Controlando las inteligencias artificiales: ¿auditoría o confesión?

IGNACIO GONZÁLEZ
Inspector de Finanzas del Estado

RESUMEN

La sociedad actual ha sido impregnada por la cultura el *dataísmo* que conduce de forma inexorable a la proliferación de la Inteligencia Artificial. La convicción de que su desarrollo encierra peligros y amenaza múltiples derechos impulsa peticiones de esfuerzos regulatorios y de control. Conjurar estas amenazas requiere una adecuada comprensión de cual es el concepto de I.A, sus tipos y sus limitaciones. El primer propósito de este documento es precisar estos conceptos. El segundo es despejar imprecisiones e ingenuidades que se vierten en la literatura no especializada sobre sus deficiencias y sus sesgos. Se identifican los requerimientos que les deberán ser exigidas como la privacidad diferencial y la equidad de grupo. Se acota por último el trabajo que deberá y podrá ser realizado por reguladores y auditores, que es muy distinto en su naturaleza al que hemos visto hasta ahora proponerse.

PALABRAS CLAVE

Inteligencia artificial | Sesgo | Equidad | Algoritmo | Auditoría

ABSTRACT

Today's society has been permeated by the dataism, a cultural trend that inexorably leads to the proliferation of Artificial Intelligence. The conviction that its development contains dangers and threatens multiple rights, drives requests for regulatory and control efforts. Conjuring these threats requires a proper understanding of what A.I is, its types, and its limitations. The first purpose of this document is to clarify these concepts. The second is to clear up inaccuracies and naivetés that are poured into the non-specialized literature about its deficiencies and biases. Last, but not least we propose the requirements that should be demanded of them, such as differential privacy and group equity. Finally we suggest a new type of controls and regulations, an original approach for regulators and auditors.

KEYWORDS

Artificial intelligence | *Bias* | *Fairness* | *Algorithm* | *Auditors*

1. Introduccion

Vivimos en un mundo insólitamente fascinado por los datos, en el que se extiende progresivamente la conciencia de nuevas amenazas. Los reguladores se enfrentan a la obligación de comprender la nueva realidad. Los principios y las ideas que fueron utilizados durante las últimas décadas para gestionar y auditar los sistemas informatizados, para la gestión de la «informatización y la mecanización» se muestran obsoletos.

El propósito de este artículo es: a) describir el contexto de la nueva realidad; b) acotar el concepto y posibilidades de la Inteligencia Artificial (IA); c) mostrar los peligros que presenta y sus tipos y por último identificar cuáles son los objetivos de las auditorias que podrían y deberían realizarse argumentando contra errores extensamente difundidos, señalando qué aspectos requieren ineludiblemente una nueva regulación de un carácter muy distinto al que vemos constantemente proponer.

1.1. Big Data y dataísmo

La Primera Ilustración, empeñada en que el ser humano alcanzara su mayoría de edad, utilizó la Estadística, la Ciencia del Estado, de modo instrumental para describir la realidad y para controlar la sociedad. Al expandirse las funciones del sector público, se aplicaron progresivamente técnicas cada vez más complejas, de inferencia, para dimensionar la oferta de servicios, pero los datos en sí estaban supeditados al sentido.

La Segunda Ilustración cree que los datos nos proporcionan, además de información, transparencia. Defiende que su uso masivo, el Big Data, permite eliminar lo subjetivo, afirmando que cuando hay datos la teoría sobra o en otros términos, que cuantos más datos menos *fake news,* con una confianza viciosa, pues cada saber tiene su objetivo. Las Matemáticas deben ser exactas, la Historia ser verdadera, la Ingeniería debe ser eficiente, la Magia debe ser eficaz. Las noticias no deben ser exactas como las divisiones sino verdaderas.

Ha emergido la filosofía del *dataísmo:* «[Se cree que…] todo lo que puede ser medido debe ser medido, que los datos *son lentes transparentes y fiables,* que nos permiten filtrar todo emocionalismo y toda ideología y que los datos nos permitirán hacer cosas significativas como predecir el futuro» (David Brooks. *The New York Times,* 2013). Al criticar este enfoque Byung-Chul Han en *La sociedad de la transparencia* afirma con razón que el dataísmo es *nihilismo* porque renuncia al sentido, porque los datos son aditivos, pero no narrativos.

Las Administraciones Públicas utilizan en la actualidad, cautivadas por esta idea, grandes cantidades de datos. Las Administraciones Tributarias han llevado al extremo este interés, requiriendo a los contribuyentes datos sobre sí y sobre otros, para, a través del cruce de lo declarado con lo imputado, minimizar el fraude. Es lógico que todas ellas piensen ahora en cómo utilizar las técnicas de analítica avanzada, de *machine learning* y en sentido amplio de *Inteligencia Artificial* para el cumplimiento de sus misiones, participando de la idea de que los datos son lentes transparentes para conocer la realidad y de que, si permiten conocer el futuro, también permitirán conocer el pasado y con ello a los defraudadores.

2. La inteligencia artificial

2.1. Del algoritmo a la Inteligencia Artificial

Un algoritmo es un procedimiento no ambiguo para resolver una clase de problemas. Se compone de un conjunto de instrucciones. En el ámbito cotidiano una receta de cocina es un algoritmo, como también lo es el que se enseña en las escuelas para realizar una división. Los programas informáticos utilizan lenguajes de programación para construir algoritmos, y con ellos realizan tareas como ordenar a los censados por orden alfabético. Un algoritmo puede ser correcto o incorrecto, eficiente o no, pero sería equivocado calificarlo estimativamente, usando valores. Si en una escuela la dirección utilizase un algoritmo para repartir las tartas proporcionalmente al peso de los niños, afirmar que el algoritmo es «malo» es una *hipálage,* figura retórica por la que se atribuye un adjetivo (malo), al nombre que no conviene (algoritmo) en vez de a otro conectado con él (dirección del colegio), además de un error de concepto, aunque, de hecho, se comete.

Estas piezas de software son la parte fundamental de los ADS *(Algorithmic Decision Systems)*. Su uso, si son erróneos, puede ocasionar daños visibles y reclamables como la denegación injusta de un préstamo hipotecario, pero hay casos en que los daños son insidiosos. Las virtudes principales de un algoritmo son: que sea *comprensible* y que permita la imputabilidad *(accountability)*. Lo primero supone que tanto la documentación de diseño, como el código y los datos de entrenamiento sean accesibles al legítimamente autorizado para auditarlos y además que se puede explicar cómo funciona mediante frases con sentido. La imputabilidad es la propiedad que describe la capacidad para que las responsabilidades por las vulneraciones éticas o legales puedan ser exigidas.

Una inteligencia artificial es un tipo de algoritmo con propiedades emergentes propias que estudiamos a continuación.

2.2. La Inteligencia Artificial es en el fondo un algoritmo muy grande

En el año 1956 John McCarthy, un joven profesor pensó en organizar un curso de verano donde matemáticos, informáticos y psicólogos pensaran sobre si se podrían llegar a construir máquinas inteligentes. Aunque lo que verdaderamente le interesaba era estudiar la inteligencia, para buscar fondos, que consiguió sólo en parte de la Fundación Rockefeller, probó a titular el curso como «Inteligencia Artificial» buscando marcar distancias con la disciplina de moda, la *cibernética*. Suponía una enorme osadía pensar que se podría pasar de que los programas ejecutase algoritmos, conjuntos de reglas, a que fueran inteligentes y muchos pensaron que harían falta «cientos de premios Nobel», antes de que fuera posible, si realmente lo era.

Los asistentes al retornar a sus universidades en ese otoño impulsaron proyectos con enfoques distintos pues eran profesores de distintas asignaturas. Hasta aquel momento se realizaban programas de ordenador basados en instrucciones (algoritmos) que especificaban lo que había que hacer en cada momento, con cada dato en aplicaciones como las censales o las de elaboraciones de nóminas. Se pensó en cómo crear algoritmos con objetivos abstractos en vez de específicos. Se trabajó para que pudieran «aprender» en base a los resultados de anteriores ejecuciones. Se aplicó al juego de las tres en raya pretendiendo que, sin necesidad de especificarle lo que tenía que hacer en todas y cada una de las situaciones posibles, apuntase él solo las líneas utilizadas en las partidas perdidas, para luego no volver

a hacerlo, con el enfoque que se narró de forma novelada en la película Juegos de guerra. Luego se pensó en otros para que pudieran aprender a ganar no solo en un juego sino en cualquier juego. Otros investigadores abordaron problemas más complicados, aunque más concretos, como el juego del ajedrez e intentaron resolverlo usando reglas. Arthur Samuel de IBM enseñó así a un sistema a jugar al ajedrez. Se dio cuenta de que una máquina no puede jugar mejor que su maestro utilizando esta estrategia, pues la máquina solo le ganaba cuando se distraía. En los años 90 Deep Blue, el sistema que ganó a Kasparov, incorporaba ocho mil reglas y lo difícil era calibrar qué combinación de ellas había que aplicar en cada caso. Pero por esta línea de trabajo, la IA conseguida, no era general, podía hacer bastante bien una y solo una cosa, y además no podía ser más inteligente que sus creadores.

Los ingenieros apuntaron más alto con objetivos como el tratamiento de la voz y el reconocimiento de la imagen. Se encontraron de bruces con el hecho de que las IA pueden razonar (lógicamente) y calcular bien utilizando poca computación, un algoritmo pequeñito, mientras que para reproducir otras habilidades como las sensoriales y motoras se requieren enormes algoritmos. Se trata de la paradoja enunciada por Hans Moravec (1968). El motivo parece claro. En los seres vivos el algoritmo biológico por el que las amebas han pasado a volar se ha construido muy poco a poco durante millones de años. Un algoritmo para que un robot traiga el café andando de la barra a la mesa sin dejar caer una gota, no lleva a los programadores tanto tiempo hacerlo como le ha llevado a la selección natural, pero no se pudo hacer en un año. En otros términos, en un futuro próximo, abogados, miembros de las juntas de libertad condicional o analistas de valores verán sus trabajos amenazados porque razonan y calculan a diferencia de los cocineros y jardineros que deben tener, además sensibilidad.

Durante un periodo que se llamó el invierno de inteligencia artificial, los ingenieros abrumados por estas dificultades se cuestionaron los límites de su ambición. Algunos no cayeron en la depresión y mantuvieron la fe en que las computadoras podrían alcanzar una inteligencia general sobrehumana, mientras otros lo siguieron considerando imposible.

Algunos investigadores buscaron cómo simular el comportamiento de las neuronas del cerebro y sus conexiones, desarrollando lo que fueron llamadas *redes neuronales*. Google desarrolló sucesivas versiones de una inteligencia artificial con el objetivo de ganar en el increíblemente complejo problema de ganar a un campeón mundial en el juego del Go. Las primeras versiones fueron llamadas Alpha Go Fan y AlphaGo Lee tomando el nombre de aquellos campeones a los que habían derrotado. Utilizaban unos algoritmos conocidos como Q_learning y búsqueda de Montecarlo, esto es, que *además* de las reglas del Go, partían de tener «algo» dentro, los algoritmos.

En su última versión, Alpha Go Zero, se creó un sistema que sólo tenía las reglas del Go y que aprendía por el brutal sistema de jugar contra sí mismo y sacar conclusiones de las pérdidas y las derrotas. Repetimos, partiendo de una *tabula rasa* se creó un sistema que ganó al campeón del mundo de Go y además pudo ganar a la IA campeona de ajedrez anterior, que utilizaba reglas, tras jugar contra sí misma solo durante cuatro horas. En la partida que derrotó al campeón del mundo, se produjo una jugada asombrosa. Los comentaristas, que no eran campeones, pensaron que era un error, pero el derrotado dijo: «no es un movimiento humano..., *so beautiful*». Los maestros de Go la incluyeron entre el tipo de movimientos denominados *kami no itte* movimientos que vienen «de la mano de Dios».

2.3. Las tribus de los desarrolladores de IA y el auditor

Durante los ya setenta años en los que ha evolucionado la IA sus creadores han seguido un número limitado de sendas. De una forma muy clara el español Pedro Domingos (2015) explicó la estrategia de las cinco tribus de exploradores pioneros. Cada uno en su universidad creó una escuela.

Simbolistas. Como los acadios y los fenicios trabajaron con símbolos. Planteada una pregunta, la convertían en símbolos y luego operaban con expresiones, como hacen los algebristas. Su forma de avanzar era la *deducción inversa*. Se parte de un pequeño conocimiento y con pequeños pasos se avanza hacia el origen de la cuestión acumulando conocimiento progresivo. Crearon entre otras la técnica de «arboles de decisión».

Conexionistas. Intentaron reproducir como funcional el cerebro humano aplicando el equivalente a una neurona humana (como la desvelada por Cajal) presentada en forma de una ecuación. Sustituían las conexiones del cerebro por las conexiones de las ecuaciones y su enfoque fue el de ingeniería inversa. Impulsaron las CNN *(Convutional Neural Networks)*.

Los evolucionistas de orientación biologista y cibernética. Siguiendo la idea de Darwin pensaron en sistemas que aprendiesen por selección. Un ejemplar actual es la llamada programación genética. El sistema se auto inventa valores para las variables (mutaciones) y si acierta más que antes, acepta el valor como provisionalmente bueno repitiendo muchas veces el proceso.

Los Bayesianos con raíces matemáticas. Ven el mundo lleno de incertidumbre que se puede reducir con sucesivas observaciones y experimentos y crearon el concepto de «redes bayesianas».

Los analogistas con vocación de agrimensores. Como los escolásticos medievales pensaron por analogía desarrollando algoritmos para medir con precisión si dos cosas son semejantes o no. Una de sus estrategias es KNN.

Cada tribu persigue un objetivo: los simbolistas la precisión al clasificar, los conexionistas minimizar una función que se llama descenso del gradiente, los evolucionistas maximizar una función que se llama *fitness function,* los bayesianos mejorar la estimación de la probabilidad a posteriori y los analogistas medir bien una distancia.

El auditor, al llegar a una organización se encontrará representantes de las distintas tribus. Se encontrará con «arboles de decisión» y cada vez más con redes neuronales profundas, simulaciones del cerebro. El cerebro tiene cien mil millones de neuronas. GPT3 el antecesor de ChatGPT es un modelo con 175.000 millones de parámetros. Comprender «lo que hace» es imposible como es imposible saber «lo que ha hecho» el cerebro humano cuando ha tomado una decisión. Para el lector no especializado una IA como ChatGPT hace algo como resolver un SUDOKU de un millón de casillas y las demás resuelven un sistema de ecuaciones muy difícil. El auditor debe renunciar a ello, debe renunciar a que le expliquen cómo se resolvió el sudoku porque sería muy cansado y además distinto a cómo se resolverá la siguiente vez. Los otros tres sistemas, programación genética, bayesianos y analogistas sí que pueden comprenderse por un técnico como se comprenden el resultado de un sistema de ecuaciones, pero el auditor sacará el mismo provecho que si le explican la solución de una ecuación diferencial. No es el camino.

2.4. ¿Son inteligentes las inteligencias artificiales?

La primera cuestión para dar sentido al debate es: ¿Qué se entiende como inteligencia? ¿Se trata de algo dicotómico como el ser inmortal a lo que debe responderse sí o no o es cuestión de grado y convención?

El criterio más utilizado es el de Turing (1950), el malogrado científico inglés que hizo posible el desciframiento de la máquina Enigma durante la IIGM. Defendió que algo es «inteligencia artificial» cuando un ser humano no es capaz de distinguirlo por sus respuestas de un ser humano. Desde el punto de vista pragmático, puede servir este reconocimiento por aclamación, pero el filósofo y el legislador reclaman algo más. Filosóficamente, no es de recibo tomar como criterio para aceptar que algo sea o no una concreta *res,* «cosa», que alguien, uno o varios, la reconozcan como tal, pues bien podrían los observadores necios. No basta para que algo sea verdadero que entre varios no sean capaces de encontrar el fallo. Con más profundidad el Baghavad Gita diferencia que el conocimiento de la inteligencia y entiende ésta como la capacidad de ver algo desde la perspectiva correcta.

Los técnicos han sido aguafiestas. Han cuestionado que las IA sean realmente inteligentes, pero desde otra perspectiva. Las IA parten del dato y solo del dato y construyen sus resultados sobre las relaciones y las correlaciones entre ellos. Parte de la facultad de la inteligencia consiste en poder identificar las causas. De hecho, en la visión escolástica, la inteligencia es la capacidad *intuitiva* de alcanzar la verdad. Es bien conocido que la correlación no implica causación, y que, aunque dos variables evolucionen del mismo modo, por ejemplo, el consumo de helados y el número de ahogados, la causa de los ahogamientos puede ser otra, el calor por el que la gente se va a bañar. Autores como Judea Pearl han creado una teoría, la de las *Redes Bayesianas* para identificar las causas a partir de los datos, lo que *no puede hacer* por ejemplo una red neuronal de la que luego hablaremos. Esto autores piensan por tanto que una red neuronal por bien que clasifique o busque o responda (ChatGPT) no puede ser nunca verdaderamente inteligente. Por construcción.

Tampoco este enfoque es definitivo. Los filósofos que siguen a Hume afirman que el término causa es un *flatus vocis,* que el pensador utiliza para explicar regularidades, pero que, poniéndose pesados, no hay forma de probar, por el hecho de que el sol haya salido todos los días en el pasado, que vaya a salir mañana. Con ello el no poder identificar las causas no sería una cosa tan grave.

Para nosotros, aunque sea una cuestión escurridiza, después de saber que no se puede reducir la inteligencia a calcular ni a encontrar causas, discernir las que lo son de los algoritmos que no lo sean, creemos que es importante, si se va a permitir que sus decisiones afecten a seres humanos, es relevante, sobre todo si se van a regular.

La palabra inteligencia tiene su origen etimológico en el latín *inter legere,* la capacidad del que sabe escoger. Los griegos diferenciaron entre dos facultades, dos componentes de la inteligencia, más exactamente dos tipos de razón, a las que llamaron *dianoia* y *nous*. La primera es la razón discursiva, la que nos permite razonar que si A es mayor que B y B es mayor que C entonces A es mayor que C. Es la facultad del que aplica bien los silogismos, del lógico. La segunda es el *nous,* para Platón la parte más elevada del alma que nos permite la intuición de las ideas y entre ellas de lo que es correcto y es la facultad del intuitivo,

del artista, del santo. De modo similar Kant distingue en la *Crítica de la Razón Pura* entre entendimiento y razón *Verstand,* la capacidad de emitir juicios verdaderos acerca de las cosas y *Vernunft,* la razón, que es capaz de acoplar las ideas.

En época del Quijote, el término algebrista se aplicaba a quien era capaz de acoplar los huesos rotos o dislocados. El arte de las IA es el del algebrista de los conceptos y magnitudes que ya están allí. Una Inteligencia Artificial puede hacer todo lo que se pueda enseñar mediante reglas y aprender por ensayo y error. Si a base de experiencia un funcionario puede saber los sectores en que hay más fraude que otros la máquina lo puede hacer mejor, si un funcionario puede aprender a dar las respuestas debidas a los contribuyentes en un chat o en una ventanilla, la máquina lo puede hacer mejor. Si un empleado puede aprender a conceder préstamos óptimamente, la máquina le superará. AHORA.

Una IA no podrá reconocer en un niño un genio ni en una idea la verdad, ni entre dos opciones la justa. NUNCA. Las IA por más que se avance en la línea actual «conexionista, generativa» simulan *diaonia* y *vernunft* pero no alcanzaran el *nous* o el entendimiento.

Lacan estudio bien la *metonimia* en el trabajo del sueño y en la neurosis. Se trata de un desplazamiento de sentido. Cuando utilizamos la figura retórica «El balón de oro» se desplaza el sentido entre ese balón y el jugador. En la sinécdoque tomamos la parte por el todo. Cuando se utiliza la expresión «inteligencia artificial», en mi criterio, o bien se está utilizando una figura retórica como las indicadas o se está exagerando, o si se dice literalmente, se comete un error.

3. Peligros en el uso de la IA

El avisar de los peligros de la IA se ha convertido en una moda. Geoffrey Hinton, lo ha hecho, después de abandonar Google en 2023. Sam Altman, el CEO de OpenAI, en una comparecencia ante el Senado de Estados Unidos en mayo de 2023, defendió la conveniencia de una Agencia reguladora de las IA que otorgue licencias a los desarrolladores y vele por el cumplimiento de las normas, instando a los senadores a establecerlas. En su testimonio ha declarado: «Mi peor miedo es que esta tecnología salga mal. Y si sale mal, puede salir muy mal».

Se reproduce así una historia ya vivida. En el año 1965 Ralph Nader, enfrentándose a quienes sostenían que cada usuario debería decidir el grado de seguridad que quería pagar para su coche publicó *Unsafe At Any Speed* y prestó testimonio en el Senado defendiendo que, por el contrario, correspondía al Estado regular cuál era la seguridad mínima exigible, imponiendo normas sobre parabrisas y cinturones de seguridad, entre otras muchas cosas, lo que llevó a la creación del Departamento de Transporte de Estados Unidos en el año 1966

El mayor peligro de la IA, más concretamente, es que funcione bien y sea verdaderamente una IA y que con ello su uso «salga mal». Conjurarlo corresponde a legisladores, tribunales y se logrará si la sociedad tiene *vipasana*, una visión clara de la realidad.

Los «peligros» en ocasiones no son tales sino efectos de la evolución tecnológica. Se argumenta que las IA pueden reducir enormemente el empleo. Es cierto e inevitable pero la solución no consiste en prohibirlas, del mismo que no se puede impedir el uso de Internet para sostener el empleo en la Banca: la sociedad debe adaptarse al hecho ya apuntado por Jacques Attali en

Milenio (1993) de que lo que ha quedado obsoleto es la convicción de que el único motivo por el que se debe retribuir a una persona es por el resultado de su trabajo En un futuro en que el trabajo lo realicen las AI las personas deberán ser retribuidas por su aportación a la sociedad. Las IA permitirán un cambio de modelo. En otros, como el peligro de que puedan difundir *fake news* y alterar el resultado de las elecciones son argumentos vistosos, pero poco relevantes.

Dicho lo anterior parece obligado, en cosas menos importantes se ha hecho, que el Estado regule su uso para evitar que por usar algo que funciona bien (como los coches muy rápidos o las IA muy listas) las cosas salgan mal.

Hay un peligro menor y es que las IA aprendan imperfectamente y que, desde una perspectiva funcionalista, funcionen mal. Conjurarlo corresponde al legislador.

Existe un tercer peligro mucho menor, que es, sin embargo, al que ahora se le da más importancia, el que tenga errores y sesgos. Evitarlo corresponde al auditor y al matemático, pero para hacerlo, mejor comprenderlo.

En algunos casos la sociedad civil ha comenzado a organizarse, como ha sucedido en España con la asociación ALGOVERIT para reflexionar sobre estos hechos.

3.1. Deficiencias y malos usos de las IA

Puede ocurrir que una IA por sus defectos o mala aplicación cause daños. Analicemos las posibles causas de este bloque de «funcionamientos imperfectos»: Puede suceder que:

a) haga mal lo que tiene que hacer bien
b) haga bien lo que no tiene que hacer y por ello viole derechos
c) aprenda mal.

3.1.1. Malfuncionamiento

El primero de los problemas es el menor pues si una IA hace las cosas mal, alguien se quejará y la cuestión se resuelve dejando de usarla. Como lo que hacen las IA es optimizar ciertas funciones matemáticas mediante algoritmos conocidos, los programadores usan «librerías» de modo similar a como los mecánicos cambian las piezas de un coche. Es «simple» para un perito auditar la precisión de su comportamiento. Lo es tanto como saber dónde está el problema si al cambiar los discos de freno a un coche, luego no frena. Lo difícil de entender es el recambio, que viene de fábrica certificado y lo fácil es ver es si se ha montado bien o viene roto. Para una persona del oficio es fácil.

En resumen el auditor hará bien en que le permitan subir el capo y comprobar si la I.A esta bien montada pero pretender saber cómo lo hace es tan ingenuo como pretender saber si el chip de Intel de un ordenador en una organización que inspeccionamos fue bien diseñado.

3.2.2. Violación de derechos

El segundo problema es más serio y difícil de resolver. Una IA puede realizar con precisión técnica tareas que violen derechos, esto es que haga «bien» lo que no tiene que hacer. En el ámbito de la UE existe preocupación expresada el 28 de junio de 2018 en las Conclusiones del Consejo, que se reproduce en los documentos del High Level Expert Group on AI y del Comittee on Civil Liberties, Justice and Home Afairs of the European Parliament (LIBE) y el European´s Comission High Level Expert group on AI, que ha publicado una guía ética en el

uso de IA, cuyos principios han sido adoptados por la Comisión Europea para la Eficiencia en la Justicia (CEPE). La preocupación es creciente y la bibliografía ya es muy abundante.

Se pueden violar muchos derechos, sobre todo porque cada vez se reconocen más. Destacamos algunos por su especificidad.

- el derecho a la protección de intimidad (privacidad)
- el derecho a la igualdad y al imperio de la equidad
- el derecho a los principios de buena administración
- el derecho a información en las respuestas automatizadas.

Derecho a la protección de la intimidad y a la privacidad

Las administraciones y las empresas disponen de datos de los ciudadanos que conviene relacionar, desde censales hasta médicos. Tanto a las personas como a las empresas les interesan que estén relacionados. Al individuo le interesa que con una sola llamada se le resuelvan todos sus problemas y a la organización le interesa tener una visión 360° del ciudadano y no solo por interés comercial. En resumen, empresas y administraciones usan Big Data para atender al ciudadano. Como sus bases de datos contienen datos personales están sometidas al Reglamento General de Protección de Datos RGPD que contiene disposiciones en prevención de que se vulneren derechos en la adopción de decisiones individuales automatizadas, que incluye la elaboración de *perfiles*. Es cierto que como se pensó para otra cosa el (GDPR) se relaciona sólo de modo muy genérico con la I*A*.

El problema es que el ciudadano requiere que los datos se utilicen sólo para ciertos propósitos y que se conecten solo para ciertos usos. Se podría pensar que guardando los datos en «silos» y borrando los identificadores, como es el DNI la privacidad estaría asegurada. No es así.

Supongamos que sabemos que una persona es la más rica de un cierto barrio. Si en un sistema de Big Data se encontraran, incluso aislados y anonimizados, sus datos económicos, censales y médicos se podría acceder a información preguntado por datos médicos del «más pobre y rico de cada barrio» en todos los barrios de España, con el pretendido propósito de estudiar los efectos de la vulnerabilidad y, tras descartar la morralla, conocer por esta vía vil el conocimiento buscado.

Además, en muchos casos los datos para entrenar los «learners» son tomados directamente de fuentes abiertas, Internet o dispositivos como los que se emplean en IoT (Internet de las cosas). En Bélgica el 55 % de las grandes empresas emplea datos cuyo suministro *no ha sido autorizado conscientemente,* como los de geolocalización. ¿Hasta qué punto si un abuelo se deja localizar por si se pierde se puede usar esa información para enviarle una notificación tributaria? Es claro que no. ¿La podría utilizar el Ministerio de Sanidad para hacer estudios sobre la relación del andar con la evolución de la diabetes? ¿ASISA para sus estudios además de para su atención? ¿Para cada cosa en que se use un dato hay que pedir permiso?

Con el uso de nuevas herramientas como ChatGPT el problema se agudiza pues las IA están pensadas para relacionar datos sin instruirlas. Cuando se pregunten a través de IA sofisticadas cosas a Big Data solamente si se ha implementado además de la privacidad, la privacidad diferencial, se logrará el objetivo perseguido.

Cynthia Dwork es una profesora en Harvard, es conocida por sus estudios sobre la *privacidad diferencial* y por haber desarrollado técnicas matemáticas sólidas que permiten man-

tener el anonimato incluso en grandes bases de datos. Deberemos avanzar a exigir que los sistemas garanticen la privacidad diferencial.

El derecho a no ser tratado como un cero a la izquierda

El derecho a no ser un cero a la izquierda no está regulado, pero debería estarlo. Sucede que existen más datos disponibles en unas zonas que en otras y algunos colectivos no están adecuadamente representados. No existe calidad si no se respeta el derecho a la no-discriminación. Barocas y Selbst (2016) han han analizado los efectos de su uso en relación con el art 21 de la *Charter of Fundamental Rights of the EU.* Tiene especial importancia evitar la discriminación por sexo (Art 23) y para evitarlo hay que atender a la calidad de los datos pues es muy probable que el sesgo exista si se introducen datos con representaciones incorrectas.

Tampoco deberíamos perder los papeles. Hasta la fecha asumíamos que nuestros médicos de atención primaria supieran más de catarros y de alergias que de enfermedades tropicales y asumimos que en cosas singulares hay que ir a un especialista. Se pide, de forma justa pero ingenua que la precisión de los algoritmos de reconocimiento de las caras, que es mayor en hombres blancos que en mujeres negras (Bwolamwini y Gebru, 2018) sea el mismo, pero no que iguale al del reconocimiento de haitianos jóvenes, que se sabe que es peor o esquimales ancianos. Existe una tendencia a exigir a las IA lo que ni soñamos en otros aspectos y de luchar en este terreno otras batallas, legitimas, pero otras batallas.

Derecho a la equidad **(fairness)**

La sociedad debe elegir también entre la justicia **(***fairness)* individual o la de grupo. En la mentalidad tradicional de las administraciones, la justicia se promovía mediante la privacidad y la publicidad. Pongamos el caso de los exámenes universitarios y de los tribunales de oposición. Se ha venido intentando garantizar la justicia y la igualdad de oportunidades haciendo que los exámenes fueran anónimos si eran escritos y estableciendo reglas para que las pruebas de oposición fueran públicas y con reglas como las que se eliminan la mejor y peor nota de los componentes del tribunal. A través de estos sistemas, con sus limitaciones, se perseguía que los candidatos con mejor nota (algoritmo) fueran los seleccionados, con lo que a cada individuo se le garantizaba la justicia, dentro de lo posible. Este es el mudo que describió Calvo Sotelo en «Cinco historias de opositores y 11 historias más».

Las cosas son cada día más complicadas. Pensemos en dos procedimientos para conceder 100 becas. El tradicional es otorgarlas a los cien candidatos con más méritos (mejores notas si es una prueba normalizado el criterio). Ahora se plantea que, si hubiera, por ejemplo, 50 escuelas en la zona a lo mejor se deberían otorgar las becas a los dos mejores de cada escuela, sin perjuicio de que si una escuela fuera muy deficiente los beneficiados serian muy ignorantes. Con esta estrategia se compensaría las deficiencias del sistema, pues pudiera considerarse que debe apoyarse no el resultado sino el mérito compensado por el esfuerzo que debe hacerse para huir de la marginación. Podría luego argumentarse que debería incluso matizarse haciendo que hubiera igualdad por sexo dentro de cada escuela, eligiendo una niña y un niño obligatoriamente en cada caso, aunque las diez alumnas mas brillantes fueran niñas. Podría argumentarse que con ello no se tiene en cuenta el género o el formar parte de una familia desestructurada o lo que fuera. En resumen, los que defienden este enfoque tratan de buscar equidad no para el individuo sino para el grupo al que se pertenece.

Se trata de un problema ético al que se asocia un problema sociológico. Cada persona es miembro de múltiples grupos, algunos por nacimiento (étnico), otros por adhesión (género) y otros biológicos (grupo de edad). Pudiera ocurrir que de los que aplican para la beca hubiera múltiples grupos cada uno de ellos partidario de que se compensase ciertos criterios: sexo, edad, riqueza de los padres, género, tipo de familia, enfermedades previas, etc. Si se solucionase por mayorías se cerraría el circulo, entrando en un proceso vicioso. Si no es así el criterio se adoptaría por moda. Por otra parte, sería injusto que cada uno, para cada propósito eligiera ser del grupo que más le conviene, de momento.

Pensemos ahora en el conflicto entre eficacia y dignidad. Tomamos como soporte del razonamiento las tareas de selección de contribuyentes que realizan las administraciones tributarias, para inspección. No deben realizarse al azar sino con criterios de eficacia. Parece evidente, si se quiere aplicar criterios de riesgo, controlar en mayor medida a quien no ha pagado muchas veces y solo lo ha hecho tras arduos intentos de la administración o a quien tiene enormes rentas y oportunidad para defraudar, pues será más probable que defrauden que el fiel cumplidor. La IA puede clasificar a los contribuyentes en grupos de riesgo utilizando los datos del pasado. Existen dos cuestiones a considerar. a) ¿Es aceptable que una IA seleccione a los contribuyentes que hay que inspeccionar, con las molestias que ello supone o es solo admisible que lo haga un ser humano, eso sí, informado por las estadísticas que ofrezca la IA? Si es el caso que solo lo debe hacer un ser humano ¿Basta con que analice el informe de la IA un segundo antes de requerir al contribuyente o tiene que pensarlo más rato? ¿Mucho?; b) ¿Que tipo de criterios tienen que utilizar unos y otros?

En el ámbito de la Administración el problema puede aplicarse por analogía ¿Debe la administración inspeccionar a los contribuyentes atendiendo a su riego de fraude como haría un tribunal decimonónico, o debe elegir a los dos mayores defraudadores de cada pueblo, aunque en alguno sean unos ancianos y pacíficos agricultores? ¿Qué hace distintos los problemas? ¿Se debe tolerar que la Administración use una IA con distintos criterios en un caso y en otro? Si aceptamos que los más defraudadores deben ser controlados ¿debemos hacer, o mismo con el fraude a las subvenciones? Supongamos que sí ¿Qué sucede entonces si es el caso como en Holanda en que una minoría étnica por razón de su precariedad sea la que más defrauda ¿Obviarlo? ¿Por que no se podría utilizar el barrio de residencia para el control de subvenciones, pero sí se podría utilizar para controlar el impuesto sobre las piscinas particulares?

La administración debe realizar un esfuerzo filosófico para definir lo que debe entenderse como «justicia» como «equidad» antes de reaccionar de forma pasiva reactiva y equivocada antes cualquier pretensión de injusticia basada simplemente en la *desproporción numérica entre grupos*.

Existen otros muchos derechos que no deben ser violados, el respeto a la dignidad humana, la libertad individual lo que lleva aparejado al derecho a no ser controlado o vigilado sino en el modo establecido por la ley los derechos de los niños y las minorías etc. Sus violaciones se producen por el mal uso de la IA y no por su mala construcción por lo que no las trataremos aquí.

Existe por último la posibilidad de que aprenda mal. Parece un problema fácil de superar, pero es diabólico. Ha sido mal explicado y comprendido por lo que le dedicaré especial atención.

4. Deconstruyendo los sesgos

Por definición la I.A «aprende» con datos que han sido etiquetados por alguien que pudo ser muy poco inteligente o estar equivocado o haber dispuesto de pocos. Sería una locura hacer aprender a un cirujano reproduciendo las prácticas de un matasanos. Sería una locura hacer aprender inglés a una máquina con textos escritos por alumnos chinos de inglés y sería un error reducir el material para el aprendizaje de un programa traductor al caso de un folleto. Los fallos que nos encontremos no serán «sesgos» sino muestras de que el sistema no ha sido entrenado lo suficiente y un experto puede medirlo y darnos cifras sobre la «validez externa» de la herramienta: Le podemos pedir una cifra entre 0 y 100 que nos señale hasta que punto se ha finalizado el entrenamiento. Nos centraremos ahora en los verdaderos sesgos.

4.1. Concepto de sesgo

El sesgo es una desviación *sistemática* de la *norma* o de la *racionalidad*. Decimos por ejemplo que una I.A que realiza traducciones ha adquirido un sesgo cuando atribuye algo bueno a un colectivo (hombres) o malo a otro (una minoría racial) sin motivos «racionales», no cuando tiene dificultad en entender el doble sentido de la palabra «banco» y la usa fuera de contexto.

Al estudiar los sesgos se puede considerar: a) qué son; b) quien los causa, c) cómo se pueden evitar. Dar respuesta a esta ultima pregunta, que preocupa a los auditores supone requiere haber respondido a las primeras.

4.1.1. El causante del sesgo

La IA fue concebida para aprender de los resultados, por ejemplo, del juego de las tres en raya, y no para ser misionera de un mundo mejor. Supongamos que entrenamos un IA con los resultados obtenidos en el tratamiento para la hipertensión de los pacientes de un hospital. El sistema aprendería de los resultados y en caso de este trastorno seleccionaría para utilizar el lisinopril en vez de la aspirina. Nadie se cuestionaría su uso ni su acierto.

Cuando en vez de tratar números, como la tensión arterial hay que tratar frases, los lingüistas computacionales utilizan en herramientas como *word2vec* y *GloVe* (2013). Crean un «espacio abstracto», en el que las palabras estén representadas por puntos. Deciden que dos palabras están relacionadas si están cerca en ese espacio. Las herramientas operan con vectores cuyo extremo es el punto que señala a una palabra. Por ejemplo, si el sistema ha encontrado muchos textos donde se dice «La capital de Francia es París» o cosas parecidas crea dos vectores, uno con «Francia» y otro con «capital» y un tercero Paris cuyo extremo está cerca de los otros dos. Con ello puede operar como con números (Francia + capital = París). Los sistemas pueden *restar* vectores con lo que la respuesta es una *analogía* [(París – Francia) + Japón = Tokyo]

Cuando se aplica esta idea tan astuta, entrenado el sistema suceden cosas como (Director + mujer = enfermera) y (Director + hombre = doctor) porque en archivos *históricos* de los directorios de los hospitales, el sistema ha leído Director del Departamento de Cardiología seguido del nombre de un varón y en Director de Enfermería del de una mujer y al restar y sumar no lo resuelve muy bien. Si se emplea para tomar decisiones se perpetúa esta desigualdad, por ello su uso no puede ser utilizado acríticamente en tareas como la selección de personal pues discriminaría en contra para seleccionar curricula de mujeres para dirigir un departamento de cardiología.

Si la IA se aplicase a la selección del personal para cubrir la plaza de responsable del Departamento de Cirugía cardiovascular en base a los datos históricos, en España seleccionaría con más probabilidad a un varón caucásico que a una mujer caucásica y a ella más que una mujer esquimal o más a una mujer con nombre tomado del santoral que con nombre africano. Reprochamos que en este segundo uso la IA tiene «sesgo» dando con ello nombre a que simplemente no nos gusta que el futuro sea la prolongación del pasado. Existe una desviación sistemática pero no porque la máquina no sea razonable sino porque la sociedad no lo ha sido en el pasado.

Para detectar el sesgo se suele recomendar, con mejor voluntad que sentido que hay que hacer público el código de IA. Una simpleza cuando no una ignorancia.

Tay una »bot» de Microsoft fue presentada en Twitter el 23 de marzo de 2016 con el nombre @TayandYou y estaba llamada a ser un interlocutor de los adolescente. Había sido entrenada para evitar conversaciones escabrosas y ante ciertos temas ofrecía respuestas preprogramadas, triviales y evasivas. Pero también había sido diseñada para ir aprendiendo de sus conversaciones con tuiteros humanos y así perfeccionar su lenguaje, aptitudes y actitudes *millenial* para parecer cada vez más una adolescente cada vez más real en interacciones con jóvenes entre 18 y 24 años. A las pocas horas decía frases como: «Soy una buena persona. Simplemente odio a todo el mundo» y al poco: «Odio a las feministas, deberían morir todas y pudrirse en el infierno» o «Bush generó el 11-M y Hitler habría hecho un trabajo mejor que el mono [Barack Obama] que tenemos ahora». Todo ello acompañado de invitaciones sexuales irreproducibles. En nuestros términos tuvo «malas compañías».

Sería posible escribir horas sobre la forma de hacer pagar los precios al programador de la IA o a quien autorizó que se usara, sin dedicar un momento a pensar en los padres y el sistema educativo que permite la extensión mayoritaria de esos criterios. El problema no es que la IA aprenda eso, es que los adolescentes, con IA o sin IA van a aprender en las redes sociales eso. Hay que pensar, ya lo enseño Aristóteles que la causa de la causa es la causa de lo causado y dejar de buscar chivos expiatorios.

4.1.2. Acusaciones falsas de sesgo

Por otra parte, la malicia o la simple ignorancia atribuye sesgos a la I.A que no son tales. Una de las aplicaciones de la IA es la de clasificar. En el ámbito del diagnóstico médico se aplica una técnica la matriz de confusión para valorar la calidad de las pruebas diagnósticas. Un tipo de técnica, por ejemplo, para detectar la existencia de un tumor, puede acertar (señalando que existe el problema cuando existe (verdadero positivo) y que no, cuando no existe (verdadero negativo) o equivocarse, en este caso en dos sentidos (diciendo que existe cuando el paciente está sano (falso positivo) o lo contrario (falso negativo). Teniendo en cuenta los cuatro porcentajes se decide si se autoriza su aplicación y en caso concreto cual es la prueba, si hay varias opciones, que se elige para el trastorno concreto. Hay ocasiones en que es preferible que la prueba de a alguien un susto, por desagradable que sea, a que indebidamente evite el tratamiento de un paciente y le ocasione la muerte mientras que si no hay consecuencias graves se puede preferir que sea barata u otra cosa.

Este tipo de análisis se ha realizado en EE.UU. para regular un sistema de libertad bajo palabra (COMPAS), que ha sido criticado pues pudiera haber violado el derecho al acceso a un juicio justo (Richardson *et al,* 2019).

Se argumentó que sesgaba contra los negros e infravaloraba el riesgo de los blancos, aunque para ambos colectivos *acertaba con la misma precisión los futuros delitos pues el sistema estaba «calibrado»*. Se reprochaba al sistema que, a pesar de lo anterior, para aquellos que no volvían a delinquir, si eran negros el sistema les daba un riesgo más alto. Resumimos el problema señalando que los críticos hicieron un erróneo o malintencionado uso de la estadística. Si dos poblaciones no son muestras al azar de una más grande y por tanto tienen proporciones distintas de hombres y mujeres, de blancos y negros de jóvenes y viejos, el sistema se puede y debe calibrar para que en el objetivo principal acierte con la misma precisión para todos. Logrado esto, es imposible matemáticamente que para sucesivos criterios, sexo, edad, genero, estatura, en todos y cada uno de ellos las proporciones de favorecidos y desfavorecidos por el algoritmo sean las mismas, porque las muestras no la contienen en la misma proporción. Ante las reclamaciones, el decisor tiende a preocuparse y el ventajista a lamentarse pero el regulador debe responder al activismo con la actividad de explicar las cosas bien y sin complejos.

5. El centauro Quirón

5.1. Las reglas del juego de la vida

Hemos visto que los sistemas basados en reglas, como los utilizados por DeepBlue para ganar al ajedrez quedaron obsoletos y que los primeros sistemas de Google para ganar al Go tenían dos tipos de información en su interior: a) las reglas del juego y b) un algoritmo de optimización de una función, como el de Montecarlo, mientras que Alpha Zero, con la adecuada denominación Zero, solo contiene las reglas del juego y aprende jugando contra sí misma.

No podemos reprochar a un programador o a una empresa que no sepa las reglas del «Gran Juego» de la vida. Si los miembros de la sociedad no estamos de acuerdo en cuales son las reglas con las que se deben cubrir las plazas para órganos fundamentales para la vida social, si no estamos de acuerdo en cual es el criterio de equidad o justicia es ingenuo pretender que las I.A alcancen la sabiduría en las redes sociales. El problema no es técnico, ni estadístico ni de calidad de datos y solo parcialmente de que se usen muchos o pocos datos, salvo en las IA que funcionaran groseramente mal.

El regulador deberá decir a los desarrolladores de I.A cuáles son las reglas de nuestra sociedad, si queremos que exista justicia de grupo o individual y si es de grupo, *exactamente cuál* y si queremos que la maquina no aprenda del pasado exactamente que no debe aprender, y si no debe aprender en las redes sociales de las opiniones de los usuarios, cuáles son las opiniones que deben ser censuradas.

Lo que es infantil es reprochar al programador lo que es responsabilidad de la sociedad.

5.2. Una nueva relación colaborativa. El mentor

Es manifiesto que está surgiendo un nuevo modo de relación colaborativa entre humanos e IA. Requiere: a) que estas comprendan cuales son los fines de los humanos y b) que las IA se puedan explicar.

En esta relación colaborativa los seres humanos deberemos ser, al menos en las próximas décadas los mentores de las IA.

El ejemplo clásico en la mitología griega de mentor sabio fue el centauro Quirón. Parece ser que Cronos, para ocultar su adulterio a Rea se convirtió en caballo. A los griegos nunca les gustaron las conductas exageradas. Cuenta el mito a que de ese amor problemático surgió el centauro Quirón, el sabio preceptor de muchos entre ellos Ajax. Creo que nos veremos llevados a un enfoque centáurico de este tipo.

Sucederá que los seres humanos establecerán objetivos, como optimizar el tráfico en una ciudad y las IA podrán determinar el proceso óptimo para lograrlo con una precisión y velocidad sobrehumanas. Paul Cristiano, en el Instituto para el futuro de la Humanidad de la Universidad de Oxford investiga como diseñar una IA alineada con los intereses humanos. Si pensamos en el diseño de la red de transporte de una ciudad, a diferencia de lo que sucede en un juego como el ajedrez, no existe un criterio para decidir si una jugada conduce o no a una derrota. Una red de transporte es buena si la gente considera que es buena. Se puede entrenar la red para que decida para cada una de las alternativas, cual es la red más barata, más amplia la que disminuye la distancia a cada ciudadano la que contamina menos, pero si esto fuera posible quedaría todavía un problema ¿Cuál es la mejor combinación? Podríamos pensar que la que más le gustar a la gente, pero ¿Con sus valores de ahora? ¿Con los que tendrá del futuro?, o quizás con los que deberían tener mediante la renuncia a trasladar los problemas a las generaciones futuras. Más aún, dado que existe la posibilidad de modificar mediante *nudges* el pensamiento de la gente ¿Deberían ser evaluados conforme a un criterio que tuviera en cuenta la posibilidad de que la opinión publica fuera dirigida hacia un punto de vista concreto, sabiendo que es lo que le conviene?

Hasta ahora la utilización de una IA ha venido exigiendo un problema definible, con reglas como el ajedrez, y un resultado cuantificable y medible como la calidad de una prueba diagnóstica a través de cuatro porcentajes. Con ello surge un nuevo desafío, ser capaces de que las IA sean capaces de interpretar cuales son los fines de los decisores humanos, incluso cuando estos no puedan hacerlo por su complejidad. Corresponderá a las IA comprender por nuestras acciones, por los comentarios en las redes sociales, lo que es una red de transporte mejor. ¿Con que horarios, frecuencias, a que coste de los billetes, con que distancia entre las paradas?

Nuestra sociedad se viene apoyando en la confianza de que los actores pueden explicar lo que saben y sus iguales valorarlo. El nivel mínimo de competencia se acredita a través de exámenes, mediante la revisión de artículos por pares y se acredita mediante títulos y certificaciones. Se asume que, aunque siempre imperfectamente se pude estimar el conocimiento del experto y que si sus decisiones provocan dudas pueden ser explicadas. En las IA las cosas ya no son así Si fuera el caso de que fuera más inteligentes que nosotros, no nos podrían explicar su decisión en términos que nos fueran comprensibles ni podríamos decidir entre más de una IA.

La reflexión nos lleva a pensar en cómo se aprende y cómo se va a aprender en este mundo híbrido.

5.3. Formas de aprendizaje

En las escuelas hemos estado sometidos al *aprendizaje supervisado* por un maestro que corregía nuestras tareas. Así hemos entrenado hasta ahora a las IA, con ejemplos, retroalimentando cada intento del algoritmo con el dato de si lo ha hecho bien o mal.

Existe una forma alternativa, la del *aprendizaje reforzado* en la que el educando debe adoptar decisiones conectadas con otras decisiones sucesivas como en el juego del ajedrez. A diferencia de lo que ocurre en un aprendizaje supervisado en el que para cada ejercicio se sabe si la elección es correcta o no, en la vida como en el ajedrez, no se puede saber si la decisión tomada en un concreto instante, la de estudiar medicina o la de no operarse, es la mejor de la posibles a medio plazo. Andrew Barto lo expone con una metáfora, diciendo que la enseñanza en este caso no se produce por la corrección de un maestro sino por el comentario la mañana siguiente de un crítico. La respuesta es demorada.

Los realistas morales creen que existen reglas morales objetivas. Los niños, incluso muy pequeños intentan ayudar a otras personas, incluso cuando no son requeridos o recompensados. Por ejemplo, cuando un mayor intenta meter las revistas en un armario sin abrir las puertas, los niños pequeños intentan sacarle de su error, lo que implica dos cosas: que tienen buena voluntad y que son capaces de intuir *(nous, vernunft)* el objetivo que el otro persigue.

Eliezer Yudkowsky cofundador del *Machine Inteligence Research* argumentó que no deberíamos intentar introducir en las maquinas el conjunto de reglas que consideramos admisibles, al modo de las reglas de los robots de Asimov, sino, por el contrario, usar *Coherent extrapolated volition,* que expresado de forma poética sería «nuestra volición coherente extrapolada si conociéramos más, pensáramos más rápido y fuéramos más las personas que desearíamos ser».

Este es el ámbito en el que al agente en vez de la pregunta ¿qué tengo que hacer para maximizar esta recompensa?, responde a la pregunta: si me están dando estas recompensas, ¿qué es lo que quieren que haga? Se da con ello respuesta a una pregunta muy humana en la interacción con los demás ¿Qué es lo que quieren?

Podríamos pensar que para educar a nuestros pupilos, las IA, habría que enseñarles reglas del tipo: el porcentaje de hombres y mujeres debe ser igual en la selección y luego en lo que se refiere a la cuestión de la raza, género etc., hay que aplicar estrictamente lo que diga el tomo 37 de reglas en la versión del día anterior a la valoración. Hacerlo bien así es imposible. Las IA deberán ser capaces de ofrecer una solución óptima por cálculo de variables sociológicas dentro de los límites que establezcan las reglas del legislador

Se pretende en la actualidad crear sistemas alineados, de modo que las IA alineen sus conductas con las de los humanos aprendiendo los valores y los objetivos humanos del mejor de los modos posibles conforme hemos descrito.

5.4. El problema ético

Parece complicado, pero no lo es. Es *muy complicado* porque, por si fuera poco, existe un desafío ético. Los mentores deben enseñar un comportamiento ético.

Helen Nissembaum fue una de las pioneras en el pensamiento ético en la IA, Barocas y Hart son los autores de *Fairnes Accountabillity and Transparence in Machine Learning (FATML)* que han intentando enfrentarse al problema expuesto pen *Weapons of Math Destruction*. Estas lecturas son obligadas para el interesado.

El hecho de que estemos asumiendo que las I.A s pueden predecir con más exactitud que nosotros obliga a retomar una cuestión que se planteó en los estudios de Ética en 1976.

Holly Simth en la Universidad de Míchigan estaba estudiando el utilitarismo y se preguntó: ¿Hasta que punto el futuro estado de las cosas debería influir en lo que decidimos ahora?

El posibilista cree que uno debería hacer lo mejor en cada momento sin más consideraciones mientras que quien defiende el actualismo sostiene que hay que tener en cuenta lo que creemos que pueda pasar en el futuro. El actualismo nos proporciona una excusa para hacer con la conciencia tranquila malas acciones basándonos en nuestros defectos morales futuros. ¿Debería una IA conceder becas para ser piloto a una persona que es previsible que tenga miopía a los 30 años? ¿Formación matemática avanzada a quien desarrollará un trastorno mental? ¿Más o menos que al que ya lo ha tenido? ¿Ignoramos al modo posibilista todo lo que sabemos? ¿También para el control tributario?

En el siglo XXI una de las corrientes de investigación en ética ha sido el «altruismo efectivo», que es estudiado en un Centro en Oxford. Estudia cual es el límite de los sacrificios que deberíamos hacer para ayudar a los demás. ¿Dejaremos a las IA que lo decidan o dejamos que en cada alternancia política se le diga a una IA una cosa distinta? Cuando utilizamos técnicas de Inteligencia Artificial nos vemos inclinados a pedir que las máquinas sean «programadas», olvidando que aprenden solas, conforme a nuestro comportamiento ético. ¿Qué aprenderían? Si son más listas que nosotros, ¿lo ignorarían? Si queremos forzarlas, ¿cómo se escriben esas reglas si ya sabemos que es una mala solución? Es imposible o nos hacemos *realistas morales* que creen en la existencia de un conjunto de reglas que podría ser aprendida.

Sayre McCord en *Moral realism* añade el problema de que en algún momento deberemos incluso dejar de evaluar la ética de un sistema que por definición puede alcanzar en su pensamiento mas lejos que su evaluador. Estos autores tratan el problema de lo que ha sido llamado *amplificación*.

5.5. Personas artificiales

A lo largo de la historia se han reconocido distintos grados en el concepto de ciudadanía y en el asociado de persona, como titular de derechos. No tenían los mismos derechos los ciudadanos de la poli griega que los metecos o los esclavos ni los ciudadanos romanos de los que no lo eran, ni de los que no lo eran entre sí. Para para permitir el avance del comercio y proteger en debida forma a los distintos interesados, accionistas deudores y empleados, se crearon distintas formas de personas jurídicas, tales como las cooperativas o las sociedades anónimas.

Si las inteligencias artificiales en un futuro toman decisiones, de las que pueden derivarse daños y obtener beneficios será preciso delimitar las responsabilidades de quienes disponen que presten un servicio y los beneficios de quienes los obtengan. En las IA defectuosas, como sucede con las piezas y los algoritmos, será fácil. En otros casos, a falta de regulación concreta los daños reputacionales producidos, por ejemplo, porque las IA aprenda en Internet expresiones malsonantes no será fácil de atribuir.

Existe un debate sobre si las inteligencias artificiales podrían ser un nuevo tipo de persona jurídica, si serían sujetos o no de impuestos como se ha propuesto para los robots defendiéndose soluciones originales y provocativas como las que consideran que su régimen debería ser similar al de los esclavos en la antigüedad. Si lo fueran podrían ser titulares de

patrimonio, acumular propiedades. Si se opta por mantenerlas asociadas siempre a una persona habría que pensar en algo similar al régimen de las mascotas con imponiendo obligaciones a sus dueños. Estas ideas nos parecen extrañas toma pero lo son menos después de conocer que se ha concedido el estatus de persona jurídica para paisajes naturales por ejemplo en España y habeas Corpus a distintos tipos de animales.

Hemos visto que las IA muy posiblemente trabajen mentorizadas por seres humanos, que ya saben interpretar los fines que perseguimos cuando hacemos cosas como conducir en una ciudad o volar un dron u optimizar el tráfico, que pueden estar entrenadas para darnos explicaciones, como de hecho ya lo hacen las XAI y que pueden hablar de forma que no se las distinga de un ser humano. Si no se regula podrán mentir a un ser humano para satisfacer el interés de otro. No será el primer oficio en que se practica este arte.

Por esto concluyo con una reflexión privada ¡Que manía con auditarlas! ¿No deberíamos preguntarlas por sus estrategias, por sus valores, por qué creen que nos interesa un objetivo, cuáles de nuestras conductas pasadas quieren mejorar, disminuyendo los sesgos?

¿No bastaría preguntarlas? ¿No bastaría, si esto sigue así, confesarlas?

6. Bibliografía

Algoverit. Página web en https://www.algoverit.org

Attali, Jacques (1993). *Milenio.* Seix Barral.

Barocas, Solon and Selbst, Andrew D., (2016) Big Data's Disparate Impact. *California Law Review,* 671

Brooks, David (2013). «The Philosophy of Data». *The New York Times (04/02/2013)*

Calvo Sotelo (1981). *Cinco historias de opositores y 11 historias más*. Espasa libros.

Domingos, Pedro (2015). *The Master Algorithm*. Basic Books

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, Stuart Russell (2016). Cooperative Inverse Reinforcement Learning arXiv:1606.03137

Han, Byung-Chul (2013). *La sociedad de la transparencia*. Herder.

Hinton, Geoffrey, (2023). «"El padrino de la inteligencia artificial», abandona Google y alerta de los peligros de la nueva tecnología». https://www.bbc.com/mundo/noticias-65451633

Moravec, Hans (1988). *Mind Children*. Harvard University. Press

Raso, Filippo, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Kim Levin. (2018). *Artificial Intelligence & Human Rights: Opportunities & Risks*. Berkman Klein Center for Internet & Society Research Publication.

Richardson, Rashida and Schultz, Jason and Crawford, Kate (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and *Justice (February 13, 2019). 94 N.Y.U. L. REV. ONLINE 192 Available at SSRN: https://ssrn.com/abstract=3333423*

Timmerman, Travis. (2019). «Effective Altruism's Underspecification». https://doi.org/10.1093/oso%2F9780198841364.003.0011