticias de una televisión local y cuando el presentador dijo las palabras "Alexa ordena una casa de muñecas" los propietarios de dispositivos Echo que estaban viendo y escuchando las noticias, se dieron cuenta de que sus propios dispositivos pidieron casas de muñecas. Pero hay algo éticamente más preocupante que un fallo operativo o "glitch" de un algoritmo. Cuando un asistente virtual tanto de Amazon Echo™ o Google Home™ graba la voz de sus propietarios: ¿quién tiene acceso a los datos?, ¿y si esos datos pueden *hackearse*?, ¿dónde queda la privacidad con un asistente virtual que nunca duerme y controla todo tu espacio íntimo dentro de tu hogar? Po si fuera poco hay muchos más interrogantes que surgen con otro

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

producto que es la extensión natural de la IA que está detrás del asistente virtual Alexa: Echo Look™. Si Alexa vivía en los altavoces, ahora Echo Look™ vive en una cámara. Y esto es aún más peligroso para nuestra privacidad. Comercialmente Echo Look™ sirve para ayudarnos en nuestro estilo buscando la ropa que mejor nos queda de la línea de ropa y productos que se vende a través de Amazon, pero la IA y el aprendizaje máquina de Echo Look™ puede ahora no solo distinguir objetos (qué zapatos llevas) sino también características (zapatos rojos). Esto le confiere un potencial inmenso de recolectar datos masivos de tu cara (expresiones faciales), cuerpo, habitación, objetos… lo que le convierte en algo más que un simple asistente personal que te recomienda un estilo de vestir o qué chaqueta ponerte para ser una herramienta de evaluación de tus estados de ánimo, tu cuerpo, tu entorno, etc. *Internet de Todo* puede ser el espía perfecto para cualquier voyeur, acosador, hacker, servicio de inteligencia, organización criminal… que difumine por siempre los límites que separan la vigilancia legítima de la ilegítima, y erosione la privacidad hasta el punto de que deje de existir. La transformación tecnológica y digital del mundo pueden tener consecuencias graves para el valor y el derecho a la privacidad. Pero también hay un aspecto curioso del desarrollo de estos asistentes virtuales o bots. Si algún día el desarrollo tecnológico nos conduce a una IA fuerte que permita poblar el mundo con agentes artificiales auto-suficientes similares en prestaciones a Amazon Echo™, Amazon Echo Look™ o Google Home™ pero de una inteligencia mayor, la pregunta es: ¿se les debería garantizar protección legal? En la primavera de 2016, Microsoft hizo público un twitter chatbot llamado MS Tay. Este chatbot se caracterizaba por tener un algoritmo que le permitía aprender cómo interactuar con otros usuarios de Twitter y producir "comentarios" más allá de la supervisión y control de sus programadores. El problema es que en 24 horas la comunidad de Twitter que interactuó con MS Tay consiguió boicotear su proceso de aprendizaje. La interacción de los usuarios de Twitter consiguió que MS Tay se convirtiera en un negacionista del Holocausto, homófobo, transfobo y misógino. Sin embargo, si quisiéramos garantizar y dotar de derechos a la IA y sus derivaciones uno de los primeros derechos a conferir sería el de la libertad de expresión. Tanto MS Tay como Amazon Echo™ o Google Home™ tendrían el amparo de la ley para expresarse como quisieran aunque esto pudiera, en cierto sentido, incomodar a los seres humanos. Y de hecho, así lo defienden teóricos legales (Masaro, Norton y Kaminski 2017).

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

Con múltiples objetos de electrónica de consumo bots y asistentes digitales conectados simultáneamente a " la nube" y entre sí, se puede predecir tu orientación sexual, tu personalidad, tu cuenta corriente o clase social, tus movimientos por la ciudad, tu red social o círculo de amistades, en definitiva, crear un perfil digital preciso de quién eres. Plataformas digitales, redes sociales, apps, smartphones están cambiando la forma en la que nos relacionamos unos con otros y con el mundo. Quizá esta sea la tendencia más peligrosa, en lugar de una IA malévola u algoritmo que gobierne nuestras vidas. Porque la forma en la que percibimos la trasformación tecnológica y digital del mundo atribuyendo inteligencia a un sistema de GPS, un autómata con forma humanoide o a un sistema de IA -que de acuerdo con el estado actual del arte no tienen inteligencia real- va produciendo un cambio paulatino, pero constante, en nuestra forma de entender y asimilar la realidad circundante mucho más dramático que los escenarios de la televisión, el cine y los medios de masas que vaticinan la aparición de un punto de *singularidad* o super-inteligencia artificial que acabará con la humanidad.

3. Ética de datos y ética algorítmica: Una hoja de ruta para navegar la transformación tecnológica y digital del mundo

Esta sección va a tratar de diferenciar las distintas áreas de reflexión ética sobre la transformación tecnológica y digital del mundo. Esta transformación tecnológica y digital incluye el avance de nuestra capacidad para crear maquinas con IA o el diseño y programación de algoritmos que deciden por nosotros, los retos éticos que pueden emerger de esta transformación y la manera en la que pueden afectar a la organización socioeconómica y vida de las personas.

La primera distinción obvia a tener en cuenta es entre ética de datos y ética algorítmica. La ética de la ciencia de datos es una nueva rama de la ética aplicada que investiga los retos éticos de la ciencia de datos. Durante un taller auspiciado por *The Alan Turing Institute* un grupo de autores y académicos se reunieron para establecer la agenda de investigación de este nuevo instituto constituido formalmente como el centro de referencia nacional de la ciencia de datos en el Reino Unido. Como resultado de este taller se publicó un artículo de investigación que llevaba por título: ¿Qué es la ética de datos? (Floridi y Taddeo 2016). En

ISSN 1989-7022 DILEMATA, año 9 (2017), nº 24, 185-217

este artículo se contextualiza el nacimiento de la ética de datos como el legado fructífero de más de 30 años de investigación en la ética de la información y de la computación en respuesta al desarrollo de las tecnologías digitales. Los datos a lo largo de esta historia se han entendido como la materia prima sobre la que abstraemos el mundo en categorías, medidas y otros formatos representacionales (números, caracteres, figuras, imágenes, bits…) que en cierta medida constituyen los fundamentos del conocimiento y la información. Pero, ¿qué son los datos? Etimológicamente la palabra dato proviene del latín *dare* que significa "dar". Por consiguiente, los datos son el producto bruto que está *dado* que se puede obtener de los fenómenos y que se miden y registran de múltiples formas. Sin embargo, en el lenguaje coloquial los "datos" hacen referencia a aquellos elementos que se pueden tomar o coger, que se extraen de observaciones, experimentos, computaciones (Borgman 2007). Técnicamente hablando lo que entendemos como "datos" son realmente "capta" –derivado del latín *capere* que significa "coger". Jensen (1959, ix) lo dice mucho más claro:

> *Es un accidente desafortunado de la historia que el término datum… en lugar de captum… haya llegado a simbolizar la unidad de medida de la ciencia. Dado que la ciencia trata, no con "lo que está dado" por la naturaleza, sino con "lo que se ha tomado" o seleccionado de la naturaleza por el científico de acuerdo con su propósito.*

La ética de datos o la ética de la ciencia de datos se tendría que haber llamado ética de capta o la ética de la ciencia de los capta, respectivamente. Y términos tan populares hoy en día como "Big Data", se tendría que decir correctamente como "Big Capta". En el artículo de Floridi y Taddeo se dividen los retos éticos que se presentan con la ciencia de datos en tres grandes áreas:

1. La ética de datos (cómo los datos se adquieren, tratan y almacenan)
2. La ética algorítmica (cómo la IA, aprendizaje máquina y robots interpretan los datos)
3. La ética de las prácticas (desarrollar códigos de buenas prácticas para profesionales en esta nueva ciencia de datos)

Como vemos, en estas secciones de investigación Floridi y Taddeo separan la **é***tica algorítmica* y **ética de datos** del marco general de la ética de la ciencia de datos. Una de las principales aportaciones o recomendaciones del artículo de Floridi y Taddeo es que se debe implementar un marco de gran escala o enfoque amplio para valorar el impacto e implicaciones de la ciencia de datos en lugar de servirse de aproximaciones estrechas de miras o ad hoc. La ética de la ciencia

ISSN 1989-7022
DILEMATA, año 9 (2017), nº 24, 185-217

de datos puede llegar a examinar como marco general cada una de las secciones en las que se incluyen la **ética algorítmica** y la **ética de datos**, pero estas últimas son áreas autónomas en sí mismas con objeto de estudio y metodología propias. Reflexionar sobre el cambio de paradigma epistémico que implica la gran cantidad de datos que se generan tanto a nivel personal como organizacional o institucional y los problemas derivados (**ética de datos**) que nos conducen a que demos menos importancia a las teorías y nos centremos ahora en los datos y, por consiguiente, tengamos en cuenta las implicaciones éticas en el tratamiento, gestión y almacenamiento de esta gran cantidad de datos; no equivale a reflexionar sobre cómo un agente de IA dirigido por un cierto tipo de algoritmo viene a procesar los datos afectando a las personas o ciertos grupos (**ética algorítmica**) como hemos visto en el apartado segundo de este texto más arriba. Pero tanto la **ética de datos** como la **ética algorítmica** son secciones dentro del marco más general de la **ética de la ciencia de datos**.

Por otra parte, cuando un sistema o agente de IA ya no solo tiene como software un algoritmo que merezca ser monitorizado o auditado por su potencial impacto en la vida de las personas, sino que este sistema o agente de IA tiene capacidades de auto-mejorarse o aprender de manera no-supervisada, empezamos hablar de la **ética de la IA**. Los 23 principios salidos de la conferencia celebrada en Asilomar, más arriba mencionados, o las iniciativas y organizaciones fundadas para crear una IA beneficiosa para la sociedad, algunas también mencionadas, son un ejemplo claro de acciones dentro de la ética de la IA. La ética de la IA trata de servir de constreñimiento mediado por la reflexión ética cuidadosa de la innovación en IA con el objetivo de crear una IA segura bajo el control de los seres humanos y asegurándose de que todo sistema de IA esté alineado de manera sensible con los valores humanos.

A finales del 2016 el gobierno de los EE.UU., el Parlamento Europeo y la Cámara de los Comunes del Reino Unido escribieron, de manera independiente, un libro blanco o informe para valorar las implicaciones de la IA en la sociedad[2]. Como hemos podido en el apartado segundo más arriba, la IA, y en especial la toma de decisiones basadas en algoritmos, tiene implicaciones éticas, legales y económicas para la sociedad y la vida de las personas. No se sabe si los equipos y/o autores encargados de la redacción de dichos informes se coordinaron o han sido totalmente independientes, pero que hayan salido al mismo tiempo pone de manifiesto la

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

preocupación "ética" por el hecho de que la rápida transformación tecnológica y digital del mundo conduzca a una buena sociedad.

4. ¿Cómo crear software y algoritmos morales?

Hasta ahora hemos visto una definición operativa de algoritmo (código software que procesa un conjunto limitado de instrucciones), las implicaciones éticas de una sociedad cada vez más gobernada por algoritmos con casos paradigmáticos donde las decisiones basadas en algoritmos discriminan socialmente, económicamente, privan de libertad y cometen abusos de control. Por otra parte, también hemos diferenciado entre ética de datos y ética algorítmica, áreas independientes de ética aplicada que se englobarían dentro del marco general de la ética de la ciencia de datos. En esta última sección hablaremos un poco de otra nueva rama de la ética aplicada: la ética de máquinas. La ética de máquinas o ética robótica tiene entre sus principales objetivos llevar la "ética al silicio", o por decirlo de una manera menos poética, programar y diseñar software y algoritmos morales para construir máquinas éticas autónomas. La ética de máquinas o ética robótica tiene dos grandes acepciones. La primera hace referencia al intento de crear "consciencia moral" en los robots o sistemas de IA. Agentes morales artificiales con la misma capacidad que los seres humanos de razonar y tomar decisiones de carácter moral. Aunque en la actualidad existen diversos grupos de investigación que trabajan para crear un "sentido" moral en agentes artificiales ya sean (ro)bots, sistemas de IA etc. (véase, Malle y Scheutz 2014), uno de los mayores obstáculos para esta forma de entender la ética de máquinas es que todavía no sabemos muy bien cómo se da el proceso de evaluar y tomar decisiones morales en los propios seres humanos. La segunda acepción hace referencia al diseño pro-ético de la tecnología robótica. En muchos aspectos, la ética de máquinas compartiría objetivos comunes con la ética de datos y la ética algorítmica e incluso el marco general de la ética de la ciencia de datos. Pero lo que distingue a la ética de maquinas (o robótica) de la ética de datos y la ética algorítmica es el deseo de que la interacción de las personas con la tecnología esté basada en un marco ético genuino para que principalmente los seres humanos no sufran ningún daño. Para conseguir esto se pretende llevar la "ética al silicio". La ética de datos, ética algorítmica o incluso ética de la ciencia de datos serían salvaguardas y controles éticos externos de la tecnología. La ética de máquinas o ética robótica

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

quiere diseñar y construir máquinas y robots que tengan la capacidad moral en sí misma, un sentido moral implementado en su sistema operativo. Esto nos plantea un reto importante: el estatuto moral e incluso legal de los robots. Si en un futuro no muy lejano logramos crear consciencia o auto-consciencia en las máquinas, un paso previo para un sentido moral pleno, esto directamente cambiaría nuestra relación con los artefactos, los robots o el software inteligente. Si llega el día en el que un robot pueda tener una conducta autárquica con respecto a su programador y realice un acto ilegal: **¿Habrá que castigarle?, ¿otorgarle derechos de protección legal justa**? Recordemos lo dicho más arriba sobre Amazon Echo™, Google Home™ o MS Tay y la protección de su derecho a la libertad de expresión defendida por autores como Masaro, Norton y Kaminski (2017). En esta dirección, la de conseguir crear software y algoritmos morales, se han producido ciertos avances en lo que se viene en llamar "programación moral" o ética de máquinas en general. Cuando uno habla de ética de maquinas lo primero que se le viene a la cabeza son las tres leyes de la robótica de Asimov[3]. Continuando con el trabajo de Asimov una serie de estándares basados en reglas para robots se han propuesto en Corea del Sur por el Ministerio de Comercio, Industria y Energía. En Europa, EURON (European Robotics Research Network) anunció planes para desarrollar una guía para robots en cinco áreas: seguridad, protección, privacidad, trazabilidad, e identificabilidad. Pero a pesar de ser lógicas (aparentemente) y bien intencionadas, las tres leyes de Asimov que han inspirado protocolos y diversos procedimientos no son el mejor de los pasos o comienzos para pensar en sistemas autónomos éticos. El razonamiento basado en programación lógica permite modelar la permisibilidad moral (doctrina del doble efecto) y procesos duales de la mente en el razonamiento moral (razón vs emoción). El trabajo de Luis Moniz Pereira y Ari Saptawijaya (2016) se centra en las habilidades fundamentales que se requieren para el razonamiento moral en las máquinas. Como recomendaciones básicas, pero ineludibles, si se quiere implementar moralidad en las máquinas es definir la estrategia a seguir. Es probable que haya varias formas de aproximarse a la programación moral de software y algoritmos, pero son dos las estrategias generales más eficaces: el enfoque arriba-abajo y el enfoque abajo-arriba.

El enfoque arriba-abajo se caracteriza por la supervisión completa del programador sobre el tipo de ética o moral que quiere que un agente moral artificial exhiba. Es decir, el programador debe seleccionar un tipo de ética o moral que se caracterice por

ISSN 1989-7022

DILEMATA, *año 9 (2017), nº 24, 185-217*

unos principios generales, poco flexibles, constreñidos que se instalan en el agente moral artificial y este guía su comportamiento sobre la base de estos principios. Dentro de este enfoque arriba-abajo encajan bien doctrinas o teorías éticas de la familia consecuencialista como el utilitarismo, y el deontologismo del tipo kantiano. Este tipo de enfoque se basa en un formalismo lógico de estos principios. Por su parte, el enfoque abajo-arriba se caracteriza por no dejar determinado de antemano qué se ha de codificar o qué reglas son las adecuadas para que el agente moral artificial actúe, sino que deja que este aprenda del entorno y de la interacción con otros agentes morales artificiales. Las doctrinas éticas o morales que son adecuadas para este tipo de concepción de la programación moral son aquellas basadas en un razonamiento de o por casos, y de la interacción con otros agentes, a saber, la ética de las virtudes aristotélica. Si queremos construir un agente moral desde el enfoque arriba-abajo, recordemos que las teorías éticas propias de este enfoque son la familia consecuencialista y el deontologismo, debemos asegurarnos que este agente moral artificial disponga de sensores que permitan percibir la situación, generar posibles cursos de acción dentro del contexto en el que se encuentra y computar las consecuencias de cada una de las alternativas en términos de "utilidad". Por su parte, si queremos construir un agente moral desde el enfoque abajo-arriba, basado en la ética de las virtudes, debemos crear una red neuronal conexionista que tenga entre sus principales parámetros la posibilidad de ajustar las conexiones entre los nodos de acuerdo a ciertos valores en función del aprendizaje y la interacción multiagente. Es posible crear enfoques híbridos donde los agentes morales artificiales podrán actuar sobre la base de reglas morales absolutas como "no matar", pero también comportamientos morales flexibles de acuerdo o en función de la situación como, por ejemplo, "matar si matando salvas más vidas que no haciéndolo" o "actuar de acuerdo al desarrollo de una serie de habilidades que conducen al mejor comportamiento". Todavía queda mucho por hacer en la ética de máquinas, pero esta es una de las fronteras a explorar si queremos que la transformación tecnológica y digital del mundo sea ética y más amable con las personas. Un camino en esta dirección es propiciar la creación por parte de la sociedad civil, pero también por parte de las administraciones públicas, de consultoras y auditorias para identificar las implicaciones sociales, legales y éticas de los entornos de tratamiento intensivo de datos con algoritmos y otras técnicas de Big Data que puedan discriminar y suponer un alto riesgo de amenaza a los derechos y libertades civiles de las personas.

ISSN 1989-7022
DILEMATA, año 9 (2017), nº 24, 185-217

5. Discusiones

En este escrito hemos presentado una breve síntesis de la complicada tarea que tiene por delante la ética algorítmica. Cuando los algoritmos y sistemas automatizados toman decisiones en nombre de las personas necesitamos un diseño pro-ético de la tecnología o en su defecto monitorizar y cuantificar el impacto ético de la tecnología para evitar la discriminación que los algoritmos y sistemas automatizados pueden ejercer. La transformación tecnológica y digital del mundo está generando un enorme progreso en sistemas de IA cuyos algoritmos son utilizados por los gobiernos, las fuerzas de seguridad de los estados, compañías privadas, sistemas públicos de sanidad, bancos etc. Estos sistemas y los algoritmos pueden tener fallos que pueden afectar gravemente a las personas dado el gran número de aplicaciones que tienen. Como hemos visto a lo largo de este escrito los algoritmos pueden discriminar sobre la base de tu etnia o raza, por razones económicas y/o sociales. El gran problema ético que surge del hecho de que las compañías tecnológicas que desarrollan estos sistemas y algoritmos normalmente no revelan cómo funcionan -para así evitar que otras compañías hagan espionaje industrial y los copien- es la imposibilidad de rendir cuentas y/o responsabilidad. Los algoritmos y sistemas de IA son poco trasparentes o totalmente opacos. De esto se sigue que los algoritmos se convierten en "cajas oscuras" muy complejas, difícilmente corregibles si presentan fallos y dada su ubicuidad e invisibilidad uno cree que cualesquiera efectos que producen (por muy negativos que resulten ser) hemos de aceptarlos porque así es como son las cosas y nada puede hacerse para evitarlo (conformidad y resignación). Esta falta de transparencia (opacidad), la complejidad/ubicuidad/invisibilidad y la conformidad/resignación ante los efectos que los algoritmos producen hace imposible aplicar reglas éticas particulares. La aplicación de la ética para garantizar todos los derechos de las personas y en particular aquellos que se pueden conculcar a partir de los usos innovadores de los datos personales no es algo abstracto y difícil de conseguir. Lo vemos todos los días en el comportamiento de las personas y en el comportamiento de las organizaciones y diversas figuras jurídicas ficticias. Las personas y las organizaciones se rigen por normas sociales y legales motivadas por principios éticos de justicia, equidad etc. ¿Debería existir un cuerpo regulador independiente y externo que regule las prácticas del Big Data? Realmente sería un gran paso el que todo el mundo siguiera unos estándares éticos. El problema es que nadie sabe cuáles han de ser estos estándares éticos. No hay precedente histórico,

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

no ha habido ningún caso en la filosofía ética normativa en el que haya habido una posición unánime. De hecho, tampoco es este el objetivo de la filosofía ética normativa. Es casi imposible hacer que todo el mundo esté de acuerdo. Pero para eso tenemos las políticas públicas. El problema es la flexibilidad de estas políticas públicas. Si por ejemplo, se quiere hacer investigación en Big Data o en diseño de algoritmos, se recopilan los datos y la investigación se hace en una universidad, habrá que pedir permiso y la aprobación del comité de ética de la universidad. Pero si haces esta misma investigación para una compañía privada, no tienes que pedir permiso a nadie. Se pueden hacer cosas en el sector privado que ni se pueden soñar poder hacer en la academia. No se puede saber qué es lo correcto o incorrecto -éticamente hablando- pero que un grupo de gente se dedique a obtener el mayor conocimiento posible para beneficiar a la sociedad y que otro grupo de gente se dedique a obtener el mayor beneficio económico de la gente y el primer grupo este sujeto a normas y regulaciones estrictas y el segundo no, realmente es paradójico. Sobre todo si pensamos que hasta hace bien poco el conocimiento se producía en las universidades, pero ahora es al revés. Las compañías privadas tienen y generan más conocimiento que las universidades. Google sabe más de ti y de otras muchas más cosas de lo que puede saber el profesor de psicología de la universidad o los departamentos de historia y economía juntos. Y esto puede causar un gran daño a la humanidad, sobre todo porque la regulación del sector privado es mucho más laxa que la regulación del sector público. Es por ello que el equilibrio es muy importante. ¿Qué tipo de equilibrio? No lo sabemos, pero la ética algorítmica es la práctica para poder llegar a este equilibrio. Es el principal objetivo de la ética algorítmica auditar bajo principios éticos consistentes la transformación tecnológica y digital del mundo. Uno de estos principios éticos esenciales podría ser: prohibir a toda máquina inteligente o semi-inteligente tomar decisiones que afecten a las personas sin la supervisión última de un ser humano. Solo los seres humanos, como agentes morales per se, pueden entender las sutilizas y complejidades de las situaciones que afectan a otros seres humanos. La tecnología es maravillosa, pero debe ser moral.

ISSN 1989-7022

DILEMATA, año 9 (2017), nº 24, 185-217

Bibliografía

Arrow K. (1972): "The Theory of Discrimination", En *Discrimination in Labor Markets*, O. Ashenfelter and A. Rees, (eds.) (Princeton, NJ: Princeton University Press. pp.3-34.

Borgman C. (2007): *Scholarship in the Digital Age*. Cambridge, MA., MIT Press.

Bostrom N. (2014): *Superintelligence: Paths, Dangers, Strategies*. Oxford, Oxford University Press.

Echeverria J (1999): *Los Señores del Aire: Telepolis y el Tercer Entorno*. Barcelona, Destino.

Evans D. (2011): "The Internet of Things How the Next Evolution of the Internet Is Changing Everything".WhitePaper.Cisco.(http://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf)

Floridi L. y Taddeo M. (2016): "What is data ethics?" *Phil. Trans. R. Soc. A*, vol. 374, no. 2083 20160360

Ford M. (2015): *Rise of Robots: Technology and the Threat of Jobless Future*. New York. Basic Books.

Frey C. y Osborne M. (2013), "The Future of Employment: How susceptible are jobs to computerisation?" University of Oxford. [http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf] Recuperado 30 de marzo de 2017.

Goodman M. (2016): *Future Crimes: Inside the Digital Underground and the Battle for Our Connected World*. New York, Anchor Books.

Harari N. (2016), *Homo Deus: A Brief History of Tomorrow*. London. Harvill Secker.

Hil R. (2016): "What an algorithm is?" *Philosophy and Technology* 29 Nº 1 pp. 35-59.

Jensen H. (1950): "Editorial note" En H. Becker (1952), Through values to Social Interpretation Durham, Duke University Press. pp. vii-ix.

King G., Pan J. y Roberts M. (2017, en prensa): "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument." *American Political Science Review*, XX-XX.

Lewis M. (2014): *Flash boys: A Wall Street revolt*. New York, WW Norton & Company.

Malle B y Scheutz M (2014): "Moral competence in social robots" *IEEE International Symposium on Ethics in Science, Technology and Engineering*, pp. 1-6.

Masaro T., Norton H. y Kaminski M. (2017): "Siri-ously 2.0: What Artificial Intelligence Reveals about the First Amendment." *Minnesota Law Review*, Arizona Legal Studies Discussion Paper No. 17-01.

Morozov E. (2014): *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York. New York, Public Affairs.

O´neil C. (2016): *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York, Crown Publishing Group.

Powles J. y Veliz C. (2016): "How Europe is fighting to change tech companies' 'wrecking ball' ethics", *The Guardian*, 30 de Enero, https://www.theguardian.com/technology/2016/jan/30/europe-google-facebook-technology-ethics-eu-martin-schulz (último acceso 02/02/2017)

Steiner. C. (2012): *Automate This: How Algorithms Came To Rule The World,* New York, Portfolio/Penguin.

Taplin J. (2017), *Move Fast and Break Things: How Facebook, Google, and Amazon Cornered Culture and Undermined Democracy*. New York. Little, Brown Company.

Turkle S. (2012): *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York, Basic Books.

Wallach W. (2015): *A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control*. New York, Basic Books.

Yudkowsky E. (2008): "Artificial Intelligence as a Positive and Negative Factor in Global Risk". En: Nick Bostrom y Milan M. Ćirković (eds) *Global Catastrophic Risks*, New York: Oxford University Press, pp.308–345.

Notas

1. Reglamento (UE) 2016/679 del Consejo y Parlamento Europeo del 27 de Abril del 2016 sobre la protección de las personas físicas en relación al procesamiento de datos personales y sobre el libre movimiento de dichos datos y por lo que se deroga la Directiva 95/46/EC. El 24 de mayo de 2017 entra en vigor este reglamento y pasado un año y medio, el 25 de mayo de 2018, su aplicación efectiva.

2. Executive Office of the President National Science and Technology Council Committee on Technology. (2016). Preparing for the Future of Artificial Intelligence. Washington D.C. USA. Recuperado de https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing _for_the_future_of_ai.pdf;
European Parliament. JURI Workshop on Robotics and Artificial Intelligence 17-10-2016. (2016, October 12). Recuperado de http://www.europarl.europa.eu/committees/de/eventsworkshops.html?id=20161017CHE00181;
House of Commons Science and Technology Committee. (2016a). Robotics and artificial intelligence (No. Fifth Report of Session 2016-17). London, UK. Recuperado de http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf

3. Las tres leyes de la robótica del escritor de ciencia-ficción Isaac Asimov fueron publicadas por primera vez en 1942 en un ensayo corto que llevaba por título "Runaround". Estas tres leyes de la robótica dicen: 1 Un robot no hará daño a un ser humano o, por inacción, permitir que un ser humano sufra daño. 2 Un robot debe obedecer las órdenes dadas por los seres humanos, excepto si estas órdenes entrasen en conflicto con la 1ª Ley. 3 Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la 1ª o la 2ª Ley. Posteriormente, Asimov extendió las tres leyes para incluir un número adicional de leyes: Ley Zero Un robot no hará daño a la humanidad o, por inacción, permitir que sufra daño . Ley menos-uno un robot no hará daño a seres sintientes o, por inacción, permitir que un ser sintiente sufra daño. Cuarta Ley un robot debe establecer su identidad como un robot en todos los casos. Cuarta Ley alternativa un robot se debe reproducir al menos que interfiera con la primera, segunda, o tercera ley. Quinta Ley un robot debe saber que es un robot.

ISSN 1989-7022