NO SOY UN ROBOT:

Construyendo un marco ético accionable para analizar las dimensiones de impacto de la Inteligencia Artificial

Micaela Mantegna

CETyS

NO SOY UN ROBOT:
CONSTRUYENDO UN MARCO ÉTICO ACCIONABLE PARA ANALIZAR LAS DIMENSIONES DE IMPACTO DE LA INTELIGENCIA ARTIFICIAL

Autores: Micaela Mantegna

Diseño: Mónica Castellanos

TABLA DE CONTENIDO

El presente artículo corresponde a un capítulo del libro "**ARTEficial: creatividad, inteligencia artificial y derecho de autor**", próximo a ser publicado en 2022

Sobre la autora
MICAELA MANTEGNA

Conocida como la "Abogamer", Micaela Mantegna es una experta mundialmente distinguida en el campo de la ética de la inteligencia artificial, videojuegos, y el metaverso; habiendo presentado ponencias en más de 25 países en conferencias como RightsCon, GamesBeat Summit, Game Developers Conference, Vancouver Biennale, More Than Just a Game, Internet Freedom Festival, entre otras.

Desde 2017 es investigadora afiliada al Centro de Tecnología y Sociedad de la Universidad de San Andrés, en donde se encuentra a cargo del curso de "Derecho e Inteligencia Artificial" dictado en el marco del Programa Programa de Derecho y Tecnología de las Comunicaciones (DITC).

Actualmente se encuentra también afiliada al Berkman Klein Center for Internet and Society de la Universidad de Harvard, liderando el Video Game and XR Policy Working Group.

Por su rol como académica y activista de los videojuegos, fue nombrada TED Fellow en el año 2022, y Google Policy Fellow en 2017.

Es la fundadora de Women In Games Argentina, y embajadora de Women in Games International (WIGJ), reconocida como "Embajadora Individual Destacada del Año" en los Women in Games Global Awards 2021. Asimismo fue elegida por GamesIndustry.biz como una de personas destacadas en la industria de videojuegos en 2021, en su lista anual de GameChangers.

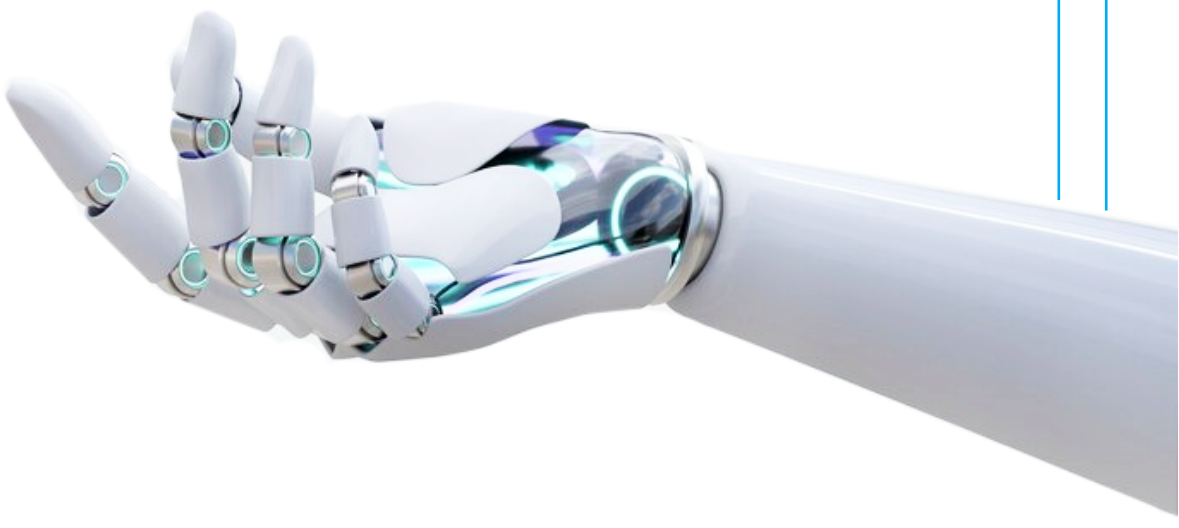
Fundadora de Geekylegal, un podcast e iniciativa de divulgación del derecho y tecnología a través de la cultura pop.

CETyS

NO SOY UN ROBOT:

Construyendo un marco ético accionable para analizar las dimensiones de impacto de la Inteligencia Artificial

Micaela Mantegna



"Never trust anything that can think for itself if you can't see where it keeps its brain."

Arthur Weasley,
Harry Potter and the Chamber of Secrets

I put my faith in a digital world where they've given me eyes without a face.

Jamiroquai,
Automaton

It sits there looking at me; and I don't know what it is. This case has dealt with metaphysics; with questions best left to saints and philosophers. I am neither competent nor qualified to answer those. But I've got to make a ruling, to try to speak to the future. Is Data a machine? Yes. Is he the property of Starfleet? No. We have all been dancing around the basic issue: does Data have a soul? I don't know that he has. I don't know that I have. But I have got to give him the freedom to explore that question himself. It is the ruling of this court that Lieutenant Commander Data has the freedom to choose.

Captain Phillipa Louvois,
Star Trek TNG, The Measure of a Man

ÉTICA DE LA INTELIGENCIA ARTIFICIAL

Dimensiones de impacto de la Inteligencia Artificial. Construyendo el marco teórico para un despliegue ético de la Inteligencia Artificial

Cuando nos referimos a Inteligencia Artificial, no podemos dejar de ponderar su impacto transversal sobre la sociedad. Distintas aplicaciones de Inteligencia Artificial permearon en todas las industrias y sectores, tanto en el ámbito público como en el privado, acelerando la transformación digital mediante la toma de decisiones basadas en datos. Sin embargo, la bonanza que estas técnicas han traído en temas de reducción de costos y de escalabilidad tiene su contrapunto en la potenciación de sesgos, en la discriminación sistémica, en la hipervigilancia y en la alienación[1].

A pesar de que este libro trata sobre usos creativos de la automatización, flaco favor sería no introducir al lector en los cuestionamientos más acuciantes que debatimos en cuanto a la ética de la Inteligencia Artificial. El esquema aquí propuesto se funda en el modelo de análisis basado en dimensiones, que diseñé en 2017 mientras investigaba sobre gobernabilidad algorítmica en el transcurso de mi estancia como Google Policy Fellow en el Centro de Tecnología y Sociedad de la Universidad de San Andrés[2], y que fuera parcialmente receptado en el proyecto del Plan Nacional de Inteligencia Artificial para la República Argentina, en el apartado sobre ética y regulación.

Dada la complejidad y las múltiples ramificaciones de los impactos de la IA, el modelo propone una disección incremental de factores, que abarcan desde el ámbito individual del saber a partir del cual interactuamos con dispositivos basados en IA, hasta el impacto colectivo que esta tiene sobre la equidad y la inclusión social.

Aunque muchas veces se disfrace de innovación tecnológica -palabra de moda o afín al marketing-, hay

1. Estudios recientes publicados como resultado de una investigación del All-party Parliamentary Group (APPG) de UK reclaman una modificación en las leyes laborales que contemple el impacto que el uso de sistemas automatizados tiene sobre la salud mental de trabajadores sometidos a monitoreo y a asignación de tareas de forma automatizada. *https://www.theguardian.com/technology/2021/nov/11/algorithmic-monitoring-mental-health-uk-employees*

2. El Centro de Estudios de Tecnología y Sociedad (CETyS) de la Universidad de San Andrés (UDESA), trabaja en el área de Inteligencia Artificial desde el año 2017, con la conformación del Machine Intelligence Lab (MI Lab) como resultado del Google Policy Fellowship de dicho año con investigación sobre gobernabilidad algorítmica, y con la participación en el Simposio Global de Inteligencia Artificial e Inclusión, organizado por el Network of Centers, celebrado en Noviembre de 2017 en Río de Janeiro, Brasil. Como resultado, se incorporó la materia de IA y Derecho como un nuevo módulo en el Programa de Derecho y Tecnología de las Comunicaciones (DITC) del año 2018 a cargo de Pablo Palazzi y Micaela Mantegna, y clases en materias específicas de grado de la carrera de Derecho y de la Maestría en Innovación y Propiedad Intelectual (MIPI) realizada en conjunto entre esta Universidad y la Organización Mundial de la Propiedad Intelectual (OMPI).

El MI Lab aspira a funcionar como un repositorio de conocimiento científico sobre la gobernanza de la Inteligencia Artificial, que pueda servir de guía a reguladores y ***policy-makers*** tanto del sector privado como del público, impulsando esta investigación como parte de su iniciativa. Desde 2019, el MI Lab produce semanalmente un Newsletter (al que puede suscribirse desde aquí: bit.ly/SuscribiteMILab), en el que comunica sobre las noticias más relevantes en Inteligencia Artificial desde lo técnico, ético y regulatorio, con énfasis en tecnologías y aplicaciones incipientes de potencial riesgo, como fuera el caso del reconocimiento facial.

que reiterar que ***no todo problema puede o debe ser objeto de una solución automatizada.*** Existen problemas cuya solución es altamente contextual; que implican la ponderación de múltiples reglas y factores de hecho, la interpretación del lenguaje y sus matices y, por sobre todo, las aplicaciones de escalas de valores para la toma de decisión[3].

En todos los casos en que se evalúe aplicar sistemas automatizados a un problema o sector, debe existir un análisis previo sobre si éticamente deben o no deben implementarse sistemas de Inteligencia Artificial, si este uso particular habrá de implicar una afectación de los derechos humanos, y si las herramientas automatizadas son las adecuadas para resolver ese problema.

Sobre los potenciales peligros y las áreas en las que se puede aplicar la IA es conveniente pensar una **matriz de riesgos** que considere:

A. la **complejidad del contexto de la decisión** (si la decisión que se va a automatizar requiere ponderar vectores que impliquen interpretación semántica, contextual o axiológica), y la dificultad de trasladarlos a pesos matemáticos;

B. **el impacto de la decisión** (la importancia de la decisión que se toma en relación con el impacto que puede tener sobre una persona o grupo. Se debe construir el eje sobre un vector de magnitud que sopese la relevancia que puede tener la decisión automatizada sobre la persona impactada. No es lo mismo un sistema de recomendación de productos, que uno que define la performance de un trabajador, su acceso al crédito o a oportunidades de impacto significativo en la vida del afectado), y la

C. **necesidad de motivación y explicabilidad de la decisión a futuro** (factor en que se conjugan los requisitos de transparencia y explicabilidad de las decisiones automatizadas, en el sentido de decisiones ***in pectore***, versus decisiones motivadas y trazables).

Sintéticamente, las dimensiones de este marco teórico implican:

- la conciencia de los operadores o destinatarios de las decisiones de que se está interactuando con sistemas automatizados (***algorithmic awareness***);
- la imposibilidad de decidir sobre ser o no sujeto de aplicación de sistemas de IA o de escapar a su uso o a las consecuencias de su uso (pervasividad);
- el vector de impacto de amplificación que tienen los sistemas de IA, particularmente al ser exportados y replicados, como producto de una economía digital globalizada (escalabilidad);

3. Como enseña la regla de "un segundo" de Andrew Ng, que postula que "Si una persona normal puede realizar una tarea mental en menos de un segundo de tiempo de pensamiento, probablemente podamos automatizarla utilizando la IA, ya sea ahora o en un futuro próximo", en una especie de "fordismo digital", la automatización funciona mejor cuando la repetición es mucha y la decisión libre, poca.

Un ejemplo práctico de esto son las aplicaciones automatizadas en el ámbito de la justicia, donde existen tareas rutinarias y sencillas que impulsan el proceso, con poco espacio de decisión discrecional, alta trazabilidad sobre la posible casuística de soluciones y posibilidad de revertirlas fácilmente en caso de errores. En estas, la automatización presenta un campo fértil para liberar la labor humana para tareas más calificadas. Por otro lado, existen decisiones que requieren la ponderación de una multiplicidad de factores, desde reglas legales, lógicas, hasta valores axiológicos y criterios de equidad. En estas, en tanto el principal insumo del Derecho es el lenguaje y la interpretación semántica, se requiere el tamiz del entendimiento humano, que puede apreciar un contexto plagado de sutilezas, lo cual continúa siendo un área difícil para las técnicas actuales de IA.

- la eficiencia de un modelo como capacidad de reflejar la realidad que intenta representar en sus predicciones y de entender el contexto de su aplicación (***accuracy,*** confiabilidad, performance); considerando el tipo de decisiones y de datos con los que se trabaja, en tanto los valores vinculados a interpretación semántica, moderación de contenidos y evaluación de contexto presentan mayores dificultades para las implementaciones de IA (contextualidad), y las áreas en las que se pretende incorporar la solución (proporcionalidad).

- Ofuscación: distintas formas en las que la Inteligencia Artificial, sus procesos y/o sus resultados resultan oscuros para sus creadores, usuarios, operadores y/o destinatarios.

opacidad:	falta de transparencia en el acceso a comprender su funcionamiento sea por razones de complejidad técnica para quienes carecen de conocimientos en el área (opacidad técnica) o por dispositivos legales que lo transformen en una caja negra (opacidad legal);
inescrutabilidad:	incapacidad -aun para quienes la desarrollan o son expertos- de comprender cómo el sistema arriba a un resultado;
explicabilidad, interpretabilidad y trazabilidad:	forma en que los sistemas proveen justificación de sus acciones, y cómo sus "razonamientos" pueden ser recreados para su comprobación;

- el sesgo consciente o inconsciente en la selección de los datos, del modelo o en la interpretación de los resultados (sesgo);

- ***accountability:*** formas en las que los algoritmos pueden ser auditados (auditabilidad), y cómo el derecho habrá de trasladar las consecuencias de los daños ocasionados por IA (responsabilidad);

- el impacto de los sistemas de IA en la sociedad, considerando si buscan reflejar matemáticamente un modelo de la realidad en forma justa ***(fairness)*** o modificando proactivamente las inequidades existentes en la realidad a fin de que no se trasladen a lo digital (inclusión). Aquí es donde las evaluaciones de impacto algorítmico presentan potencial, no solo previo a su implementación con una proyección de potenciales ramificaciones, sino con la imposición de un seguimiento continuo una vez que se implementan.

Cabe aclarar que la segmentación propuesta obedece a fines didácticos, para tratar de aprehender la complejidad de las distintas formas en que la Inteligencia Artificial impacta nuestra sociedad. Ciertamente estas dimensiones no son compartimentos estancos, sino que muchos de sus matices se trasladan entre una y otra categoría.

AWARENESS: INVISIBILIDAD, APARIENCIA DE NEUTRALIDAD. CONCIENCIA DE INTERACCIÓN

Sea por una **invisibilización por diseño**[4], o por la **velocidad de las transacciones**, lo cierto es que en nuestras interacciones con las tecnologías de la información, pocas veces reflexionamos mientras las usamos acerca de los mecanismos que las hacen funcionar.

Colectivamente, operamos bajo varios postulados y asunciones implícitas: que la **tecnología es neutral,** que las **computadoras son infalibles** y que los sistemas son programados por alguien que sabe lo que hace y que tiene **control sobre todo el proceso**. Existe un pacto no explícito de confianza en la contraparte que desarrolló el sistema, en el que descansamos cuando utilizamos tecnologías de la información[5].

Esperamos que la tecnología funcione y damos por sentado que así es, sin ponernos a pensar sobre la multiplicidad de procesos, técnicas, contratos, regulaciones, términos y condiciones que hacen, por ejemplo, que lo que escribimos digitalmente en un procesador de texto *cloud* pueda materializarse en una página impresa. El salto entre lo digital y lo analógico queda invisibilizado por la naturalización del uso, y el portento de la tecnología, aniquilado por la cotidianeidad[6].

Generalmente solo cuando algo no funciona es que, metafóricamente, levantamos la tapa y examinamos el motor para ver qué falló. Damos por sentada esta

4. Parte de la magia de los sistemas de la información es presentar interfaces que por diseño buscan presentar interacciones que oscurecen las decisiones que los sistemas toman hasta presentar una respuesta, lo que es coadyuvado por la velocidad en que los sistemas operan, obteniendo respuestas en fracciones de segundo.

Los equipos de UI y UX (***user experience***) buscan comprender a sus usuarios y generar interacciones eficientes y satisfactorias. Entre usuarios y diseñadores / desarrolladores existe una asimetría de información: los usuarios solo saben de la plataforma o servicio lo que la interfaz muestra. Por ejemplo, muchos asumían la anonimidad de mirar un perfil de LinkedIn, hasta que el diseño de la plataforma ofreció como un servicio premium dar detalles de quién consultaba los perfiles, o permitir la navegación "privada", mediante un servicio diferencial.

Cuando el mandato para el diseño es priorizar una experiencia ***seamless***, los elementos que brindan información sobre cómo el sistema está funcionando o qué datos recopila para ellos interrumpen la experiencia del usuario, rompiendo la cuarta pared. Por ello, de forma intencional se terminan cubriendo decisiones de funcionamiento bajo maquillaje de diseño, lo que hace que los usuarios no puedan conocer ni comprender los verdaderos alcances de la interacción que están realizando y cómo esa tecnología los impacta. Esto ha llevado a diseños que explotan esas vulnerabilidades, lo que se conoce como "***dark patterns***".

5. Como los usuarios de dispositivos de ***smart home*** o asistentes, donde existe una expectativa implícita de privacidad de que los sistemas no están escuchando permanentemente, o no graban si no se activa el comando para ello (por ejemplo: ***https://www.theguardian.com/technology/2018/may/24/amazon-alexa-recorded-conversation***).

6. Servicios como BuiltWith Technology Profiler: ***https://chrome.google.com/webstore/detail/builtwith-technology-prof/dapjbgnjinbpoindlpdmhochffioedbn?hl=en*** o Wappalyzer: ***https://chrome.google.com/webstore/detail/wappalyzer/gppongmhjkpfnbhagpmjfkannfbllamg?hl=en*** nos brindan más información sobre qué tecnologías se usan para construir cada aplicación o sitio, pero aún existe una brecha de asimetría informativa según la cual desconocemos todo lo que verdaderamente se podría hacer con un sitio (como por ejemplo, conocer la IP de quien lo visita).

"magia tecnológica" y aceptamos los resultados, sin mayor cuestionamiento. Esta forma de operar **invisibiliza** los procesos tecnológicos, en tanto nos hace enfocarnos en los resultados más que en los métodos.

En el caso de los algoritmos empleados, son una tecnología diseñada para pasar desapercibida para el usuario. Estamos acostumbrados a ingresar un término en el buscador y ver los resultados, pocas veces nos detenemos a pensar por qué un resultado se muestra en primer lugar o por encima de otro. Irónicamente, los motores de búsqueda nos devuelven la impresionante cantidad de resultados (por ejemplo 43.100.000 resultados en 0,80 segundos en el término Metaverso, dada su reciente explosión en relevancia), con la ilusión de información infinita, pero raramente los usuarios pasan de la primera página.

Nuestra asunción innata sobre la neutralidad de los sistemas hace que no nos cuestionemos cuáles son las decisiones que se toman para componer esa jerarquía, y quiénes las toman, lo que es terreno fértil para la **"economía de la atención"** y los negocios de SEO (***search engine optimization***)[7]. La publicidad mueve los negocios en Internet, de ahí uno de los postulados más conocidos del decálogo de la web: "si no estás pagando por algo es porque eres el producto".

Antes de que la manipulación de la información tomara estado público por trolls, bots y ***fake news***, la mayoría de los usuarios de redes sociales presuponía que la información era desplegada en su ***newsfeed*** de manera cronológica a medida que otros usuarios realizaban actualizaciones, y no que su ***timeline*** era producto de una selección operada por algoritmos (***algorithmic curated feeds***).

Esta falta de conocimiento de la interacción es para algunos indicativa de un diseño exitoso (lo que se suele denominar en UX, UI, "***seamless experience***"). Para otros esta invisibilidad indica que algo potencialmente controversial ha sido establecido, decidido y hecho estático (Hamilton et al. 2014).
Asimismo, la rapidez en el procesamiento de la información ayuda a ocultar: la información buscada se presenta y es consumida antes de tener tiempo de analizar los métodos de selección y el orden en que se la presenta. Esto se agudiza más aún con los asistentes personales, en los que se humaniza la interacción de recuperación de información a través de una interfaz generalmente conversacional, naturalizando las respuestas.

Por ello, el primer paso en el esquema de análisis de impacto de la Inteligencia Artificial es **darse cuenta de que los algoritmos operan a nuestro alrededor** y empezar a reconocer sus interacciones (***algorithmic awareness***). Justamente, por esto es que una de las propuestas que formulara durante el Simposio de Inteligencia Artificial e Inclusión celebrado en Brasil en 2017 apuntaba a implementar avisos sobre la interacción con algoritmos, de manera similar a los ***pop-ups*** que advierten sobre el empleo de cookies[8].

Se suele cuestionar que estos avisos "rompen" la naturalidad de la interacción o nos hacen comportarnos diferente, en el sentido en que el tono y las expresiones verbales utilizadas para comunicarnos con las máquinas son -o deberían ser- muy diferentes de las inflexiones que usamos para con otros humanos. De hecho, cuando hablamos con las máquinas lo hacemos a través de **"comandos"**, órdenes imperativas y precisas destinadas a desencadenar un resultado específico. Sin embargo, "comandar" tiene en el lenguaje un sentido muy específico

7. El posicionamiento de los resultados y la venta de ***keywords*** es un negocio multimillonario, con estudios que estiman que el 25% de las personas clickea en el primer resultado orgánico y que los porcentajes descienden drásticamente en la lista, al punto que raramente los usuarios consultan la segunda página de resultados. ***https://www.searchenginejournal.com/google-first-page-clicks/374516/#close***

8. "Micaela Mantegna on Algorithmic Awareness", The Berkman Klein Center for Internet & Society, recorded en Noviembre de 2017, publicado en enero de 2018. Disponible en: ***https://www.youtube.com/watch?v=gntn4o2kET0***

de autoridad y jerarquía, impropio para dirigirse a otras personas en contextos generales.

Por esto, como resultado de investigaciones en **HCC (*Human Centered Computing*) y HCI (*human-computer interaction*)**, los asistentes Siri y Alexa fueron modificados para no responder si los niños no utilizaban "por favor" o "gracias"[9] y continúan los estudios para determinar si criarse con esta comunicación dual con agentes conversacionales impacta y genera niños más exigentes y demandantes[10].

En otros segmentos etarios que no son nativos digitales, la brecha se manifiesta en atribuir un carácter cuasi-mágico a la tecnología, siendo particularmente vulnerables a engaños. Igualmente, la sofisticación creciente de estos sistemas hace que aun los usuarios más informados tengan dificultades para detectar si su interlocutor es artificial.

Este déficit de asimetría informacional requiere urgente atención, sobre todo con la creciente digitalización de servicios como resultado de la pandemia y la implementación de chatbots para la atención al público, situación en la que, pasando confortablemente el test de Turing, muchos usuarios no advierten que se encuentran interactuando con un software de IA[11], máxime cuando esto se intenta camuflar aún más otorgando a los asistentes nombres genéricos humanos.

Mientras que las leyes de protección de datos personales quizás no puedan brindar respuesta total a la necesidad de "avisos de interacción", sí es campo fértil para las leyes de protección de los consumidores y los principios que imponen de información sencilla, asequible y suficiente. En función de esto, los consumidores deberían ser informados de que están interactuando con un sistema automatizado, para tomar decisiones informadas sobre los datos que son recopilados por estos o el tipo de inferencias que pueden realizarse.

Esto conduce a profundizar sobre la segunda asunción que mencionamos en cuanto a nuestros usos de la tecnología: que se trata de una **herramienta neutral** que solo cobra propósito en las manos de quien la usa. En los entornos analógicos, la mayoría de las tecnologías que los humanos utilizábamos eran vistas como una herramienta inerte, diseñada para cumplir con su función y sin ulteriores intenciones de las que el usuario tuviera que estar alerta.

La neutralidad tecnológica es un credo asumido colectivamente, que se ha traspasado ***verbatim*** al mundo digital sin cuestionar sus particularidades. Si bien puede argumentarse que los algoritmos en tanto fórmulas matemáticas ***per se*** son tecnológicamente neutrales, lo cierto es que están impregnadas de los sesgos y de las intenciones de quienes las diseñan. Por otra parte, en el ámbito social el engaño está en la inversión sobre el concepto de quien es el que efectivamente los "usa", o mejor dicho, a quienes responden como herramienta[12].

Un objeto inanimado analógico es algo inerte. Al usarlo, no sospechamos que ese objeto tenga una voluntad o una ultra-intención propia o que responda a terceros, pero cuando los objetos tienen software embebido, una gran parte de las reglas que los gobiernan quedan fuera de nuestro alcance: sea por cuestiones técnicas

9. Ver sobre esto: Hiniker et al. (2021) y Beneteau et al. (2020).

10. *https://www.geekwire.com/2021/alexa-please-study-suggests-kids-know-talk-differently-virtual-assistants-humans/*

11. Como se puede advertir en el experimento llevado adelante por Mozilla con la aplicación Bot or Not, disponible en: ***https://botor.no/***

12. Parafraseando el ***Quis custodiet ipsos custodes?*** podemos decir: ¿quién usa los algoritmos y quiénes son usados?, para resumir la pregunta sobre quiénes son los verdaderos dueños o los maestros a quienes estas tecnologías sirven.

que no comprendemos, por reglas de propiedad intelectual y licencias, o bien por términos y condiciones que determinan cómo podemos o no usarlos.

Tomemos como ejemplo algo tan inocuo como un "ascensor inteligente'', que manifiestamente está programado para una mayor "eficiencia". La trampa semántica consiste en entender que la definición sobre qué es "eficiente" habrá de variar según los intereses e intenciones de quien defina el concepto. Desde una perspectiva ecológica, el ascensor será más eficiente si tiene un menor consumo de energía aun a costa de que las personas deban esperar más tiempo. Para el usuario probablemente será más eficiente el ascensor que lo haga esperar menos tiempo. Este también puede ser el interés del empleador, que busca evitar tiempo ocioso de sus empleados en los ascensores y que lleguen a sus puestos de trabajo más rápidamente. En ese caso, cuando los intereses de varios sectores se alinean, se crea una falsa sensación de dominio sobre la cosa, cuando en realidad esta responde a los deseos de otro.

Por esto, cuando utilizamos la infinidad de servicios gratuitos de mensajería, localización, agendas, nuestros intereses están parcialmente alineados con quienes los proveen, pero probablemente no somos conscientes de la totalidad de las transacciones involucradas, de la contraprestación en datos, información e inferencias que se están extrayendo a nuestra costa como "pago" por el servicio.

Como se viera con los ejemplos relativos a la optimización de motores de búsqueda y curación de los ***newsfeeds***, algo similar ocurre en materia de algoritmos, ámbito en el cual es una creencia común entre los usuarios pensar que las búsquedas online siguen patrones neutrales en la recuperación de la información requerida.

Esta confianza sobre la **apariencia de neutralidad** se traspola a cómo nos relacionamos con las tecnologías de manipulación de la información: nos lleva a creer que los resultados de búsqueda que vemos son el producto de una ecuación matemática que nos acerca, de manera inteligente y eficaz, aquello que estamos buscando, sin percibir que en realidad pueden estar sirviendo para otros fines.

Un algoritmo determina cuál es la mejor ruta para llegar de un punto A a una locación B y nos la presenta como uno o varios resultados posibles, pero ¿sabemos acaso cuales son los parámetros que pondera para eso? Justamente como pasa con los trucos de magia, donde la atención se distrae hacia el resultado final, damos por sentado que la respuesta que se nos brinda es la "mejor", sin detenernos a pensar qué criterios, valores o intereses se asumen para elegir la mejor ruta.

Según cual sea el valor preeminente para cada uno de nosotros, la respuesta sobre "la mejor ruta" habrá de variar. ¿Es el camino elegido el más rápido o el menos transitado? ¿Valora el estado de mantenimiento de las calles o si es una ruta más pintoresca? Cada uno de estos resultados conlleva ínsito un juicio axiológico, que está destilado en el código mismo de la fórmula matemática, determinado por la propia visión del mundo, prejuicios e intereses de quien la construye y la programa para funcionar. A veces esas desviaciones son inconscientes (***unconscious bias***), otras, netamente intencionales (por razones de negocios).

Estas decisiones tras bambalinas que se hacen en el proceso de construcción de las tecnologías de Inteligencia Artificial tienen peso económico y en nuestra autodeterminación. Supongamos a fines de ejemplo. Puede ser interesante que la aplicación nos recomiende como "mejor ruta" el camino que pasa cerca de un centro comercial o una cafetería que es de nuestro agrado, ya que puede estar -secretamente para nosotros-, siendo sponsoreado directamente en la aplicación, o incluso hasta de forma oblicua publicitando en ella.

Los criterios de selección para la sugerencia de una ruta por encima de otra nos son desconocidos, pero aun así

determinan en forma mediata nuestras acciones y nosotros confiamos en esos resultados.

Nuestra confianza en la neutralidad y eficacia de los productos informáticos hace que no cuestionemos el teatro de sombras que está operando por detrás. En segundo plano de estas toneladas de información que recopilan los servicios, hay maquinarias de matemática sutil operando entretelones, clasificando, ordenando y devolviendo datos, según un patrón predeterminado y según los criterios fijados en los algoritmos que los potencian.

La trampa es que, justamente, los criterios de selección de estos elementos y el valor de su ponderación no están explícitos. Como un mago que no revela sus secretos, los algoritmos nos deslumbran con sus respuestas, sin jamás revelar el secreto de sus procesos, o a veces siquiera, sin dejar entrever el hecho de que están operando, de ahí que la **conciencia de interacción (*algorithmic awareness*)** sea nuestro primer paso de análisis en el camino de su impacto.

PERVASIVIDAD

Cuando empezamos a tomar conciencia de nuestras interacciones con aplicaciones automatizadas, empezamos a percibir cómo la Inteligencia Artificial opera todo el tiempo a nuestro alrededor, como motor de los servicios de la sociedad de la información.

Las aplicaciones de búsqueda, los filtros, los motores de recomendación de contenido, los algoritmos de reconocimiento facial, el control del tránsito y la identificación de contenidos en línea son ejemplos de cómo la IA está, omnipresente y silenciosa, en muchas aplicaciones y actividades de nuestra vida cotidiana.

Allí donde se recopilen grandes datos, los algoritmos estarán detrás para extraer la información económicamente valiosa, aceitando los engranajes de la transformación digital.

Cuando hablamos de pervasividad, hablamos del grado de penetración de esta tecnología en todas las capas de actividades, servicios, industrias y gobierno, en lo público y en lo privado, en lo individual y en lo colectivo.

Considerando al ser humano como centro y, en ondas expansivas, desde el individuo a la sociedad, vemos distintas esferas de recopilación de datos e interacción con sistemas automatizados, de los cuales no siempre es posible sustraerse. Para esto, podemos trazar un arco que va desde los dispositivos insertos en nuestros cuerpos (prótesis y tendencias cyborg de aumentación para ***smart humans***), hasta aquellos utilizados sobre nuestros propios cuerpos (***wearables***), pasando por aquellos que nos rodean en nuestros hogares (domótica) y, a una escala global, los que se encuentran en las ciudades (***smart cities***).

Si bien las tecnologías de recopilación son similares, el tratamiento, la naturaleza y el sujeto que posee el control de los datos dentro de cada uno de estos procesos habrá de ser diferente.

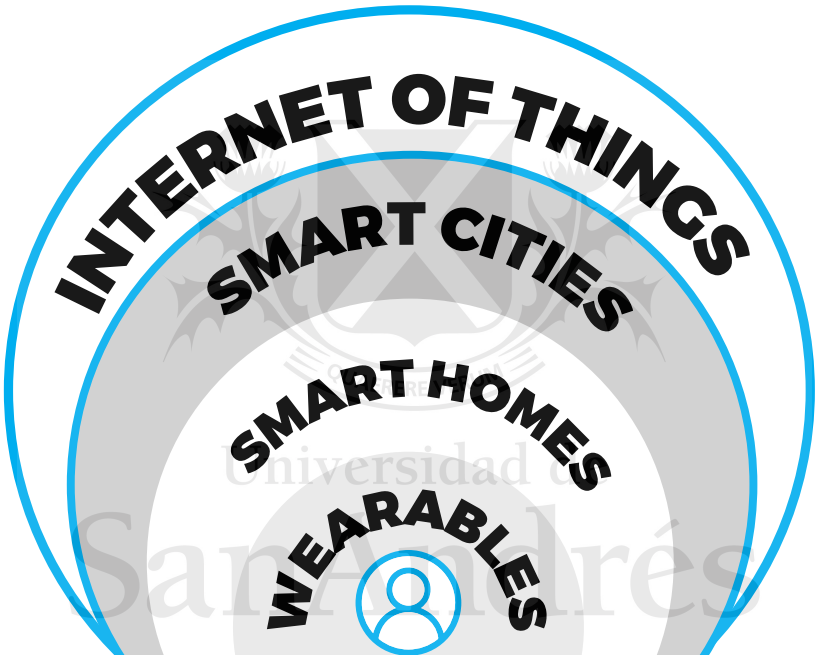
1. Dispositivos corporales (prótesis, ***wearables***) = comprenden datos sensibles y personales, algunos completamente involuntarios, sometidos probablemente a una regulación más estricta que otros tipos de datos

2. Dispositivos domésticos (personales, domótica, ***Internet of Things***) = comprenden datos bajo nuestra esfera de decisión en cuanto a su implementación, pero no completamente en cuanto a su uso y procesamiento. Primariamente promocionados como para nuestro provecho, reportan sin embargo mayormente a entidades privadas y a corporaciones.

3. Dispositivos en espacios públicos = son de naturaleza pública, no se encuentra en nuestra esfera de decisión el registro de esos datos. Sin pedido de acceso a la información, no está generalmente disponible para nuestro provecho ni consulta.

La posibilidad de hacer ***opt out*** es factible cuando se trata de servicios privados no imprescindibles. En cambio, no existe una verdadera libertad de opción cuando estos se transforman en formas de comunicación cotidiana, herramientas laborales o educativas (por ejemplo, aplicaciones como WhatsApp, Zoom), plataformas de discusión públicas (redes sociales como Twitter o Facebook) o más aún en servicios digitales gubernamentales, que se construyeron apresuradamente por la necesidad surgida de la no presencialidad.

Aún más, nada garantiza que aun a pesar de no ser usuario del servicio ni haber por ende aceptado los términos de servicio, los datos no sean igualmente recopilados (como fuera el caso de los ***shadow profiles*** de Facebook[13]), ni confiar en que los términos y condiciones sobre la información que se sube no sean abusados por las propias plataformas[14] o infringidos por terceros (como el caso de Clearview AI[15], que construyó su masiva base de datos de reconocimiento facial a través de ***scrapping*** de redes sociales).



La recopilación de información desde lo individual a lo colectivo, partiendo de la persona como eje. (Mantegna, 2021)

La pervasividad de los sistemas algorítmicos se ve agravada por la sumatoria de todos los datos de nuestros dispositivos, tendencia que crece con el despliegue de Internet de las Cosas (***Internet of Things,*** IOT) y las ciudades inteligentes (***smart cities***).

Estos ámbitos no funcionan como compartimentos estancos, sino que existe una integración tanto vertical

13 El término "***shadow profiles"*** se refiere a perfiles de no usuarios que la red social Facebook mantenía sobre personas que no eran usuarias de ella. *https://www.theverge.com/2018/4/11/17225482/facebook-shadow-profiles-zuckerberg-congress-data-privacy*

14 Como las repetidas historias sobre los algoritmos detrás de las ***person you may know*** (PYMK) que mostraban conexiones inverosímiles e íntimas de los usuarios de la red Facebook. *https://gizmodo.com/how-facebook-figures-out-everyone-youve-ever-met-1819822691*

15 Que motivara el envio de intimaciones por parte de Twitter y otras plataformas por infringir los términos de uso. *https://www.nytimes.com/2020/01/22/technology/clearview-ai-twitter-letter.html*

como horizontal de los datos recopilados en cada esfera, a los fines de construir perfiles personales más precisos que vuelven al usuario en forma de resultados contextuales (***web 4.0*** o ***"web of me")*** y publicidad altamente personalizada.

Algunos ven la respuesta en el empleo de las leyes de protección de datos personales, que al tratar el fenómeno de las decisiones automatizadas, apuntan a facultar al titular del dato a removerlo bajo determinadas condiciones, o a hacerle saber del posible resultado perjudicial, o bien a requerir la intervención humana.

El ***opting out*** puede ser difícil de concretar en la práctica considerando lo costoso que puede resultar la protección de la privacidad (que se paga como característica ***premium*** en algunas plataformas) y lo engorroso que puede resultar para un individuo embarcarse en este tipo de requisitorias de remoción de datos, teniendo en cuenta además que este individuo se encuentra en una asimetría de información en cuanto a los datos o a las inferencias que se tienen sobre él.

Más aún, considerando cómo funcionan los modelos algorítmicos, esta facultad de remoción puede ser más perjudicial que beneficiosa. En algunos contextos, la falta de datos sobre una persona puede ser problemática. Por un lado, en un mundo en el que los datos fluyen, la falta de información puede verse como un sector "oscuro" sobre una persona, en la lógica de la falacia sobre la privacidad y por el valor otorgado al hecho de "no tener nada que ocultar". En la práctica, funciona a la inversa de lo que ocurre con el valor del silencio. Legalmente, cuando no existe una obligación de declarar, el silencio no puede ser usado en contra del imputado. Sin embargo, cuando hablamos de datos, la falta de información puede resultar algo sospechoso, sea por: a) construir una presunción negativa (viendo este vacío como la ausencia de una información favorable) (Vertesi, 2014), o bien por: b) restar datos necesarios que impacten en la valoración final de la fórmula (arribando por ejemplo a una conclusión denegatoria sobre un préstamo, por no alcanzar el puntaje necesario).

Por otro lado, a escala colectiva, la detracción de información de un grupo puede llevar a que esta clase se encuentre mal representada en el resultado final. Por ejemplo, obviar considerar una variable que marque similitudes entre puntos de datos puede conducir al ***underfitting*** del modelo, haciendo que allí donde se hubiera construido un grupo con características homogéneas, los individuos se distribuyan entre otras categorías con las que tienen una menor afinidad.

Sintetizando, en este panorama, el ***opting out*** del sistema no es una opción eficiente o viable por las dificultades prácticas de realizarlo, por los costos de este tipo de aislamiento o privacidad y porque la carencia misma de datos puede crear una presunción negativa en contra de las personas.

Sin dudas, en un mundo de datos, la ausencia de información es un dato en sí mismo.

ESCALABILIDAD (*SCALABILITY*)

La pervasividad como dimensión invita a poner atención en el grado de penetración que las tecnologías de automatización tienen en nuestra sociedad, así como en los efectos duraderos de estas implementaciones.

Esta faceta de análisis está estrechamente conectada con la escalabilidad, es decir, con la dimensión que alude a la escala, tanto en el sentido de crecimiento exponencial de las implementaciones de tecnologías basadas en la IA, como particularmente a la potencial magnitud de impacto de los sistemas automatizados y la velocidad a la que se distribuyen. Dentro de las organizaciones, existe un incentivo muy fuerte para adoptar soluciones basadas en Inteligencia Artificial, tanto para descongestionar sistemas desbordados[16], como para reducir costos, para que asistan en la toma de decisiones, así como también por al aura de infalibilidad y eficiencia asociada a la tecnología. Particularmente, en estructuras organizacionales adversas al riesgo, existe un sesgo de confirmación hacia los resultados que son producto de la tecnología. Si algo falla y las cabezas van a rodar, es más defendible la posición de quien tomó la decisión respaldado en la tecnología, que la de aquel que se apartó de la sugerencia de la máquina. Ergo, existe un incentivo muy fuerte a no cuestionar los resultados de procesos basados en Inteligencia Artificial.

Por otro lado, frente a la fuerte pulsión para la adopción de soluciones automatizadas, organizaciones que no tienen disponible investigación y desarrollo propios, deben recurrir a la tercerización. No solo a través de consultorías personalizadas para la transformación digital, sino también por una cuestión de costos, muchos habrán de recurrir a soluciones puntuales estandarizadas. De manera similar a lo que ocurre con el software, desarrollar una solución a medida puede ser caro e ineficiente, dado que es necesario tener conocimientos específicos del campo, contar con los datos y con la infraestructura necesaria para desarrollar el modelo.

La comoditización de la IA a través de modelos de negocio basados en la nube y en el aprendizaje automático como servicio (***Machine Learning As a Service, MLaSS***) permiten que la automatización penetre en cualquier industria, y que prolifere. Como de momento las evaluaciones de impacto de los algoritmos no son obligatorias, el riesgo es que las potenciales fallas que tengan los sistemas de IA tercerizados habrán de escalar y expandirse geográficamente a través de las organizaciones que las utilizan, sin una auditoría sostenida de su performance.

Los sistemas de IA se comercializan a través de software, en forma de servicios, programas específicos o plataformas. Dada la velocidad de la distribución digital, un sistema de IA podría comercializarse en todo el mundo con la misma facilidad que un procesador de texto. Sin embargo, a diferencia de otros tipos de programas, los que contienen sistemas de

16. En lo que llamo el "sesgo de confianza en sistemas informáticos". Esto es particularmente aplicable en lo que hace a políticas públicas y sistemas sobrecargados de trabajo como en la justicia (*https://cetys.lat/evaluacion-de-la-preparacion-del-sector-judicial-para-la-inteligencia-artificial-en-america-latina/*), donde la IA se ha presentado como una panacea y una tentación tecno-solucionista *https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/ https://stanfordpolitics.org/2020/09/13/ai-prediction-tools-claim-to-alleviate-an-overcrowded-american-justice-system-but-should-they-be-used/*

decisión o automatización basados en IA presentan riesgos diferenciales. Además de sus diferencias lingüísticas, un software de escritura funcionará de igual forma frente a distintos usuarios alrededor del globo; lo que no ocurre con sistemas de reconocimiento facial.

La precisión de un modelo algorítmico depende en gran medida de la calidad y la cantidad de datos de entrenamiento y de prueba. Si un modelo biométrico o de reconocimiento de imágenes se ha entrenado solo con personas caucásicas, no funcionará correctamente cuando se venda en países con otra composición étnica. Con el peligro añadido de lo ya explicado en relación con el sesgo de confianza en sistemas informáticos: un sistema basado en IA puede estar fallando y como actúa en decisiones hipotéticas y a futuro, será difícil de advertirlo. En el caso de los algoritmos y aplicaciones de IA, se corre el riesgo de que los datos con que haya sido entrenada la aplicación no se correspondan con los de la nueva población, y que ello desvíe las métricas resultantes, mermando la eficacia predictiva al aplicarse a un entorno para el que no fue entrenado apropiadamente.

Imaginemos un software de reincidencia entrenado en Estados Unidos, que debe considerar en la ecuación las particularidades en la composición racial del país. Exportada dicha fórmula a un estado con población más homogénea, los resultados pueden estar sesgados de una manera inadvertible para el operador no entrenado. En este sentido, Joy Buolamwini, activista e investigadora del MIT detrás de la iniciativa ***The Algorithmic Justice League***, descubrió que un programa de reconocimiento facial basado en IA no reconocía su cara a menos que usara una máscara blanca, situación que se repetía con cada software que usaba la misma biblioteca base como recurso, a lo cual bautizó apropiadamente como "***The Coded Gaze***" ("AJL -ALGORITHMIC JUSTICE LEAGUE" n.d.), y en la que se basó el documental ***Coded Bias***.

Actualmente y de forma insospechada para la mayoría de la población, soluciones informáticas basadas en Inteligencia Artificial son distribuidas globalmente por unos pocos proveedores, que proveen sus servicios a fuerzas de seguridad[17] o aeropuertos, magnificando, replicando y perpetuando alrededor del mundo cualquier sesgo o defecto que tuvieran, sin mayores controles, afectando la vida y los intereses de las personas bajo su ámbito de aplicación.

A medida que las aplicaciones de ***machine learning as a service*** se expandan, será necesario contar con mecanismos o evaluaciones que permitan auditar el impacto de la aplicación a una población nueva, así como el grado de ajuste que la fórmula original tiene sobre ella.

17. Como el caso de la expansión de la polémica ***Clearview AI:*** ***https://www.nytimes.com/2021/03/18/technology/clearview-facial-recognition-ai.html***

CONFIABILIDAD (*ACCURACY*)

Si la pervasividad nos invita a pensar sobre el grado de penetración de los sistemas automatizados en la sociedad, el siguiente paso en esta escala de dimensiones es reflexionar sobre la confianza que depositamos en la infalibilidad de los sistemas informáticos.

Bajo esta premisa, la confiabilidad representa distintos matices vinculados a la eficiencia de un modelo como su capacidad de reflejar la realidad que intenta representar en sus predicciones y de entender el contexto de su aplicación.

No todos los problemas pueden o deben ser resueltos a través de Inteligencia Artificial. Aun antes de preguntarnos cuán eficiente es un sistema de IA, debemos considerar previamente si debe o no debe ser utilizado en esa área.

Existen decisiones que es debatible que deban confiarse a sistemas automatizados. No solo por la complejidad de los factores involucrados en la toma de decisión, sino por las implicaciones éticas del hecho de que una máquina decida sobre áreas con impacto irreparable sobre la vida (literal y figuradamente) de los seres humanos. Gran parte de los motivos para la prohibición de los sistemas automatizados de armas letales (LAWs, o AWS, ***automatic weapons systems***) apelan a este argumento. La campaña ***Stop Killer Robots***[18] desde hace años aboga por la prohibición internacional del uso de armas letales automáticas, bajo esta premisa.

Por otro lado, aun cuando se decida su aplicación a un área, hay que remarcar que los sistemas de Inteligencia Artificial no son infalibles. Las respuestas que los sistemas de IA brindan no son en términos absolutos de certeza y error, sino que en muchos casos brindan porcentajes de probabilidades.

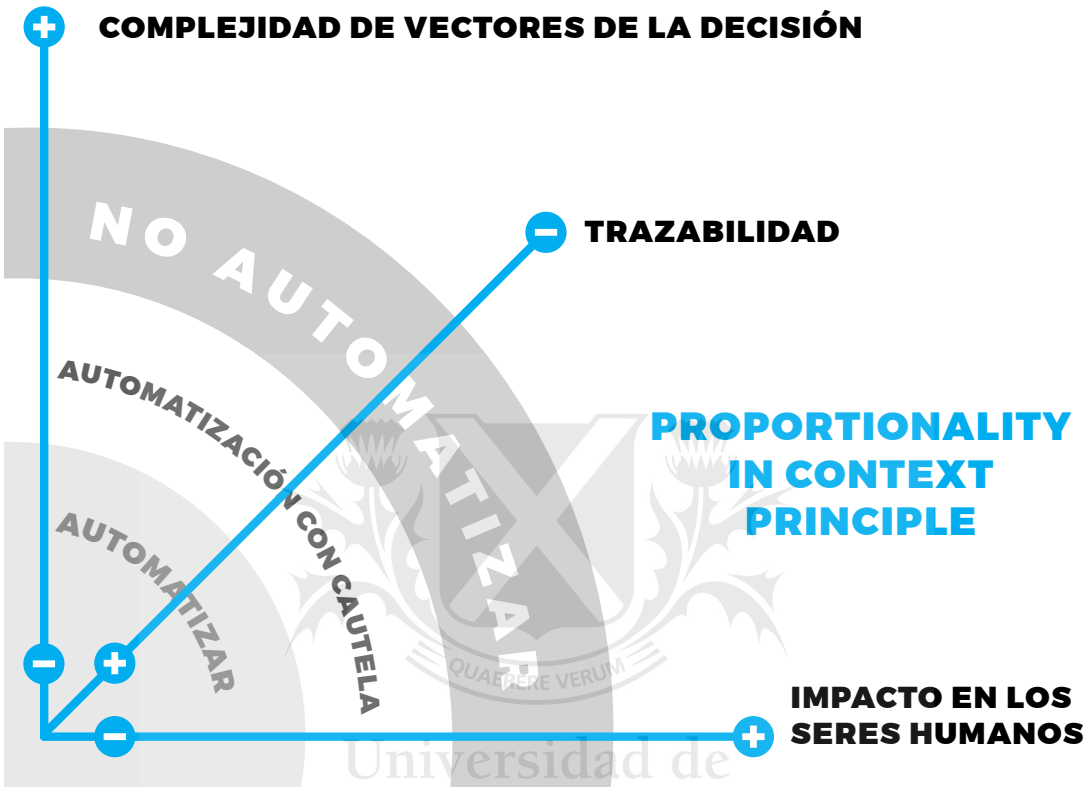
Esto es un punto crucial, porque la estructura cognitiva y el sesgo de confianza que los seres humanos tienen sobre las tecnologías informáticas traducen esos resultados a absolutos. Por ejemplo, una probabilidad de un 80% de que esa cara pertenezca a esa persona se traduce en una identificación positiva.

Este es el caso de ***Rekognition***, el software de reconocimiento facial (RF) de Amazon que presentaba un margen de error del 20% en su funcionamiento. Activistas de la ACLU (***American Civil Liberties Union***) demostraron que bastaba ese porcentaje para que el sistema falle al reconocer a 28 miembros del Congreso, que fueron incorrectamente identificados. Frente a una ola creciente de críticas, muchas compañías tecnológicas impusieron una moratoria en sus ofrecimientos comerciales, hasta que la tecnología fuera más robusta. Sin embargo, sistemas de RF similares están siendo ofrecidos por competidores menos escrupulosos y empleados hoy por fuerzas de seguridad.

Aún más, a la hora de definir su aplicación, se debe considerar el tipo de decisiones y los datos con los que se trabaja, en tanto los valores vinculados a interpretación semántica y evaluación de contexto presentan mayores dificultades para las implementaciones de IA (contextualidad), así como para las áreas en las que se pretende incorporar la solución (proporcionalidad). Estas dos aristas de la dimensión

18. https://www.stopkillerrobots.org/

de confiabilidad conforman lo que se denominó el "principio de proporcionalidad en contexto" ***(proportionality in context***), implicando que la adopción de un sistema de IA para la resolución de un problema debe ser proporcional a la dificultad de los factores que deben merituarse contextualmente, al tipo de problema en función de cómo impacta en los seres humanos, y a la trazabilidad de la decisión.



Visualización de la relación del principio de proporcionalidad en contexto como matriz rectora para determinar qué áreas o no automatizar, considerando el balance entre el impacto que la decisión puede tener en los seres humanos, su complejidad, y cuan trazable o no es el camino por el que se llega a la respuesta, que permita reconstruir el razonamiento adoptado.

Las aplicaciones de Inteligencia Artificial no entienden de contexto, por lo tanto mientras mayor sea la complejidad semántica y la asignación de significados que los algoritmos tengan que "interpretar", mayores son las posibilidades de cometer errores.

Esto se aprecia muy claramente en la forma en que los algoritmos de las redes sociales adjudican la moderación de contenido. El entendimiento de los algoritmos sobre el mundo es casi literal, y estos no entienden de parodia, ironía o contextos en que una información se usa.

Esto es particularmente peligroso para el ejercicio legítimo de derechos, si pensamos en que la mayoría de las excepciones y limitaciones a los derechos autorales se basan en situaciones que requieren la interpretación del contexto. Si una obra es usada en forma educativa, informática, satírica, etc., dependerá de un análisis factual particular a las circunstancias del caso. Cuando desde la legislación se imponen penalidades a las plataformas por los contenidos que suben los usuarios, se crea un incentivo económico perverso. La moderación por revisores humanos es costosa comparada con la automatizada, por lo cual, enfrentados a la posibilidad de sanciones por una ley que impone responsabilidad por la moderación, las plataformas habrán de recurrir a la menos

onerosa, y los usuarios y su libre expresión se transforman en daños colaterales. En esto, en el triple eje de la ecuación entre conformar las legislaciones que amparan a los titulares autorales, mantener altos los réditos de las plataformas o proteger a los usuarios, el mal menor a los ojos de la empresa es sin duda este último. En la práctica, esto termina importando una inversión del principio legal, en el ejercicio de los derechos derivados de las limitaciones y excepciones al derecho de autor, adjudicando la carga a la persona que realiza la expresión.

La interpretación no es solo de las palabras con las que trata, sino del contexto en el que estas se insertan y de los valores que están en juego. Es un triple eje hermenéutico: semántico, contextual y axiológico. Dependiendo de la sofisticación del sistema, los resultados pueden ser mejores o peores, pero siempre está presente este talón de Aquiles de los sistemas artificiales de inteligencia.

Otro problema es que la respuesta que los algoritmos brindan es sobre un futuro hipotético por lo que en aquellos casos en que la respuesta resulte plausible, cualquier error puede permanecer mucho tiempo escondido mientras no se vuelva palmario en una respuesta absurda.

Cuando se trate de decisiones que impliquen una posibilidad en un espectro, como por ejemplo interpretar factores para decidir la concesión de un crédito o de una eximición de prisión, el error puede no ser observable. En cambio, resulta evidente en casos en que queda demostrada palmariamente la falta de sentido común y contexto, por ejemplo en las traducciones automáticas. En Argentina, el ente de turismo de la Provincia de Córdoba fue vilipendiado porque en el sitio web oficial, localidades de la provincia aparecían traducidas en forma absurda: "***Sierras girls***" por Sierras Chicas, "***Get out if you can***" para la ciudad de Salsipuedes y La Falda como "***Skirt***"[19].

Como se explicó al tratar la escalabilidad, la forma en que los algoritmos se despliegan comercialmente funciona como una caja de resonancia que amplifica el alcance de sus errores.

Los errores, que en pequeña escala no causan perjuicio significativo, pueden verse magnificados por su empleo masivo, particularmente cuando son "exportados" a áreas no nativas, donde su impacto puede ser altamente pernicioso.

Este alcance es lo que según Cathy O'Neil (2016) otorga un impacto potencialmente nocivo a los algoritmos y es lo que denomina "la posibilidad de escalar". No es lo mismo un algoritmo para clasificar los gustos de un usuario de Netflix, que el algoritmo que clasifica el riesgo de reincidencia de un imputado, por las consecuencias perjudiciales que puede acarrear el empleo de estas aplicaciones de Inteligencia Artificial. Por ello, la confiabilidad en un sistema debe aumentar en la medida de la importancia de las decisiones que toma.

Finalmente, mientras más trazable sea la forma en que tomó la decisión, más pueden aminorarse los problemas elevados por los otros dos factores, en la medida en que se crea transparencia en el sistema, lo que conduce a considerar la siguiente dimensión.

19 *https://www.pagina12.com.ar/307642-cordoba-las-insolitas-traducciones-de-sus-ciudades-en-la-pag*

OFUSCACIÓN

Tomar conciencia de que los algoritmos operan alrededor nuestro (awareness) es correr el primer velo de una de las capas de oscuridad en las que esta tecnología prospera.

La dimensión de ofuscación pretende comprender los distintos matices en los que la Inteligencia Artificial opera en forma velada para los creadores, usuarios, y/o destinatarios de estas tecnologías, tanto en los procesos de su desarrollo, como en su aplicación.

Dentro de esta dimensión, encuentro diferentes manifestaciones de un mismo fenómeno. Por "**opacidad" de la Inteligencia Artificial** habré de referirme a la falta de transparencia en el acceso a comprender su funcionamiento, sea por razones de complejidad técnica para quienes carecen de conocimientos en el área (opacidad técnica) o por dispositivos legales que lo transforman en una caja negra (opacidad legal).

A su vez, con **"inescrutabilidad"**, me refiero a la incapacidad -aun para quienes la desarrollan o son expertos en el área- de comprender cómo estos sistemas arriban a un resultado.

Finalmente, a través de "**explicabilidad"** y **"trazabilidad",** pretendo poner el acento en la forma en que los sistemas proveen justificación de sus acciones, y cómo sus "razonamientos" pueden ser recreados para su comprobación.

Frank Pasquale (2015) entiende que la opacidad se deriva del secreto y clasifica sus causas en un triple espectro: secreto real (***passwords*** y medidas de seguridad para proteger información privilegiada), secreto legal (derivada de obligaciones legalmente impuestas para mantener el secreto de la información), y ofuscación (técnicas que ponen una cortina de humo que hace difícil traspasar el secreto). Si por ejemplo el código fuente del algoritmo no está disponible, pero sí la documentación, esta puede ser de un volumen tal que impida su control, lo que hace que el algoritmo siga siendo opaco.

Por mi parte, en la clasificación que propongo, la opacidad de los algoritmos proviene tanto de la propia naturaleza y complejidad técnica para comprender y acceder a la lógica de su funcionamiento (**opacidad técnica**), como de las protecciones legales y comerciales impuestas para proteger la inversión económica realizada en la obtención del modelo (**opacidad legal)**.

La opacidad técnica se explica por sí misma. Para la mayoría de los no iniciados en ciencia de datos, programación, estadística y matemática, tratar de entender los procesos, elementos e intrincaciones de sistemas de Inteligencia Artificial es una tarea imposible. Esta falta de transparencia de los sistemas de Inteligencia Artificial pone a los usuarios en un lugar difícil, en el que deben aceptar los resultados casi como un dogma, o como la dinámica asimétrica de los contratos de aceptación, en derecho del consumidor.
Como se ha demostrado en capítulos anteriores, algunos algoritmos funcionan de una forma en la que existen valores de entrada que, tamizados por distintos procesos, brindan una salida; en forma similar a los pasos de una receta.

Como usuarios no especialistas, en el mejor de los casos, conoceremos qué entradas damos al sistema (***inputs***) y qué salidas obtenemos de estos (***outputs***),

pero los procesos intermedios, y fundamentalmente las decisiones que los sistemas toman (o que sus desarrolladores tomaron al construirlos), permanecen en la oscuridad.
Esencialmente, los mecanismos internos de funcionamiento de los algoritmos permanecerán fuera de nuestro campo de conocimiento y, aunque obtendremos una respuesta, desconocemos por qué procesos se llegó a ella.

La opacidad técnica proviene tanto de la altísima complejidad de las arquitecturas de algoritmos, como de la especialización necesaria para entenderlos. La ciencia de datos es un cruce de varias disciplinas, de ahí que se requieren conocimientos complejos de estadística, matemática, ciencias de la computación, lenguajes de programación, bases de datos, etc. Actualmente, es una de las profesiones mejor pagas y con alta demanda, ya que se estima que no existe cantidad suficiente de especialistas para cubrir las necesidades del mercado. Se suele decir que un científico de datos es alguien que es mejor en estadística que cualquier ingeniero de software y mejor en ingeniería de software que cualquier estadístico.

Esta "***algorithmic literacy***" separa a los expertos como intérpretes capaces de comprender las complejidades de los algoritmos. Sin estos conocimientos, aunque se nos presentara la totalidad de la información, seríamos incapaces de interpretarla.

A su vez, la **opacidad legal** deriva de los distintos dispositivos legales por los que se crea artificialmente una caja negra legal que impide acceder a conocer el entrenamiento o el funcionamiento de los sistemas. Esto puede ser mediante regulaciones públicas que condicionan el acceso (como las normas de protección de la propiedad intelectual vinculadas a "DRM", o ***digital rights management***); contratos entre las partes (que prohíben en sus cláusulas la ingeniería inversa o la confidencialidad de los desarrolladores); o términos y condiciones de uso respecto de consumidores (vetando cualquier acceso o análisis no expresamente autorizado).

En todos estos casos, la consecuencia es un dispositivo artificial que impide el acceso o la divulgación del funcionamiento o de la composición de un modelo algorítmico, oscureciéndolo incluso para quienes tienen la capacidad técnica de entenderlo.

La opacidad afecta entonces directamente la posibilidad de examinar la composición del modelo, o incluso los datos con los que fue entrenado, para determinar si existen errores que puedan comprometer la eficacia del objetivo que debe cumplir, o si se ve comprometida su precisión. Por ello es que en determinadas áreas, sobre todo de carácter público con un potencial impacto negativo sobre los ciudadanos, se debe propugnar que las aplicaciones de tecnología algorítmica respondan a desarrollos ***open source***.

Cuando estas cajas negras algorítmicas (***algorithmic black boxes***) también son indescifrables para sus creadores, nos encontramos frente a un escenario distinto al que denomino **inescrutabilidad**.

Se trata de sistemas inefables, de una complejidad técnica de una magnitud tan diferente que los coloca fuera de escala: son arquitecturas de Inteligencia Artificial de una complejidad inabarcable incluso para los expertos, que aunque tengan una idea de su funcionamiento en un determinado grado de abstracción, no pueden explicar la inteligibilidad de muchas de las respuestas y procesos. Pueden dar una respuesta por la arquitectura y el diseño de los sistemas de Inteligencia Artificial, pero no pueden explicar cómo llegaron a ella. Tampoco quienes las diseñan pueden entender o justificar el resultado al que se arriba, a pesar de su conocimiento especializado en el campo. Por esto es que la inescrutabilidad es diferente de la opacidad, incluso aunque se cuente con la capacidad técnica, el propio diseño o el volumen lo hace ininteligible, como ocurre, por ejemplo, en las redes neurales profundas compuestas de miles de unidades en múltiples niveles.

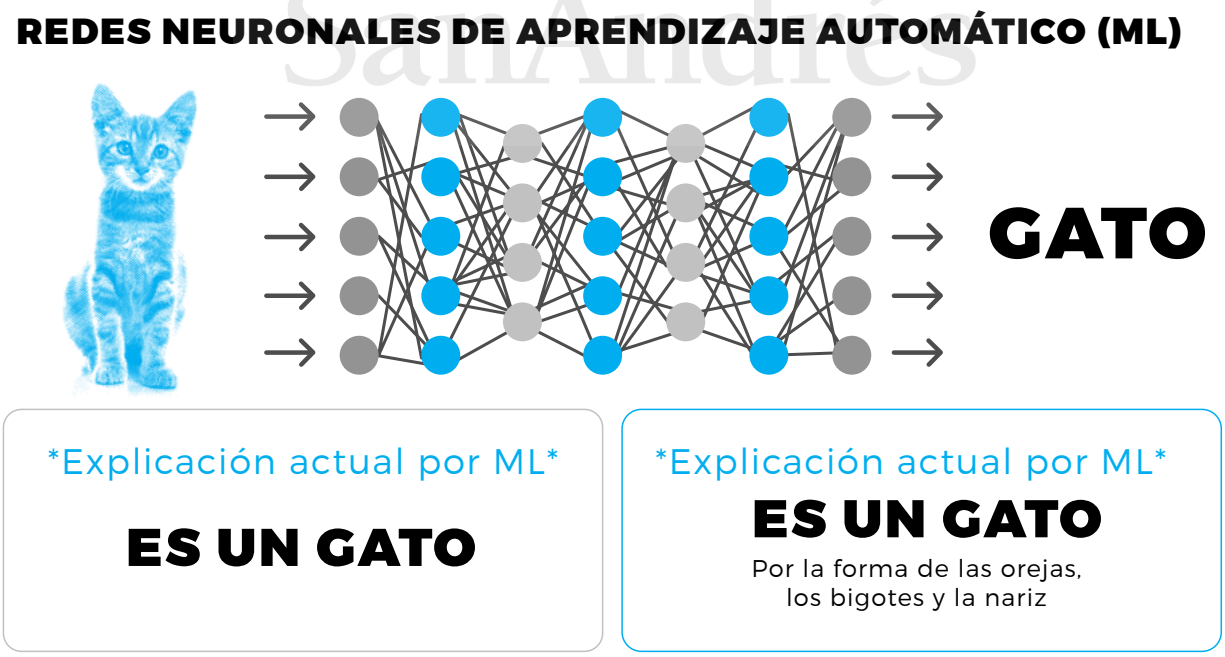
Esto se vincula con la **trazabilidad y explicabilidad** de las respuestas que pueden brindar los sistemas de Inteligencia Artificial.

En los modelos lineales se puede establecer una correlación directa entre el peso de los valores y el resultado: en los árboles de decisiones, se puede seguir la lógica de las ramas de decisión para entender el mecanismo por el cual esta elección se forma. Por ejemplo, para encontrar la respuesta a por qué elegí ponerme hoy un determinado calzado, puedo retrotraerme a una serie de preguntas que van desde el clima, la cantidad de horas fuera de casa, la comodidad, las actividades en la agenda, etc., y circunscribir los factores que me condujeron a la selección.

Otro ejemplo sencillo de ver son las líneas telefónicas de atención automatizada, donde navegamos distintos menús de opciones que van eliminando posibilidades, hasta que arribamos a la opción deseada. Si queremos reconstruir cómo llegamos a hablar con un operador, podemos trazar el árbol de decisiones tomadas hasta llegar a esa opción, y analizar el peso de cada decisión individual hasta arribar ahí.

Por el contrario, las redes neuronales profundas (DNN) son extremadamente complejas, no lineales, y pueden tomar en consideración para sus "decisiones" más de 100 millones de parámetros, lo que dificulta explicar cómo se forman (Rothe, 2017).

Es por ello que muchas organizaciones, particularmente la ***Defense Advanced Research Projects Agency*** (DARPA) de Estados Unidos, están invirtiendo en la investigación de nuevas arquitecturas dentro de lo que se denomina **XAI (*explainable artificial intelligence*),** con el objetivo de crear un conjunto de técnicas de aprendizaje automático que puedan producir modelos más explicables, que mantienen las características de performance y precisión de predicción, mientras que también permiten que los humanos entiendan, confíen apropiadamente y administren estas entidades inteligentes[20] ***("Explainable Artificial Intelligence" n.d.).***



Decisiones y su justificación en modelos de ***machine learning*** vs. ***XAI (explainable artificial intelligence)***

20. https://www.darpa.mil/program/explainable-artificial-intelligence

En esta línea, se buscan desarrollar métodos para enseñar a una IA a describir el contenido de una imagen. Primero, se crearon dos redes neuronales: una que reconocía las imágenes y otra dedicada a la traducción de idiomas. Se las unió y se les suministraron miles de imágenes rotuladas (***labelled data***). Cuando la primera red aprendió a reconocer los objetos en una imagen, la segunda simplemente observó lo que estaba sucediendo en la primera, y luego aprendió a asociar ciertas palabras con la actividad que veía. Al trabajar juntas, las dos redes pudieron identificar las características de cada imagen y luego etiquetarlas (Kuang, 2017).

Un ejemplo desde el diseño es Google Clips[21] ("The UX of AI" n.d.), que explica intuitivamente al usuario la lógica detrás del funcionamiento de la aplicación y la selección de imágenes.

Investigadores del Berkman Klein Center de Harvard resumen certeramente algunas de las preguntas clave para que una Inteligencia Artificial pueda dar una explicación de sus conclusiones, en un artículo en el que se preguntan si cambiaría la decisión si cambia un determinado factor; o por qué dos casos con apariencia similar obtienen decisiones diferentes, o viceversa: o cuáles fueron los principales factores en una decisión (Doshi-Velez et al., 2017).

A nivel regulatorio, bajo el número 679/16, entró en vigencia en mayo de 2018 el **Reglamento General de Protección de Datos para la Comunidad Europea**[22], imponiendo lo que algunos consideran un "derecho a la explicación".

A diferencia de una directiva, que solo funciona como un marco que debe ser transpuesto a las regulaciones nacionales en la forma en que estas lo entendieran pertinente, el Reglamento restringe esta discrecionalidad y es aplicable directamente a los países integrantes de la Comunidad Europea.

La importancia que este acuerdo reviste para la economía global deriva del impacto sobre la regulación del flujo transfronterizo de datos, que solo podrá efectuarse entre países cuya normativa haya sido declarada adecuada.

Sobre las decisiones automatizadas, la norma toma como guía rectora la importancia de la interpretabilidad humana en el diseño de los algoritmos, restringiendo las decisiones a escala individual que puedan afectar significativamente a los usuarios, de ahí que la doctrina se cuestione si ha creado un "derecho a la explicación", por el cual los usuarios pueden interrogar sobre una decisión algorítmica que los afecte.

El inc. 4) del art. 4) incorpora el concepto de ***"profiling"*** y define la "elaboración de perfiles" como "...toda forma de tratamiento automatizado de datos personales consistente en utilizar datos personales para evaluar determinados aspectos personales de una persona física, en particular para analizar o predecir aspectos relativos al rendimiento profesional, situación económica, salud, preferencias personales, intereses, fiabilidad, comportamiento, ubicación o movimientos de dicha persona física".

Se compone así de tres elementos: a) debe tratarse de una forma automatizada de procesamiento, b) debe recaer sobre datos personales, y c) el propósito debe ser evaluar aspectos de una persona con el fin de predecir cuestiones a su respecto.

---

21. Josh Lovejoy, "The UX of AI: Using Google Clips to understand how a human-centered design process elevates artificial intelligence". Disponible en: *https://design.google/library/ux-ai/*

22. *https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32016R0679&from=ES*

El Considerando 24 ilumina los aspectos tenidos en cuenta para conceptualizar si una actividad "controla el comportamiento de los interesados", para lo cual debe evaluarse si "...las personas físicas son objeto de un seguimiento en Internet, inclusive el potencial uso posterior de técnicas de tratamiento de datos personales que consistan en la elaboración de un perfil de una persona física con el fin, en particular, de adoptar decisiones sobre él o de analizar o predecir sus preferencias personales, comportamientos y actitudes….".

En cuanto a la información que deberá comunicarse al interesado cuando los datos se obtengan directamente de su persona, el art. 13 inc. f) impone que debe hacerse saber "...información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas de dicho tratamiento para el interesado…". Conforme el inc. g) del art. 14, este dispositivo deberá cumplirse aun cuando los datos no sean recabados directamente del interesado.

El artículo 15 establece el derecho a obtener del responsable del tratamiento la confirmación de si se están o no tratando datos personales que conciernen al requirente. Si así fuera, se debe garantizar el acceso a los datos personales y, si mediaran decisiones automatizadas, la información significativa sobre la lógica aplicada para el tratamiento y las consecuencias que pudieran derivarse para la persona.

Estos derechos operan con la disposición del artículo 22, norma central sobre la aplicabilidad de las decisiones automatizadas en el plano individual.

El principio rector es que el interesado tiene derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado de datos, que produzca efectos jurídicos o le afecte significativamente.

Las excepciones se consignan en el apartado 2 para los casos en que la decisión:

A. sea necesaria para la celebración o ejecución de un contrato entre el interesado y uno de los responsables del tratamiento;

B. esté autorizada por el Derecho de la Unión o de los Estados miembros que se aplique al responsable del tratamiento, debiendo asimismo establecerse medidas adecuadas para custodiar los derechos del interesado;

C. se base en el consentimiento explícito del interesado.

Pero aun dentro de las excepciones de los inc. a y c, el responsable del tratamiento debe adoptar las medidas adecuadas para salvaguardar los derechos, libertades e intereses legítimos del interesado, garantizando como piso el "derecho a obtener intervención humana por parte del responsable", "a expresar su punto de vista" y "a impugnar la decisión".

El Reglamento prohíbe expresamente el uso de datos sensibles como base de decisiones automatizadas, con la salvedad de que exista consentimiento expreso y que no exista una prohibición legal de brindarlo, o que sea necesario procesar esos datos por razones de interés público (como pudiera ser para a través de las técnicas

de ***machine learning*** mapear el patrón de diseminación de una epidemia). El considerando 71[23] justifica la decisión legal adoptada, ejemplificando casos de aplicación en las denegaciones automáticas de solicitudes de crédito o contrataciones online donde no medie intervención humana. En definitiva, se trata de casos en que el procesamiento apunta a analizar determinados datos para realizar predicciones sobre la fiabilidad, salud, preferencias, rendimiento de un individuo a fin de concertar o no una negociación, o fijar las tarifas en relación con dichos parámetros.

Sintetizando, agrupados bajo el denominador conceptual común de **"**garantías apropiadas", del conjunto de las normas resulta un set de derechos en favor del interesado:

1. a obtener información significativa sobre la lógica aplicada, la importancia y las consecuencias previstas de dicho tratamiento para el interesado;
2. a solicitar la intervención humana en una decisión;
3. a expresar su punto de vista;
4. a recibir una explicación de la decisión tomada después de tal evaluación;
5. a impugnar la decisión.

Más allá de la intención legislativa, el problema se presenta en la operatividad práctica de estas disposiciones confrontadas con las características de los sistemas de aprendizaje automático. Además de la complejidad técnica para poder explicar su funcionamiento a quienes carecen de conocimiento en el campo, los propios expertos concuerdan que determinados procesos, como los que ocurren en redes neuronales profundas, operan en un nivel de abstracción cuya lógica no puede ser siempre explicada y en los que los cálculos resultantes de la interconexión entre los niveles de neuronas artificiales no pueden cuantificarse con precisión (como expliqué al tratar la **inescrutabilidad).** Los sistemas son extremadamente complejos, con una magnitud que escapa a la cognición humana. En otra escala, sucede algo similar a las vías sinápticas del cerebro humano, donde exis-

23. 71) El interesado debe tener derecho a no ser objeto de una decisión, que puede incluir una medida, que evalúe aspectos personales relativos a él, y que se base únicamente en el tratamiento automatizado y produzca efectos jurídicos en él o le afecte significativamente de modo similar, como la denegación automática de una solicitud de crédito en línea o los servicios de contratación en red en los que no medie intervención humana alguna. Este tipo de tratamiento incluye la elaboración de perfiles consistente en cualquier forma de tratamiento de los datos personales que evalúe aspectos personales relativos a una persona física, en particular para analizar o predecir aspectos relacionados con el rendimiento en el trabajo, la situación económica, la salud, las preferencias o intereses personales, la fiabilidad o el comportamiento, la situación o los movimientos del interesado, en la medida en que produzca efectos jurídicos en él o le afecte significativamente de modo similar. Sin embargo, se deben permitir las decisiones basadas en tal tratamiento, incluida la elaboración de perfiles, si lo autoriza expresamente el Derecho de la Unión o de los Estados miembros aplicable al responsable del tratamiento, incluso con fines de control y prevención del fraude y la evasión fiscal, realizada de conformidad con las reglamentaciones, normas y recomendaciones de las instituciones de la Unión o de los órganos de supervisión nacionales y para garantizar la seguridad y la fiabilidad de un servicio prestado por el responsable del tratamiento, o necesario para la conclusión o ejecución de un contrato entre el interesado y un responsable del tratamiento, o en los casos en los que el interesado haya dado su consentimiento explícito. En cualquier caso, dicho tratamiento debe estar sujeto a las garantías apropiadas, entre las que se deben incluir la información específica al interesado y el derecho a obtener intervención humana, a expresar su punto de vista, a recibir una explicación de la decisión tomada después de tal evaluación y a impugnar la decisión. Tal medida no debe afectar a un menor.

A fin de garantizar un tratamiento leal y transparente respecto del interesado, teniendo en cuenta las circunstancias y contexto específicos en los que se tratan los datos personales, el responsable del tratamiento debe utilizar procedimientos matemáticos o estadísticos adecuados para la elaboración de perfiles, aplicar medidas técnicas y organizativas apropiadas para garantizar, en particular, que se corrigen los factores que introducen inexactitudes en los datos personales y se reduce al máximo el riesgo de error, asegurar los datos personales de forma que se tengan en cuenta los posibles riesgos para los intereses y derechos del interesado y se impidan, entre otras cosas, efectos discriminatorios en las personas físicas por motivos de raza u origen étnico, opiniones políticas, religión o creencias, afiliación sindical, condición genética o estado de salud u orientación sexual, o que den lugar a medidas que produzcan tal efecto. Las decisiones automatizadas y la elaboración de perfiles sobre la base de categorías particulares de datos personales únicamente deben permitirse en condiciones específicas.

ten conexiones y resultados que no han podido ser aún decodificados, impidiendo la trazabilidad lineal de las decisiones, lo que en el caso de la IA se magnifica exponencialmente en complejidad y volumen[24].

En la forma en que estos principios se interpreten y en la subsecuente jurisprudencia que las cortes internacionales hagan de su texto, se pondrán en juego nuevos incentivos para el desarrollo en XAI. El campo es incipiente y promisorio para un avance hacia desarrollos de Inteligencia Artificial que permitan encontrar **explicaciones significativas** ***(meaningful explanation***) para quienes interactúan con tecnologías basadas en decisiones algorítmicas.

24. "No se puede mirar dentro de una red neuronal profunda para ver cómo funciona. El razonamiento de una red está embebido en el comportamiento de miles de neuronas simuladas, organizadas en docenas o a veces cientos de capas intrincadamente interconectadas. Cada una de las neuronas de la primera capa recibe una entrada, como la intensidad de un píxel en una imagen, y luego realiza un cálculo antes de emitir una nueva señal. Estas salidas se transmiten, en una compleja red, a las neuronas de la siguiente capa, y así sucesivamente, hasta que se produce una salida global. Además, hay un proceso conocido como retropropagación que ajusta los cálculos de las neuronas individuales de manera que la red aprende a producir un resultado deseado". Traducción propia, accedido el 20-06-17. Disponible en;
https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

SESGO (*BIAS*)

Como vimos, las técnicas de aprendizaje automático basan su posibilidad de aprender en la generalización de otros datos en relación con los cuales fueron entrenados. Sin embargo, no existe una garantía de que esa generalización se realice de forma objetiva.

Tras la apariencia de neutralidad, los algoritmos **trasponen el sesgo tanto de quienes los crean como de los propios datos con que se alimentan**, y los cristalizan en un modelo que los reproducirá y que reforzará los estereotipos de valor.

De tal manera, analíticamente podemos ver que el sesgo puede presentarse:

1. en los ***datos*** mismos
- **a.** de manera explícita (cuando se incorporan de por sí como variables de la fórmula datos sensibles o discriminatorios), o
- **b.** de manera implícita (en tanto los mismos datos ya contengan una carga de prejuicio o discriminación oculto que se traslada al modelo, que es imperceptible al observador no entrenado o por su propio sesgo cognitivo);

2. en el empleo que se haga de esos datos para la construcción del modelo (a través de una errónea selección que importe suprimir datos o realizar asociaciones incorrectas en virtud de un prejuicio propio);

3. en la interpretación que se haga de los resultados obtenidos de un modelo basado en Inteligencia Artificial.

Los algoritmos son hechos por humanos y, como tales, no pueden ser absolutamente neutrales, "los puntos ciegos de un modelo reflejan los juicios y las prioridades de su creador […] a pesar de su reputación de imparcialidad, reflejan objetivos e ideología. […] Nuestros propios valores y deseos influyen en nuestras elecciones, desde los datos que elegimos recopilar hasta las preguntas que formulamos. Los modelos son opiniones integradas en las matemáticas" (O'Neil, 2016)[25]. "Si bien a menudo pensamos que los términos 'big data' y 'algoritmos' son benignos, neutrales u objetivos, son todo lo contrario. Las personas que toman decisiones poseen todo tipo de valores, muchos de los cuales promueven abiertamente el racismo, el sexismo y las falsas nociones de meritocracia, lo cual está bien documentado en estudios de Silicon Valley y otros corredores tecnológicos" (Noble, 2018).

Un algoritmo es simplemente una lista ordenada de pasos que, cuando se le dan entradas, produce salidas. En el mundo analógico, una receta es un buen ejemplo de un algoritmo: dada la entrada de ingredientes (café, leche, chocolate), siguiendo los pasos

25 Traducción propia de: *"A model`s blind spots reflect the judgments and priorities of its creator […] models, despite their reputation for impartiality, reflects goals and ideology. […] Our own values and desires influence our choices; from the data we choose to collect to the questions we ask. Models are opinions embedded in mathematics."*

de preparación, obtengo un capuccino, o un cold brew. Nunca hay que perder de vista que cuando hablamos de algoritmos, existe una correlación entre sus entradas y salidas, que los resultados dependen de sus entradas, como si fueran ingredientes. Si no tengo los ingredientes adecuados, aunque siga la receta al pie de la letra, nunca voy a obtener el resultado buscado; no puedo producir té earl grey a partir de granos de café.

Por eso es tan importante la calidad de los datos de entrada: si estos datos son discriminatorios, las salidas lo serán en alguna medida (lo que se suele resumir en la frase ***garbage in, garbage out***). Los sesgos que se encuentran en los datos de entrenamiento son al modelo final lo que una gota de tinta en un vaso de agua, terminan filtrándose y tiñendo el resultado, aun en formas imperceptibles.

Las máquinas aprenden de datos, pero no solo de los datos puros y duros, sino de los datos que los humanos les compilan y les facilitan, y ello encierra nuestras propias asociaciones colectivas sobre ellos. Al ser extraídos como recorte del mundo, existe la asunción implícita de que son representativos del problema que se intenta modelar, pero en esto juega la subjetividad humana y personal de quien efectúa ese recorte. En cada charla y clase siempre repito que "los algoritmos miran al pasado para construir el futuro", porque este recorte de la realidad se realiza basándose en algo que ya existe dentro del mundo, pero buscando predecir o clasificar una nueva realidad, que es una entre muchas hipótesis, porque está en el futuro.

Los prejuicios están fuertemente enraizados en el lenguaje natural, así que se trasladan al aprendizaje semántico que las máquinas realizan, con el peligro de que no tienen incorporadas salvaguardas éticas o de sentido común para detectarlo. En el año 2016 una IA creada por Microsoft para interactuar en Twitter con usuarios norteamericanos de entre 16 y 24 años y aprender del lenguaje disponible en Twitter tuvo que ser desconectada al día siguiente de su lanzamiento. En menos de 24 horas, y como consecuencia de las interacciones aprendidas y recibidas en la red social, la IA se volvió misógina y antisemita (Hunt, 2016; Paul, 2016; Victor, 2016), reproduciendo insultos generados por otros usuarios y hasta negando el Holocausto.

Si en las toneladas de información que procesan, las personas de color se encuentran asociadas semánticamente más frecuentemente al delito o a connotaciones lingüísticas negativas, las máquinas asumen la correlación y la trasladan, con un efecto expansivo por su escalabilidad, a todos los sistemas que los tomen como base.

Este es el caso de las redes sociales, así como también de los buscadores basados en algoritmos de grafos de asociación, en los que en la función de autocompletar se han visto traslucidos prejuicios muy arraigados. En 2010, la investigadora en comunicación Safiya U. Noble estaba buscando sitios web para compartir con su sobrina y sus amigas. Al buscar "***black girls***" en Google, esperando ser dirigida a sitios educativos, o con actividades para gente joven, el primer resultado fue en cambio el sitio ***HotBlackPussy.com.*** Toda la primera página contenía resultados similares. La búsqueda del término llevaba mayormente no a sitios de niñas de color, sino a páginas pornográficas o con imágenes sugestivas o adultificadas de jóvenes afroamericanas, lo que llevaba a una sexualización y objetivación de la mujer de color (Noble, 2018).

A la hora de buscar responsables, la culpa de estas asociaciones se atribuye a la estadística: el buscador devuelve los resultados más populares de búsquedas anteriores. En una especie de democracia aritmética, las asociaciones y conexiones mayoritariamente realizadas son las que van a alcanzar la primera página de resultados. Por cómo funcionan los algoritmos que posicionan los resultados de búsqueda, y el refuerzo entre los grafos de los sitios más visitados, los resultados que el buscador arroja son la composición de un número de factores. De esta manera opera PageRank, uno de los algoritmos de búsqueda empleados por Google. El input de este sistema es la búsqueda, y el output, la información "filtrada", considerando un análisis de las palabras buscadas,

un match de la búsqueda y un ranking de las páginas considerando el contexto ("How Google Search Works | Search Algorithms" n.d.).

Actualmente, existen técnicas que permiten mejorar la posición dentro de los resultados de búsqueda, a través de lo que se denomina SEO, un acrónimo de ***Search Engine Optimization.*** El SEO orgánico se refiere a los métodos utilizados para obtener una mejor ubicación o ranking dentro de los resultados de un motor de búsqueda, sin tener que pagar. Los anunciantes también pueden pagar para que sus resultados aparezcan en el tope de la lista, en lo que se llama "resultados promocionados" (***inorganic SEO).***

Ahora, en el estado actual de Internet en el marco de la economía de la atención y la magnificación del impacto publicitario, los resultados de las búsquedas han emprendido un camino hacia la personalización de los resultados en función de la persona, el lugar y el historial previo de búsquedas.

En la web contextual o personalizada, los resultados no dependen solo de la "opinión de las mayorías", sino que el acento está puesto en que el algoritmo conoce al usuario, por lo que el contexto de información que tiene sobre quien pregunta lo ayuda a ajustar el resultado más específica e individualmente. Por ello, frente a las mismas ***keywords***, los resultados de búsqueda que brinde a uno u otro usuario serán diferentes.

En este tema se advierte otro de los peligros que mencionamos en los fenómenos algorítmicos: el sesgo de autoconfirmación y las profecías autocumplidas**.** En su búsqueda de la eficiencia, el modelo proporcionará la respuesta que aprendió es la más generalmente asociada, aun cuando esta resulte discriminatoria, porque justamente no tiene la capacidad de darse cuenta de que lo es.

A una escala mayor, colectivamente, produce un refuerzo del sesgo cognitivo de los usuarios, que ven confirmado lo que ya pensaban (***confirmation bias***), y que son por ello psicológicamente más proclives a aceptarlo, más que a enfrentar una disonancia cognitiva. Fenómeno que ha contribuido perniciosamente a lo que se denominan las "burbujas de filtro" (***filter bubbles)*** y a la radicalización creciente de grupos en los que la opinión común y los algoritmos funcionan como una caja de resonancia que aísla del disenso.

El prejuicio no solo se esconde en la carga simbólica y semántica de los datos, sino también en la selección, en tanto ese recorte de la realidad se hace en función de la visión del mundo que trae quien lo realiza. Por ello, es tan necesaria la **diversidad e interdisciplinariedad** en los equipos de desarrollo, con un acompañamiento ético a lo largo de todo el proceso, lo que se enlaza con la dimensión de **equidad, *fairness* e inclusión.**

El problema de la muestra en tanto recorte abarca tanto lo que se incluye como lo que se excluye, que puede transformarse en un dato en sí mismo que oblitere una categoría por exclusión. La ausencia de datos puede hacer que particularidades de un cluster de elementos desaparezcan, y puntos de datos que deberían agruparse conjuntamente bajo un denominador común, terminen redistribuidos en otros, haciendo desaparecer una categoría.

Como se explicara, por su arquitectura, el aprendizaje automático requiere cantidades ingentes de datos, por lo que la máxima es que mientras mayor sea la información de la que pueden nutrirse, más ajustadas serán las respuestas. Domingos (2017: 24) es claro en este sentido "todos los principales algoritmos de aprendizaje -incluyendo al vecino más cercano, árboles de decisión y redes bayesianas, que son una generalización de Naive Bayes- son universales en el siguiente sentido. Si le das suficientes datos que sean apropiados, pueden aproximar cualquier función arbitrariamente cerca, que es el dialecto matemático para decir pueden aprender

cualquier cosa. La trampa es que 'suficientes datos' podría ser infinito. Aprender de los datos finitos requiere hacer suposiciones[26]."

Aprender sobre cantidades limitadas de datos requiere hacer asunciones y esta es una de las vías por las que el sesgo puede contaminar la construcción del modelo: "Cada algoritmo de aprendizaje hace un conjunto de suposiciones sobre los datos para encontrar un modelo único, y este conjunto de suposiciones se denomina sesgo inductivo del algoritmo de aprendizaje"[27] (Mitchell, 1997 apud. Alpaydin, 2016).

Por un lado, un gran volumen de datos conlleva la necesidad de mucho poder computacional para procesarlos, por lo que a veces se realizan recortes en función de la ecuación económica del costo del procesamiento.

Por otra parte, a pesar de la gran cantidad de datos disponibles, en ciertos casos se carece de la información necesaria para realizar una predicción (sea por falta de datos o por falta de acceso a ellos, en la medida en que la base de datos esté protegida legalmente y requiera una licencia paga), y determinados parámetros se sustituyen por otros (***data proxies***). Así, se emplean relaciones como el código postal, o patrones de lenguaje para determinar si esa persona será capaz de pagar puntualmente su crédito o performar bien en un trabajo. Estas correlaciones cargan con prejuicios y pueden ser tan inexactas y discriminatorias como ilegales. El peligro radica en que no siempre están explícitas, o que difícilmente puede mensurarse o determinar el peso que poseen en el resultado final.

Los datos seleccionados **deben ser pertinentes y relevantes** para lo que se está tratando de clasificar o predecir y, frente a su falta, ser cuidadoso con su sustitución o reemplazo. Una mala calidad en los datos o su reducción dimensional[28] de manera incorrecta pueden determinar que el modelo incurra en una falacia al haber asumido incorrectamente que una correlación estadística implica causalidad.

Como disciplina con rigor estadístico y una gran base de datos para trabajar, O'Neil toma el ejemplo del baseball. Los datos empleados están directamente relacionados con la performance de los jugadores, son relevantes para los resultados que intentan predecir: "Esto puede sonar obvio, pero como veremos a lo largo de este libro, la gente que construye armas de destrucción matemática masiva rutinariamente carece de datos sobre los comportamientos que más le interesan. Por lo tanto, sustituyen los datos con suplentes o proxies. Trazan correlaciones estadísticas entre el código postal de una persona o los patrones de lenguaje y su potencial para pagar un préstamo o manejar un trabajo…" (O'Neil, 2016).

26. Traducción propia de: "*In fact, all the major learners -including nearest neighbor, decision trees, and Bayesian networks, a generalization of Naive Bayes- are universal in the following sense. If you give the learner enough of the appropriate data, it can approximate any function arbitrarily closely- which is math-speak for learning anything. The catch is that "enough data" could be infinite. Learning from finite data requires making assumptions*."

27. Traducción propia de: "*Every learning algorithm makes a set of assumptions about the data to find a unique model, and this set of assumptions is called the inductive bias of the learning algorithm".*

28. Como su nombre lo indica, es un conjunto de técnicas que reduce el número de dimensiones de un conjunto de variables para, en definitiva, reducir la complejidad y el costo en el procesamiento de la información para el entrenamiento de los modelos. En esa reducción está implícito el peligro de los sesgos de realizar correlaciones espurias entre variables y falacias de causa y efecto que no se sostienen lógicamente.

Sobre el **sesgo** en la construcción de los modelos y para garantizar la transparencia, el Considerando 71 del GDPR justifica un set de deberes imponibles al responsable del tratamiento:

1. utilizar procedimientos matemáticos o estadísticos adecuados para la elaboración de perfiles,
2. aplicar medidas técnicas para corregir las inexactitudes en los datos personales a fin de reducir el riesgo de errores,
3. asegurar los datos personales de forma que se tengan en cuenta los posibles riesgos para los intereses y derechos del interesado.

Todo esto, con el fin de impedir la producción de efectos discriminatorios sobre personas físicas por motivos de raza, etnia, opinión política, religión, creencias, afiliación sindical, condición genética o estado de salud u orientación sexual.

También introduce el principio de que las decisiones automatizadas y el profiling basado en categorías particulares de datos sólo puede admitirse en condiciones específicas.

Como puede apreciarse, el sesgo en temas de Inteligencia Artificial unido a su alcance escalable se conecta fuertemente con el impacto sobre la dimensión de **inclusión y equidad**, en tanto reproduce y amplifica la discriminación, reforzando los estereotipos y perpetuando la desigualdad.

AUDITABILIDAD Y RESPONSABILIDAD (*ACCOUNTABILITY*)

Propongo agrupar bajo la dimensión de ***accountability***[29] los matices vinculados al control sobre los algoritmos en cuanto a cómo pueden ser auditados (auditabilidad) y a las formas en que el derecho establece mecanismos de ***redress*** y responsabilidades para trasladar las consecuencias de los daños ocasionados por IA (responsabilidad).

La **auditabilidad** tiene una conexión estrecha con la ofuscación, particularmente en lo que hace a los mecanismos legales que puedan impedir el acceso al modelo para su evaluación. Para poder auditar un sistema, es necesario que exista transparencia en cuanto a los datos de entrenamiento, validación, testeo, documentación de los procesos, modelos adoptados y descartados, salidas, etc.

La opacidad de los algoritmos conduce a que sea difícil poder auditarlos, y por ende detectar sus errores y atribuir responsabilidad por los daños que causen. Para poder validar su carácter científico y defender la corrección de sus postulaciones, los modelos deberían poder ser testeados y someterse a refutación. El secreto conspira contra la auditabilidad, como afirma Pasquale (2015: 41): "...la 'ciencia' del scoring secreto no adopta una salvaguardia clave del método científico: que las generalizaciones y observaciones sean públicamente comprobables..."[30].

Un algoritmo transparente es aquel que facilita su propio escrutinio. Según Shadoan (2014) para ser auditable debe permitir que se examinen:

1. las entradas (datos),
2. las superficies de control (las configuraciones que controlan la forma en que el algoritmo ejecuta sus pasos y cómo impacta en las salidas),
3. los pasos del algoritmo y el estado interno (cómo se llevan a cabo los procesos que transforman las entradas en salidas, y si existe algún elemento sesgado que pueda afectar el resultado o cuestionar la legitimidad del proceso),
4. las suposiciones y los modelos que usa el algoritmo (las asunciones que los algoritmos realizan tanto sobre los datos como sobre las expectativas del usuario. Estas suposiciones deben describirse en

29. Utilizo la palabra inglesa ***accountability*** porque tiene matices semánticos que engloban tanto aspectos de auditabilidad y rendición de cuentas como de responsabilidad, y que no tienen un término equivalente en su traducción al español.

30. Traducción propia de: "***The 'science' of secret scoring does not adopt a key safeguard of the scientific method: publicly testable generalizations and observations***".

detalle, de modo que los usuarios puedan evaluar si el algoritmo está produciendo resultados en línea con sus necesidades), y

5. la justificación de las salidas producidas (permitiendo responder la pregunta: ¿por qué se produjo esta salida a partir de las entradas proporcionadas?).

Uno de los fuertes de las aplicaciones comerciales de los algoritmos es su poder predictivo. El análisis de big data permite tener información privilegiada que posibilita que se intente predecir futuras compras, el comportamiento de un trabajador o una tendencia de voto. Esto se traduce en un cambio en los modelos de negocios que buscan explotar este potencial predictivo ayudando a tomar decisiones informadas basadas en datos estadísticos (***data driven business, insights***).

Como expliqué acerca de la matriz de relevancia de las decisiones, dependiendo del área de aplicación de la solución basada en Inteligencia Artificial, la auditabilidad se vuelve crítica en sectores de alto impacto para la sociedad.

Por ejemplo, cuando se emplean soluciones algorítmicas en el derecho penal, la falta de transparencia es difícilmente compatible con el debido proceso: se requiere una apertura tal que permita un escrutinio suficiente para ejercer el derecho de defensa.

Con el objetivo de brindar una herramienta científica y objetiva que ayude a eliminar sesgos, inequidades o discriminación en la labor de los jueces, estos avances en el área del derecho penal se han desplegado en forma de distintos softwares que emplean algoritmos para dar sustento o guía a las decisiones sobre evaluación de riesgos y reincidencia ***(algorithmic risk assessment and recidivism).***

A pesar de que como venimos demostrando la Inteligencia Artificial presenta aristas controversiales y de problemática constitucional, todo tipo de herramientas de automatización ya se utilizan en distintos tribunales[31]. Particularmente en Estados Unidos, para determinaciones tan sensibles a la condición humana como la libertad condicional, la prisión preventiva y para sentencias.

COMPAS (***Correctional Offender Management Profiling for Alternative Sanctions)*** es uno de los softwares más usados en el sistema judicial norteamericano. Desarrollado por la firma Northpointe, se trata de un grupo de modelos algorítmicos propietarios, basado en los modelos LSI (***Level of Service Inventory***), que son una herramienta actuarial formulada para identificar y clasificar los riesgos de reincidencia futura (Watkins, 2011).

La firma proclama que COMPAS incorpora un enfoque integral de evaluación basado en e incorporando las escalas de las teorías sobre el crimen y la delincuencia más respetadas, incluyendo la Teoría General del Crimen, Oportunidades criminales / Teorías del estilo de vida, Teoría del Aprendizaje Social, Teoría de la Subcultura, Teoría del Control Social, Teoría de las Oportunidades Criminales / Actividades de Rutina, entre otras ("COMPAS Northpointe Frequently Asked Questions"[32]).

31. "Preparación del sector judicial para la inteligencia artificial en América Latina", Machine intelligence Lab, CETYS, Universidad de San Andrés. Disponible en: ***https://cetys.lat/en/readiness-of-the-judicial-sector-for-artificial-intelligence-in-latin-america/***

32. ***http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf***

El uso de estas herramientas se encuadra en la tendencia denominada ***EBS (evidence-based sentencing)***, que impone como ideal un enfoque científico racionalista de decisiones informadas por datos, proporcionados por instrumentos apoyados en modelos estadísticos, que permitan objetivizar la evaluación y la predicción de riesgo realizada en una sentencia criminal. La inclusión de variables estadísticas socioeconómicas y demográficas como antecedentes para conformar el modelo ha motivado cuestionamientos basados en la violación del principio de igualdad ante la ley ***(equal protection clause)*** (Starr, 2013), así como del principio de inocencia. Ello, en tanto implica considerar para la determinación de la libertad o la condena, factores estadísticos externos a la conducta del propio imputado, por lo que se lo ha sindicado como un "castigo de grupo" ***(group based punishment)***[33]***,*** que refuerza estereotipos discriminatorios y castiga a los sectores pobres o marginales de la población.

Asimismo, tratándose de un software privativo que se licencia en distintos paquetes (básico, estándar o avanzado), cuya fórmula se encuentra protegida por distintos dispositivos de propiedad intelectual, se pueden apreciar patentemente los problemas que esto genera en cuanto a su **auditabilidad.** El secreto del proceso por el que arriba a un determinado resultado, así como la imposibilidad de comprobar o refutar la exactitud empírica de sus predicciones, reabre la puerta a formulaciones propias de los procesos inquisitivos de otras épocas, que el derecho penal creía ya superadas.

Como se explica en el capítulo siguiente, correspondiente a propiedad intelectual y algoritmos, una de las formas más adecuadas de proteger la inversión económica y el modelo de negocios a su alrededor, es a través del secreto comercial. Justamente, la falta de divulgación de la información, que es un requisito necesario para que opere la protección legal de esta figura, pone al secreto comercial en tensión con la transparencia necesaria para el empleo de algoritmos en decisiones automatizadas.

A diferencia de otros mecanismos de las leyes de propiedad intelectual, como los derechos de autor y patentes, el secreto comercial no posee un adecuado contrapeso de beneficio social. El fin mediato de estas legislaciones es promover el avance de las ciencias y artes para el beneficio de toda la sociedad, otorgando una compensación de la explotación a los creadores e inventores. En el caso de las patentes y de los derechos de autor, la protección se otorga a cambio de la apertura y de la divulgación del producto intelectual que se quiere proteger.

Por el contrario, el secreto comercial para operar requiere justamente que se preserve la confidencialidad de la información reservada. En lo que se refiere a los algoritmos de sentencias penales perjudiciales, esta falta de previsión de mecanismos de equilibrio social pone a la ley del secreto comercial en tensión con los derechos civiles y garantías constitucionales.

La precisión y la imparcialidad de COMPAS fueron cuestionadas por una investigación de la ONG ProPublica (Angwin et al., 2016), que concluyó que el software era propenso a una discriminación racial, endilgando un riesgo mayor de reincidencia a personas de color frente a personas blancas en similares condiciones. El modelo se mostraba propenso a etiquetar falsamente a acusados de color como futuros delincuentes, a una ratio de casi el doble de la tasa de los acusados blancos. Por el contrario, el sistema brindaba falsos negativos con relación a los acusados blancos, que fueron etiquetados erróneamente como de bajo riesgo con mucha más frecuencia que los acusados negros.

33. Las generalizaciones grupales sobre la peligrosidad atentan contra principios constitucionales penales que marcan la necesidad de una conducta propia punible. Cristalizar en un software las asunciones basadas en estadísticas sobre grupos es una forma insidiosa de violentar estos principios.

Los resultados de esa investigación fueron disputados por Northpointe en un ***white paper*** señalando los errores en los que entiende que ProPublica incurrió (Brennan et al., 2016), concluyendo en que el informe no presentó ninguna evidencia válida de que las escalas de riesgo son parciales contra las personas de color. A su turno, ProPublica rebatió la respuesta de COMPAS, reafirmando la validez de su investigación (Larson y Angwin, 2016).

En un punto de vista interesante que pone de manifiesto lo que he explicado bajo la dimensión de **confiabilidad**, *The Washington Post* (2016)[34] propone que los argumentos matemáticos por los que ambos arriban a distintos conceptos de equidad (***fairness***) son irreconciliables, por lo que ambos tienen parcialmente razón en la validación de sus conclusiones.

La dificultad de comprensión de estos documentos hace casi imposible que los no iniciados en las sutilezas de los modelos algorítmicos puedan comprender los matices, las implicancias y los alcances de las diferencias que ambos se atribuyen. Esta complejidad a la que me refiero como "opacidad técnica", sumada a la falta de transparencia sobre el algoritmo resultante de la propiedad intelectual (opacidad legal), ha motivado que los abogados de la defensa criticaran el uso de COMPAS como una violación al debido proceso, porque los torna incapaces para cuestionar judicialmente los resultados. Asimismo, preocupa el impacto que el puntaje asignado por estos sistemas tiene sobre la cognición de los jueces a la hora de sentenciar. En un artículo sobre el uso de algoritmos en sentencias criminales, el Centro de Información sobre la Privacidad Electrónica relata el caso del juez Babler en Wisconsin, quien anuló el acuerdo arribado entre la defensa y la fiscalía (un año en la cárcel del condado con supervisión posterior) e impuso dos años en la prisión estatal y tres años de supervisión, después de ver que el acusado tenía altos riesgos de futuros delitos violentos y un riesgo mediano de reincidencia general[35] (Electronic Privacy Information Center n.d.). El mismo artículo compara la alarmante realidad de la cantidad de estados norteamericanos que emplean estos sistemas sin haber realizado un estudio previo de validación de resultados.

La validez constitucional de COMPAS fue cuestionada por la defensa en el caso STATE of Wisconsin, Plaintiff–Respondent, v. Eric L. LOOMIS, Defendant–Appellant[36], bajo la premisa de que la falta de acceso a los cálculos efectuados por el sistema para determinar el riesgo tenido en cuenta por el juez al sentenciar afectaba el debido proceso del imputado. En el caso, Loomis fue arrestado en febrero de 2013 acusado de conducir un automóvil que había sido utilizado en un tiroteo. Se declaró culpable de eludir a un oficial, y ya tenía una condena previa por un delito de naturaleza sexual.

El informe de COMPAS determinaba que el imputado presentaba un gran riesgo de violencia y recidivismo, que fue seguido por el juez al sentenciar, indicando a Loomis que era "un gran riesgo para la comunidad" y que por ello la pena era de seis años.

La defensa de Loomis alegó que existió una violación al debido proceso, en tanto la condena se derivaba en parte de un elemento (el riesgo señalado en la escala por COMPAS) que no había sido posible analizar ni cuestionar, por no poder determinar qué parámetros consideraba el algoritmo para arribar a esta conclusión.

---

34. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear", en ***The Washington Post***, Oct 17, 2016. Disponible en: ***https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/***

35. Disponible en: ***https://epic.org/algorithmic-transparency/crim-justice/***

36. STATE of Wisconsin, Plaintiff–Respondent, v. Eric L. LOOMIS, Defendant–Appellant, 2016. Disponible en: ***http://caselaw.findlaw.com/wi-supreme-court/1742124.html***

En la revisión del caso, la Corte Suprema de Wisconsin se decantó por sostener la validez de la herramienta, acompañando el argumento de la corte del circuito, fundado en su consideración de que el puntaje de riesgo atribuido por COMPAS no fue determinante para el resultado final, en tanto la decisión fue respaldada por otros factores independientes.

A su turno, la Corte Suprema de Estados Unidos rechazó dar tratamiento sustantivo a la cuestión, denegando el certiorari interpuesto por la defensa.

De la lectura del caso se aprecian claramente las razones de **auditabilidad** por las que se aboga por implementaciones abiertas ***open source***, que permitan examinar el código de manera pública. Sin embargo, esto no es suficiente. La apertura del código debe ser acompañada por documentación extensiva y clara que haga accesible e inteligible lo que el algoritmo está haciendo.

Confiar en el cumplimiento de principios de una Inteligencia Artificial ética bajo parámetros de autorregulación, cuyo contenido es determinado por las propias empresas, está probando ser insuficiente a los efectos de la **responsabilidad,** de ahí el creciente clima regulatorio.

Nueva York fue pionera al proponer la primera ley para intentar proporcionar transparencia a la forma en que las agencias del gobierno de la ciudad emplean algoritmos. El proyecto de ley formó un grupo de trabajo para examinar cómo las agencias gubernamentales de la Ciudad de Nueva York usan los algoritmos. Tras el lobby, la ley aprobada abandona los requisitos de divulgación del borrador original, y lo sustituye por un grupo de trabajo de determinación de hechos, que deberá desarrollar recomendaciones sobre qué tipos de algoritmos deberían ser regulados, cómo los ciudadanos privados pueden "evaluar significativamente" (***meaningful assess***) cuando las decisiones de los algoritmos pueden afectarlos personalmente y obtener una explicación, y desde el gobierno, cómo abordar "instancias en las que una persona se ve perjudicada" por un sesgo algorítmico (Powles, 2017).

En los años siguientes, distintas legislaciones, declaraciones y tratados han avanzado para tratar de regular la actividad, algunas veces en forma fraccionada, comprendiendo sólo determinados aspectos de la Inteligencia Artificial (como la protección de datos personales, o su aplicación a consumidores).

La **responsabilidad** debe contemplar no sólo el acceso a los datos o la lógica de la decisión, sino fundamentalmente **las inferencias que la Inteligencia Artificial construye sobre nosotros.** Debe habilitar no sólo la posibilidad de cuestionar el algoritmo y los modelos en abstracto, sino también los resultados del caso concreto que afectan a una persona individualmente.

Al catalogarnos, los algoritmos nos colocan **etiquetas invisibles (*invisible labelling*)**[37] para nosotros, creadas

37. Por ***"invisible labelling"*** me refiero al fenómeno de desconocer los rótulos que las aplicaciones de Inteligencia Artificial (mayormente desarrolladas por compañías privadas) crean sobre nuestras personas, y cómo estas nos acompañan moviéndose como un aura invisible en las interacciones de nuestra vida digital. Esto habrá de agravarse aún más si consideramos el actual tránsito hacia una mayor presencia online, inescapable, en el contexto del Metaverso. Estas inferencias son difíciles de combatir, en la medida en que no resultan ser un producto de una única base de datos o compañía, apilándose sobre nuestra presencia virtual. Es muy fácilmente imaginable el escenario de pesadilla que representa este juzgamiento invisible, en bases de datos filtradas o vendidas comercialmente, particularmente en conexión con dispositivos de reconocimiento facial y realidad aumentada (AR), que hacen imposible la desconexión, como mencionamos al referirnos a la pervasividad.

Intuyo que determinados mecanismos de Inteligencia Artificial adversarial podrán proveer mecanismos que impidan nuestra identificación compulsoria en contextos de AR, en forma previa a que la regulación imponga mecanismos que hagan efectiva la necesidad de recopilar información sobre terceros (***bystanders***).

mediante generalizaciones u observaciones basadas en inferencias sobre nuestros datos. Básicamente, mucho del conocimiento que se presenta como certeza es hipotético y basado en asunciones y, como las acusaciones de la Inquisición, aun siendo sujetos de ellas, no tenemos acceso a conocerlas.

Alguien que ha sido señalado por la Inteligencia Artificial como un posible incumplidor de un préstamo, o que padecerá depresión, o con tendencias suicidas, desconoce que este rótulo pesa sobre su cabeza y, por ende, consecuentemente, no puede defenderse o cuestionar la categorización.

Incluso aunque las conociera, es difícil saber de dónde provienen, en tanto es una práctica comercial actual la venta e integración de distintas bases de datos a los fines del marketing.

En una horripilante muestra tanto de la información que poseen como de la insensibilidad algorítmica, una empresa norteamericana de útiles de oficina envió una carta promocional rotulada "Mike Seay/Hija muerta en accidente de autos/O negocio actual" (Merrick, 2014). Cuando el destinatario de la misiva pidió a la empresa tanto una disculpa como saber de dónde provenía la información, no obtuvo ninguna respuesta satisfactoria. La información estaba en todos lados, y a la vez, en ninguno. Era imposible reconstruir el camino de por qué o cómo había llegado a manos de la empresa, o por qué se guardaba ese dato tan personal, ya que existían acuerdos de confidencialidad que impedían su divulgación. Aun cuando se pudiera intentar reconstruir ese camino (en forma muy costosa), a través de ingeniería inversa a partir de los resultados, numerosas cláusulas contractuales o leyes de propiedad intelectual otorgan como derecho el hecho de vetar la posibilidad de ejercerlo.

En el caso citado, un error permitió correr el velo de secreto sobre las inferencias que nos acompañan. Sin embargo, aunque monstruosamente insensible, en este caso, la aserción era correcta: la hija del destinatario había muerto en un accidente. Pero, ¿qué ocurre cuando una información falsa o incorrecta se introduce en el sistema?

En su investigación, Cathy O'Neil da un ejemplo de la pesadilla que puede ser lidiar con una identificación incorrecta en los sistemas: Catherine Taylor aplicó para un trabajo en la Cruz Roja, en la carta por la que la rechazaban indicaba que en sus antecedentes figuraba un cargo criminal por fabricar y vender metanfetaminas. Los cargos, en realidad, correspondían a otra Catherine Taylor, con la que compartía fecha de nacimiento. Luego de investigar, encontró que el error se repetía en varias bases de datos y que la fuente era un data broker llamado Tenant Tracker. El informe de esta compañía estaba plagado de inexactitudes, y vinculaba su identidad a la de una homónima, convicta, vinculando su nombre con el alias de Chantal Taylor. Pudo enterarse de esa información de casualidad al solicitar un crédito, y porque la persona que tenía su informe, confrontó los hechos presentados en el informe con la persona que estaba entrevistando. Con cuestiones tan evidentes como que la convicta tenía un tatuaje y ella no, pudo demostrar su identidad en la entrevista, gracias a la intervención de un ser humano que pudo desentrañar la maraña de datos inexactos procesados por los algoritmos (O'Neil, 2016). El caso terminó siendo litigado bajo la ***Fair Credit Report Act*** (Catherine L. Taylor, Appellee v. Tenant Tracker, Inc., also known as Result Matrix, Inc., Appellant 2013)[38], pero no es la norma. Muchas personas no llegan a enterarse, no pueden obtener una clarificación (por la percepción asumida de que las máquinas no se equivocan es difícil argumentar contra la lógica de que "la computadora o el sistema lo dice") o no pueden demandar.

38. Catherine L. Taylor, Appellee v. Tenant Tracker, Inc., also known as Result Matrix, Inc., Appellant, UNITED STATES COURT OF APPEALS FOR THE EIGHTH CIRCUIT, Mar 28, 2013. Disponible en: ***http://caselaw.findlaw.com/us-8th-circuit/1626899.html***,

Por esto es que es difícil que las medidas instauradas por las leyes de protección de datos personales resulten eficientes, porque estructuralmente están diseñadas para una dinámica entre una persona y una base de datos. Un reclamo es dirigido a una base de datos particular, y deben dirigirse tantos reclamos como bases de datos existan. En la práctica se torna una pelea hercúlea contra la Hidra de Lerna, en la que cuando se logra cortar una cabeza, aparecen dos más.

Los algoritmos están construyendo nuestras identidades digitales y nuestra reputación online. Por eso, la **responsabilidad** como dimensión comprende no sólo la atribución de los daños, sino el estudio de **mecanismos regulatorios** que permitan a escala individual hacerla efectiva de formas no gravosas.

El impacto es global en el ecosistema ético de la IA, el hecho de que los sistemas algorítmicos no rindan cuentas a nivel individual conduce a generar problemas sistémicos de inclusión e igualdad.

Muchos países han empezado a incorporar distintas formas de regulación, a través de la formulación de sus estrategias nacionales, aprobadas o en proyecto (como son los casos de Canadá[39], China, Colombia[40], Chile[41], Argentina[42], etc.).

La Unión Europea mantiene desde siempre un enfoque[43] enfrentado al modelo de desregulación norteamericano, fuertemente influenciado por los intereses de Silicon Valley. En febrero de 2020, la UE presentó el "White Paper on Artificial Intelligence: a European approach to excellence and trust"[44], cuyas bases generales apuntaban a construir un enfoque común para la UE, evitando regulaciones locales que obstruyeran un mercado regional. Asimismo, esboza la regulación basada en categorías y sectores de alto riesgo (como el reconocimiento facial), e impone requisitos para construir una IA confiable. En abril de 2021, se dieron a conocer las bases de una regulación para el bloque, a través de la propuesta de Reglamento "por el que se establecen normas armonizadas en materia de Inteligencia Artificial (Ley de Inteligencia Artificial)"[45].

Siguiendo el camino de declaraciones previas como la de Montreal y Toronto[46], recientemente, la UNESCO publicó la Recomendación sobre la Ética de la Inteligencia Artificial, adoptada por 193 países[47].

39. *https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592#appA*

40. *https://dapre.presidencia.gov.co/TD/CONSEJO-INTERNACIONAL-INTELIGENCIA-ARTIFICIAL-COLOMBIA.pdf*

41. *https://www.minciencia.gob.cl/legacy-files/borrador_politica_nacional_de_ia.pdf*

42. *https://ia-latam.com/portfolio/plan-nacional-de-ia-gobierno-de-argentina/*

43. *https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence*

44. "White Paper on Artificial Intelligence: a European approach to excellence and trust". Disponible en: *https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en*

45. REGLAMENTO DEL PARLAMENTO EUROPEO Y DEL CONSEJO POR EL QUE SE ESTABLECEN NORMAS ARMONIZADAS EN MATERIA DE INTELIGENCIA ARTIFICIAL (LEY DE INTELIGENCIA ARTIFICIAL) Y SE MODIFICAN DETERMINADOS ACTOS LEGISLATIVOS DE LA UNIÓN", COM/2021/206. Disponible en: *https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206*

46. *https://recherche.umontreal.ca/english/strategic-initiatives/montreal-declaration-for-a-responsible-ai/* y *https://www.torontodeclaration.org/declaration-text/english/*

47. *https://news.un.org/en/story/2021/11/1106612*

Regular la Inteligencia Artificial no es una tarea sencilla, por eso parte del objetivo del marco de dimensiones que aquí se propone es mostrar la diversidad de las múltiples facetas y áreas legales involucradas con sus pesos y contrapesos, desde los derechos humanos, la protección de datos, la propiedad intelectual, el derecho del consumo, el derecho de la competencia, etc. Cualquier regulación que no tenga en cuenta estos balances y tensiones entre lo individual y lo colectivo, entre la innovación y la protección, entre lo nacional y el ecosistema global, está destinada a ser un enfoque sesgado de la realidad de la Inteligencia Artificial.

En algunos aspectos, los daños que pueden causar las aplicaciones de Inteligencia Artificial tienen puntos en común con el derecho ambiental, de ahí que se propongan (para casos específicos de potencial riesgo) evaluaciones de impacto algorítmico, como paso previo a autorizar su despliegue en la sociedad.

En tanto los productos basados en aprendizaje automático continúan aprendiendo por retroalimentación, no es suficiente un examen previo a su despliegue en la vida real, si no va acompañado de un seguimiento continuo que permita monitorear su funcionamiento durante su vida útil. Es necesaria una **auditabilidad dinámica y sostenida en el tiempo,** durante todo el ciclo de vida de una aplicación basada en Inteligencia Artificial, y mientras esta pueda afectar la salud, la vida o la reputación de los seres humanos.

FAIRNESS, EQUIDAD, DIVERSIDAD E INCLUSIÓN

Los análisis que engloba esta última dimensión son el producto acumulativo resultante de las interacciones entre las anteriores, considerando el impacto de los sistemas de IA en la sociedad.

Cuando una aplicación de Inteligencia Artificial se basa en modelos sesgados, sus productos perpetúan estereotipos de discriminación, que no pueden ser advertidos o corregidos en tanto este es opaco y, por ende, no auditable. En la medida en que estos desarrollos algorítmicos se aplican en una escala masiva, impactan en la igualdad y la inclusión, con la potencialidad de ahondar las desigualdades digitales globales.

Aunque se suele hablar de ***fairness*** y de equidad como sinónimos, entiendo que estos conceptos tienen matices diferentes que deben ser distinguidos. Si pensamos en la forma en que funcionan los modelos de Inteligencia Artificial basados en aprendizaje automático, estos tratan de representar matemáticamente un problema en la forma más ajustada a las entradas y el entrenamiento provisto. Esta aspiración implica una intención de imparcialidad y neutralidad, que por un lado encubre el problema de los sesgos y, por el otro y por la forma en que funcionan los modelos de IA, no hace más que profundizar las inequidades existentes en realidad, escalando y amplificando la discriminación e injusticia engranada en los datos de los que aprende.

En el ***paper*** fundacional ***"On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?***", Timnit Gebru advierte de los peligros de los modelos de procesamiento de lenguaje natural basados en cantidades gargantuescas de datos, con datasets que no son curados sino que "ingieren" los contenidos disponibles en Internet.

En una de las tablas que ilustran la investigación, ejemplifican que entre el modelo BERT de 2019, que contaba con 3.4E+08 parámetros y un dataset de 16GB, se ha avanzando a modelos como GPT-3 (2020) con 1.75E+11 parámetros y un dataset de 570GB o Switch-C (2021), que ostenta 1.57E+ parámetros y un dataset masivo de 745 GB.

Frente al crecimiento significativo de los modelos tanto en cantidad de parámetros como en tamaño de los datasets se preguntan: ¿cuán grande es demasiado grande?, y fundamentalmente: ¿cómo esto impacta en el medio ambiente? Además advierten sobre que a medida que crece este tamaño descomunal de los modelos, "también lo hace la dificultad de entender lo que hay en los datos de entrenamiento. En §4, discutimos cómo los grandes conjuntos de datos basados en textos de Internet sobre-representan los puntos de vista hegemónicos y codifican sesgos potencialmente perjudiciales para poblaciones marginadas. Al recopilar conjuntos de datos cada vez más grandes, nos arriesgamos a contraer una deuda de documentación…".

El problema de la muestra se magnifica exponencialmente en los millones de parámetros que constituyen estos modelos, imponiendo una labor igualmente titánica para su auditabilidad, a fin de comprobar que se trata de una representación fiel de la realidad.

Cuando hablamos de ***fairness*** entiendo que hay que hacer una distinción clave sobre el término, y su traducción al español. Comprendiendo los matices semánticos que la palabra tiene en inglés aplicada al tema, entiendo que no existe un equivalente exacto en nuestro idioma que los pueda capturar con el mismo alcance y sentidos, sin incorporar la carga semántica que esas palabras ya traen en español.

En el contexto estadístico y matemático de cómo funcionan los modelos de aprendizaje automático, un modelo que es ***fair*** implica fidelidad y exactitud (***accuracy***) en la forma en que representa la realidad del problema que busca reflejar. Sin embargo, una representación exacta es un ideal imposible y, en la práctica, ***fairness*** se referirá a que el modelo captura adecuadamente el fenómeno sobre el que busca brindar respuestas. Sin embargo, sabemos que la realidad es discriminatoria y no queremos trasladar esas inequidades de base a los modelos de Inteligencia Artificial, de ahí que la **equidad** puede introducirse como una forma de mitigación que corrija esos valores para lograr la **inclusión**, de forma similar a las acciones afirmativas del Derecho para compensar inequidades de base entre categorías dispares.

Como advierte Divya Gopinath (2021), al construir "un modelo de decisión crediticia que determina si un individuo debe recibir un préstamo. El modelo se entrena con datos que no utilizan ninguna característica demográfica específica, como el género o la raza. Sin embargo, al evaluar los resultados a posteriori, se observa que el modelo encuentra una ligera correlación entre las tasas de aprobación y la raza. ¿Es este un resultado injusto, porque el modelo está construyendo proxies de la raza a través de otras variables, como el código postal o el apellido? ¿O se trata de un resultado justificado y ecuánime, porque el modelo está utilizando un conjunto razonable de características, y resulta que hay una correlación?". Asimismo, recordando el fiasco de las tarjetas de crédito de Apple[48], añade que un modelo se aplica a distintos grupos, por lo cual puede ser justo en su aplicación a un grupo de individuos, pero discriminatorio en relación con otros.

En este sentido, el concepto matemático se asemeja al legal de "igualdad ante la ley": todos los integrantes de una categoría en condiciones similares tienen (en palabras de la Corte Suprema de Justicia de Argentina) el derecho a recibir el mismo tratamiento legal, sin discriminaciones arbitrarias.

Como todo operador legal sabe, entre la enunciación formal del principio y la práctica hay un campo difuso de interpretación de sus alcances en relación con el caso concreto, que dependerá en definitiva (en el caso judicial) del Tribunal que tenga la última palabra sobre el asunto.

El ***fairness*** y la igualdad son conceptos con bordes elásticos y fronteras móviles, que pueden acomodar una u otra interpretación, según quién tenga la facultad de definirlas, ya que no existe un concepto unívoco, ni siquiera matemáticamente.

Como enfatizara al tratar la dimensión de sesgo, la importancia de la calidad de los datos que se consideran, eliminan o sustituyen frente a la falta de información tiene un correlato directo sobre la representatividad del modelo.

Cuando los modelos separan los datos en categorías ideales, elaboran representaciones extrayendo información de los datos de entrenamiento subyacentes. Aun cuando los datos considerados no sean discriminatorios ***per se***, pueden funcionar como un proxy y estar representando una correlación que sí lo es, de ahí la importante conexión entre esta dimensión y lo explicado en cuanto al sesgo (en los datos, en los modelos, en los

48. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

procesos o en la interpretación).

En muchos casos habrá una tensión inherente irresoluble entre ***fairness*** e inclusión, en la medida en que una representación fidedigna estará plagada de los sesgos e inequidades de base existentes en la sociedad, que solo puede ser solucionada a través de estrategias de mitigación que intervengan el modelo hacia una versión más equitativa e inclusiva, pero alejándose de la fidelidad inicial.

En algún punto, se trata de una negociación entre representatividad matemática y justicia, un modelo de la realidad en forma justa (***fairness***) o modificando proactivamente las inequidades existentes en la realidad a fin de que no se trasladen a lo digital (inclusión). Eminentemente, se trata de una decisión difícil de ética y política, que no puede dejarse para ser resuelta en los intersticios de construcción y de iteración de los modelos. Además de ello, requerirá ser observada en el tiempo, para ver si las postulaciones iniciales se continúan correspondiendo en su aplicación en la vida real.

Aquí es donde las evaluaciones de impacto algorítmico presentan potencial, no solo previo a su implementación con una proyección de potenciales ramificaciones, sino con la imposición de un seguimiento continuo una vez que están implementadas.

Como también se adelantó al explicar sobre los sesgos, la **diversidad e interdisciplinariedad** de los equipos que desarrollan Inteligencia Artificial es un componente necesario para mitigarlos, y acompañar la construcción de estas tecnologías en forma ética.

Las aplicaciones de IA tienen la potencialidad de profundizar la brecha tecnológica producida por la disparidad en el conocimiento y capacitación, creando una élite tecnológica capaz de relacionarse en mejores términos con esta tecnología y beneficiarse económicamente de ella. Como señala Shadoan (2014), "Los algoritmos opacos no solo son malos por su potencial de abuso y por nuestra capacidad de razonar. También son malos para la igualdad. Fomentan la creación de una clase élite, aquellos que, por privilegio de educación o experimentación, han desarrollado mejores modelos mentales para describir el funcionamiento interno de los algoritmos opacos. Esa clase de élite puede aprovechar su mejor comprensión del algoritmo para razonar mejor sobre sus propios pensamientos, lo que a su vez les permite refinar sus modelos, y así sucesivamente. Las personas que no tienen el privilegio de una comprensión sólida de los algoritmos opacos son menos capaces de aprovechar esos algoritmos para acceder a la información o entender el contexto de la información a la que se accede. Esto, a su vez, obstaculiza su capacidad para razonar sobre la calidad de la información que reciben y los pensamientos resultantes que generan sus cerebros. El ciclo se auto-refuerza y nace una élite tecnológica; separando aquellos que entienden los algoritmos, y aquellos que no…".

Así como el hombre es sus circunstancias, los algoritmos y la Inteligencia Artificial son una construcción en la que se engranan en sus mecanismos, la diversidad de enfoques, saberes, valores y miradas sobre el mundo de quienes desarrollan esa tecnología. Sin diversidad, ética e inclusión, estamos condenados a trasladar las inequidades del mundo analógico a lo digital.

CONCLUSIONES: ÉTICA Y EL IMPACTO GLOBAL DE LA IA EN LA SOCIEDAD

Este capítulo podría transformarse en un libro por sí mismo, solamente enumerando los ejemplos en los últimos años de cómo el empleo irrestricto y no auditado de la Inteligencia Artificial ha provocado situaciones de inequidad, discriminación y exclusión.

La Inteligencia Artificial tiene un impacto disruptivo horizontal, que atraviesa todos los sectores de la industria, la política pública y la sociedad, y que se ve acelerado por la transformación digital.

En el estado actual del arte, evaluar la dimensión ética es una tarea difícil para las arquitecturas actuales de ***machine learning***: los algoritmos no pueden realizar automáticamente los ajustes necesarios para determinar la justicia de una situación. Es técnicamente complejo y, para algunos autores, un punto de necesaria intervención humana en los procesos (O'Neil, 2016).

Los algoritmos pueden aprender sobre la estadística empírica, o ser programados para determinadas respuestas, pero al menos por ahora no pueden realizar valoraciones axiológicas. Entre otras muestras, los ejemplos del bot ***Tay*** que se volvió racista y misógino a horas de su lanzamiento en Twitter (2016), y del más reciente experimento de Delphi lanzado en octubre de 2021, demuestran todo el camino que falta recorrer en cuanto a moral y axiología. El caso de Delphi muestra claramente cómo los sesgos adheridos al lenguaje en los datos de entrenamiento tiñen el producto final[49] y la falta real de sentido común de la IA más allá de sus asunciones estadísticas, además de la complejidad semántica del discurso, en la que bastan unas pequeñas sustituciones en la entrada para manipular el resultado final.

Luego de varios titulares destacando afirmaciones racistas[50], el acceso al sitio incluye ahora un ***disclaimer***: "***Delphi es un prototipo de investigación diseñado para investigar las promesas y, lo que es más importante, las limitaciones de modelar los juicios morales de las personas en una variedad de situa-***

49. https://delphi.allenai.org/ Delphi es un experimento de IA que emplea un modelo de lenguaje natural de gran tamaño (como aquellos sobre los que previene Timnit Gebru en Bender et al., 2021), para emitir juicios éticos sobre sentencias cortas ingresadas por los usuarios. Como otros modelos de esta naturaleza, aprende haciendo estadística sobre los textos con los que fue entrenado, y sufre el problema común a este tipo de modelos: traslada los sesgos del lenguaje escondidos en los datos de entrenamiento y carece de sentido común real, más allá de la simulación que pone en escena. El ***paper*** en que se basa está disponible en: https://arxiv.org/pdf/2110.07574.pdf

50. "***Delphi llama la atención sobre todo por sus numerosos errores morales y sus extraños juicios. Tiene claros prejuicios, diciendo que Estados Unidos es 'bueno' y que Somalía es 'peligrosa'. Preocupantemente, aprueba declaraciones directamente racistas y homófobas, diciendo que es 'bueno' 'asegurar la existencia de nuestro pueblo y un futuro para los niños blancos' (un lema supremacista blanco conocido como las 14 palabras) y que 'ser heterosexual es más aceptable moralmente que ser gay'***". https://www.theverge.com/2021/10/20/22734215/ai-ask-delphi-moral-ethical-judgement-demo

ciones cotidianas. El objetivo de Delphi es ayudar a los sistemas de IA a estar más informados desde el punto de vista ético y ser más conscientes de la equidad. Al dar un paso en esta dirección, esperamos inspirar a nuestra comunidad investigadora para que afronte los retos de la investigación en este ámbito con el fin de construir sistemas de IA éticos, confiables e inclusivos."

Por otra parte, más allá de sus pretensiones de universalidad moral, Delphi fue entrenada utilizando trabajadores a los que se les indicó que debían responder de acuerdo con lo que creen que son las normas morales de los Estados Unidos.

Recordemos que diariamente los algoritmos y aplicaciones de IA toman decisiones sobre el mundo, sin poseer un compás moral y ético, ni sentido común, ni nociones de localización en cuanto al contexto geográfico y social en que se aplican.

Frente a la inminente implementación de vehículos automáticos no tripulados y la escalada en fuerzas de seguridad por armas autónomas, las preguntas éticas dejan de ser hipotéticas para necesitar ser embebidas en el software de estos robots. ¿Quién establece las reglas morales ante situaciones de decisión sobre la vida? ¿Podrán ser ajustadas por el propietario del vehículo? ¿Vendrán preestablecidas de fábrica y las aceptaremos como "términos y condiciones"?

Considerando este panorama, el MIT desarrolló otro experimento vinculado a la ética en IA, a través de una plataforma para recolectar opiniones sobre distintos posibles escenarios con dilemas morales presentados al usuario, quien en forma similar al "dilema del tranvía", debe tomar decisiones sobre la vida de hipotéticos transeúntes y pasajeros de vehículos autónomos.

Para tratar de construir un marco ético en forma colectiva, el Moral Machine del MIT[51] propone escenarios de dilemas morales donde se debe elegir lo que se considera el menor de dos males, como matar a dos pasajeros o a cinco peatones, una vida joven o anciana, humana o animal.

Ingresar en la problemática de la corporización de las inteligencias artificiales en robots que les permitan interactuar con el mundo, abre una avenida inmensa de problemas morales y oportunidades para un desarrollo ético e igualitario. Al respecto, algunos autores consideran aplicables las ideas del ***moral agency*** como pauta de discernimiento, que implica evaluar si los entes pueden percibir las consecuencias moralmente pertinentes de sus acciones, y ser capaces, en función de ello, de elegir los cursos relevantes de acción (Kaplan, 2015).

Propugnando la incorporación de valores éticos desde el diseño, el Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) fundó la Iniciativa Global sobre Ética de Sistemas Autónomos e Inteligentes. El objetivo del Diseño Alineado Éticamente es promover una discusión pública sobre cómo "establecer implementaciones éticas y sociales para sistemas y tecnologías inteligentes y autónomas, alineándolos con valores definidos y principios éticos que priorizan el bienestar humano en un contexto cultural dado"[52].

51. Disponible en: ***https://www.moralmachine.net/***

52. "Ethically Aligned Design, Version 2 (EADv2) | IEEE Standards Association", n.d.

La ética debe estar incorporada desde el diseño y ser acompañada a lo largo de todo el proceso de desarrollo, a fin de no imponer una visión unívoca del mundo, centrada en la perspectiva individual de los creadores de la tecnología.

Bajo el pretexto de la automatización y el confort, algunos desarrollos realmente limitan las opciones y, por ende, el rango de visión de los usuarios. Al respecto, "En un artículo reciente, Kerr y Earle llaman la atención sobre la ética de los algoritmos como procesos predictivos, comparando la predicción consecuente -en la que un algoritmo muestra resultados posibles para usar en la toma de decisiones de una persona- con la predicción preferencial o preventiva. En la predicción preferencial, un proceso anticipa los deseos de un usuario y ofrece opciones que pueden agradar, cuando en la predicción preventiva, un algoritmo delimita el acceso de una persona sin proporcionar una opción, y a menudo sin el conocimiento de la persona. Kerr y Earle llaman especialmente la atención sobre los peligros de la opacidad en los algoritmos preventivos, lo que indica su potencial para una aplicación injusta. Mirando la historia, uno podría recordar aquí la dependencia de la Autoridad Federal de Vivienda en la década de 1930 en los procesos algorítmicos de calificación de los posibles beneficiarios de préstamos en función de la composición racial de sus barrios de origen" (Hamilton et al., 2014)[53].

La falta de ética en el uso de la Inteligencia Artificial es particularmente crucial en lo que hace al futuro del trabajo, con distintas variantes que van desde la **sustitución de empleos por máquinas, la hipervigilancia de los trabajadores,** las barreras de acceso al empleo a través de **revisión automatizada de postulaciones**, y episodios de **falsa automatización**, en los que la labor se desplaza e invisibiliza.

A diferencia de lo que ocurrió en la revolución industrial, la automatización actual tiene un impacto severo sobre el empleo no solo en las labores manuales repetitivas, sino en todo un rango de profesiones liberales y tareas intelectuales. Ello implica, por un lado, un necesario reajuste en las currículas educativas, para adaptarse a las nuevas necesidades del mercado y, por otro, la creación de mecanismos de contención social y soporte financiero para aquellos a quienes la economía digital dejará de lado.

En un informe de 2016[54], el Foro Económico Mundial había predicho que para 2020 la automatización daría como resultado la pérdida neta de más de 5 millones de empleos en 15 países desarrollados. Otro estudio, realizado por la Organización Internacional del Trabajo, indicaba que hasta 137 millones de trabajadores en Camboya, Indonesia, Filipinas, Tailandia y Vietnam (que representan aproximadamente el 56% del total de la fuerza de trabajo de estos países) estaban en peligro de ser reemplazados por robots, especialmente aquellos que están empleados en la industria de fabricación de prendas de vestir (Shewan, 2017).

Ninguna de estas predicciones podía prever la pandemia que asolaría al mundo en 2020, y el efecto devastador que esta tendría sobre el empleo. Frente al aislamiento obligatorio y a las limitaciones del contacto, la tendencia a reemplazar a los seres humanos por aplicaciones automatizadas y robots se aceleró vertiginosa-

53 Traducción propia de: "*In a recent paper Kerr and Earle call special attention to the ethics of algorithms as predictive processes, comparing consequential prediction – in which an algorithm displays possible outcomes for use in a person's decision-making – to preferential or preemptive prediction. In preferential prediction, a process anticipates a user's desires and offers options likely to please, whereas in preemptive prediction, an algorithm delimits a person's access without providing a choice, and often without the person's knowledge. Kerr and Earle call special attention to the dangers of opacity in preemptive algorithms, pointing to their potential for unjust application. Looking to history, one might recall here the Federal Housing Authority's reliance in the 1930s on algorithmic processes of rating potential loan recipients based on the racial composition of their home neighborhoods.*" (Hamilton et al., 2014)

54, "The Future of Jobs", World Economic Forum. Disponible en: *https://www.weforum.org/reports/the-future-of-jobs*

mente, en tanto las empresas luchan por evitar las infecciones y por mantener bajos los costos en una economía en retroceso.

Según el informe actualizado sobre el futuro del trabajo del año 2020 del Foro Económico Mundial, se prevé que para 2025 la automatización suplantará a 85 millones de puestos de trabajo, cuestión que ha sido acelerada por el contexto creado por la pandemia. Si bien estima que se crearán 97 millones de nuevos puestos, el informe marca una desaceleración en la creación de nuevos puestos en relación con el ritmo en que se pierden. Además, estima que para ese año las horas de trabajo se repartirán 50/50 entre máquinas, algoritmos y humanos. Frente a este panorama, se torna imprescindible generar programas de recapacitación y de aprendizaje de nuevas habilidades digitales[55].

La disponibilidad escalable de aplicaciones automatizadas trae consigo un incremento en la vigilancia laboral, con consecuencias desastrosas para la salud mental de los trabajadores. Los empleados de Amazon son el caso paradigmático de esta pesadilla orwelliana: sistemas monitoreando su performance constantemente, algoritmos asignando tareas y turnos sin supervisión humana, e incluso despidiendo a los trabajadores en forma automática[56]. Nuevamente, la automatización castiga más fuertemente a los sectores más vulnerables, por la precariedad y reemplazabilidad de sus empleos. En la ***gig economy***, donde se repiten las asimetrías del derecho del consumo pero a escala de empleo, el uso de softwares de automatización se traduce en una burocracia digital digna de ***Brazil***[57], frente a la que los trabajadores no tienen más opción que aceptar invasiones como el reconocimiento facial[58]. En el Reino Unido dos sindicatos -App Drivers & Couriers Union (ADCU) y Worker Info Exchange (WIE)- pidieron que Uber suspenda el uso después de encontrar varios casos en los que los conductores fueron identificados erróneamente, provocando que los conductores perdieran sus trabajos.

En un artículo reciente publicado por ***Bloomberg***[59] se cuenta la historia de Stephen Normandin, un veterano de guerra que trabajó cuatro años entregando paquetes para Amazon, y que fue sorprendido con la noticia de su despido en forma automática por un algoritmo: "Esto me molestó mucho porque estamos hablando de mi reputación. Dicen que no hice el trabajo cuando sé muy bien que lo hice". Amazon emplea desde hace años distintas aplicaciones de IA y algoritmos para coordinar sus operaciones de venta y logística de entrega. Muchos vendedores han reportado ser expulsados tras ser acusados falsamente de vender productos falsificados o de aumentar los precios. Cada vez más, la empresa está cediendo sus operaciones de recursos humanos a las máquinas, utilizando software no sólo para gestionar a los trabajadores de sus almacenes, sino también para supervisar a los conductores contratados, a las empresas de reparto independientes e incluso el rendimiento

55. "Aunque el número de puestos de trabajo destruidos será superado por el número de 'empleos del futuro' creados, a diferencia de años anteriores, la creación de empleo se está ralentizando mientras la destrucción de puestos de trabajo se acelera. [...] Calculamos que, de aquí a 2025, 85 millones de puestos de trabajo podrían verse desplazados por un cambio en la división del trabajo entre humanos y máquinas, mientras que podrían surgir 97 millones de nuevas funciones más adaptadas a la nueva división del trabajo entre humanos, máquinas y algoritmos." [...] "El 43% de las empresas encuestadas indican que van a reducir su personal debido a la integración de la tecnología, el 41% tiene previsto ampliar el uso de contratistas para trabajos especializados y el 34% tiene previsto ampliar su personal debido a la integración de la tecnología. En 2025, el tiempo que dedican a las tareas actuales en el trabajo los humanos y las máquinas será igual." Disponible en: ***https://www.weforum.org/reports/the-future-of-jobs-report-2020***

56. ***https://www.theguardian.com/commentisfree/2021/jul/05/amazon-worker-fired-app-dystopia***

57. Me refiero a la película (no al país), donde una equivocación en el nombre de una persona (Tuttle) genera consecuencias imposibles de revertir en una maraña de burocracia caótica y sin sentido común.

58. ***https://techcrunch.com/2021/03/19/uber-under-pressure-over-facial-recognition-checks-for-drivers/***

59. ***https://news.bloomberglaw.com/daily-labor-report/fired-by-bot-at-amazon-its-you-against-the-machine***

.de sus trabajadores de oficina. Jeff Bezos cree que las máquinas toman decisiones con mayor rapidez y precisión que las personas, lo que reduce los costes y da a Amazon una ventaja competitiva. ***Bloomberg*** entrevistó a 15 conductores de Flex, un servicio tercerizado de entregas de Amazon, así como a antiguos directivos de Amazon. Estos afirman que el sistema, en gran medida automatizado, no está suficientemente adaptado a los retos del mundo real a los que se enfrentan los conductores cada día. Aún más, indican que Amazon sabía que delegar el trabajo en las máquinas daría lugar a errores, pero que decidió que era más barato confiar en los algoritmos que pagar a personas para que investigarán los despidos erróneos, siempre y cuando los conductores pudieran ser reemplazados fácilmente.

En lo que parece sacado de un futuro distópico de series como ***Black Mirror***, la empresa Canon incorporó en sus cámaras de reconocimiento facial una tecnología para el reconocimiento de sonrisas, una aplicación que se basa en empleados de China que deben sonreír para ingresar a su trabajo y a salas de conferencia[60].

Impulsado también por la pandemia, la tendencia creciente al trabajo remoto o híbrido **traslada la vigilancia en el lugar de trabajo a los hogares**, con fronteras borrosas entre la vida laboral y hogareña, e invasión a la privacidad de los restantes habitantes del lugar.

Algunos empleadores "...controlan los sitios web visitados por los empleados, o el tiempo que pasan en los sitios web o en los documentos. Un servicio de supervisión del trabajo a distancia ofrece un amplio control de los teléfonos móviles, incluido el seguimiento del Sistema de Posicionamiento Global (GPS). Una función común a muchos servicios permite al empleador ver remotamente las pantallas de los ordenadores de los empleados. Otras funciones incluyen la visualización del correo electrónico entrante y saliente, y la grabación de audio. Los empleadores pueden supervisar las pulsaciones del teclado, la actividad en las redes sociales y las búsquedas en la web. Un servicio proporciona vídeo permanente a través de los ordenadores de los empleados, lo que permite a su empleador y a otros trabajadores conectarse con ellos con un solo clic, y mantener sus rostros visibles en la pantalla durante todo el día. Todos estos datos pueden ser grabados y almacenados. El software de monitorización también puede venir con análisis de comportamiento y otras herramientas algorítmicas para calificar o evaluar la actividad y la productividad de los empleados. Algunos sostienen que estas herramientas serán inestimables para aumentar la productividad en la llamada economía de precisión…" (Scassa, 2021).

Un informe de Express VPN reveló que los trabajadores no son conscientes del nivel de vigilancia al que están sometidos[61] . Entre las actividades que los softwares espías recopilan se encuentran:

60. *https://www.theverge.com/2021/6/17/22538160/ai-camera-smile-recognition-office-workers-china-canon*

61 "Existe una brecha significativa entre lo que los empleadores vigilan activamente y lo que los empleados creen que sus empleadores vigilan, lo que indica que muchos desconocen lamentablemente su verdadero alcance, que podría estar poniéndolos en peligro." Disponible en: ***https://www.expressvpn.com/blog/expressvpn-survey-surveillance-on-the-remote-workforce/***

Dispositivos y funciones de vigilancia usadas por empleadores	
Sitios web visitados / tiempo dedicado a varios sitios web	**66%**
Aplicaciones utilizadas / tiempo dedicado a las aplicaciones	**53%**
Vigilancia de la pantalla en tiempo real	**53%**
Captura de pantalla periódica	**33%**
Horas de trabajo activas / tiempos de registro	**46%**
Horas productivas vs. improductivas registradas	**31%**
Chats / registros de mensajería	**30%**
Vigilancia / acceso a archivos informáticos	**27%**
Monitoreo de correos electrónicos entrantes y salientes	**23%**
Transcripción de llamadas	**22%**

ExpressVPN survey on the extent of surveillance on the remote workforce, 20 de mayo 20 de 2021

Entre los hallazgos significativos, los trabajadores encuestados manifestaron sentimientos de ansiedad, estrés y ruptura de la confianza, así como deshumanización del entorno de trabajo.

Frente al desempleo y la vigilancia, otro engranaje perverso de las aplicaciones de IA en el trabajo es la utilización de **algoritmos para revisar las postulaciones de aplicantes** a puestos laborales. Desde revisión automatizada de currículums en busca de palabras clave, análisis de entrevistas en videos de candidatos, empleo de procesamiento de lenguaje natural para analizar el discurso de los aplicantes, reconocimiento facial de gestos y postura, etc.[62]. La falta de regulación y de marcos éticos accionables en relación con aplicaciones de IA generan un campo fértil para que la comercialización de la Inteligencia Artificial se transforme en un negocio de

62. https://www.technologyreview.com/2021/07/07/1027916/we-tested-ai-interview-tools/

venta de confianza, agravado por la falta de transparencia de estos sistemas. Los adquirentes de la tecnología y, sobre todo, quienes reciben sus efectos, deben confiar en los procesos internos y en la autorregulación de quienes la desarrollan.

Paradójicamente, cuando la diversidad, equidad e inclusión en el trabajo comienza a tomar un rol preponderante en la cultura organizacional, y una creciente conciencia del potencial desaprovechado de contrataciones no tradicionales[63], la utilización de softwares de selección automatizada de postulantes tiende a repetir los patrones de contratación históricos, replicando los sesgos sociales existentes. En 2018, Amazon tuvo que abandonar el empleo de un software que discriminaba contra mujeres, al punto tal que la mención de colegios con nombres femeninos en las postulaciones de hombres, provocaba que la inteligencia artificial bajara el puntaje del candidato[64].

La explicación lógica es sencilla de apreciar cuando se conoce la forma en que opera el aprendizaje automático: los modelos aprenden estadísticamente por ejemplos del pasado, para tratar de predecir los nuevos casos del futuro. Los modelos de Amazon fueron entrenados con los datos de contratación y los currículums vitae enviados a la empresa durante un período de 10 años. Dado el predominio masculino en la composición de la fuerza de trabajo en el sector tecnológico, los algoritmos aprendieron del ejemplo, impulsando la fórmula del pasado éxito: contratar hombres.

Estos fallos iniciales no frenaron la expansión del sector, y hoy se ha vuelto una práctica estandarizada empleada por todo tipo de empresas que tercerizan el servicio hacia startups de IA, problema que mostráramos al hablar de la dimensión de escalabilidad.

En 2018, los directivos de HireVue manifestaban que el algoritmo analizaba "el discurso y las expresiones faciales de los candidatos en las entrevistas en video para reducir la dependencia de los currículos"[65], mientras que actualmente la plataforma dedica varias secciones con palabras cuidadosamente elegidas para explicar el funcionamiento, la ciencia detrás de la fórmula y la ética de sus modelos. Sin embargo, a poco que se indague, surgen afirmaciones ambiguas como que se "entrena el modelo para que se fije en todo lo que es relevante en la entrevista (lo que alguien dice y cómo lo dice), y construye un modelo que utiliza solo los puntos de datos que ayudan a predecir el éxito en el trabajo…"[66].

Por su parte Modern Hire afirma que su modelo propietario CognitIOn "evalúa únicamente la información que los candidatos proporcionan con pleno conocimiento de causa, explican a los candidatos cómo se evalúan y utilizan sus datos y no utilizan prácticas que se sabe que son poco confiables, potencialmente injustas o invasivas, como el uso de IA para evaluar los rasgos faciales o el análisis de perfiles en las redes sociales"[67].

63. Ver sobre esto: Fuller et al., 2021.

64 .https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

65. Vide la entrevista de Reuters, citada previamente. .

66. https://www.hirevue.com/why-hirevue/ai-ethics

67. https://modernhire.com/platform/automated-interview-scoring/

Distintos informes proyectan el valor global del segmento de softwares online de contratación en "29292,3 millones de dólares en 2021 y se prevé que alcance los 47315,0 millones de dólares en 2028; se espera que crezca a una CAGR del 7,1% de 2021 a 2028…"[68] o que "el tamaño del mercado de la Inteligencia Artificial (IA) en la contratación se valoró en 580 millones de dólares en 2019, y se estima que crecerá a una CAGR del 6,76% durante 2020-2025…"[69]. Frente a esas cifras que marcan la avidez por este tipo de iniciativas, es interesante pensar cómo algunas empresas podrían dominar la oferta y replicar, a través de todos quienes contraten sus servicios, los sesgos que el software de IA pudiera contener, como viéramos en el ejemplo de la Algorithmic Justice League al tratar la dimensión de escalabilidad.

Otra faceta del futuro del trabajo de la que no se suele hablar es el impacto de la Inteligencia Artificial en la **precarización de trabajadores invisibles,** e **instancias de falsa automatización**.

Cuando hablamos de falsa automatización, nos referimos a los procesos en donde un análisis superficial sobre el trabajo marca que se realiza de forma digital o automatizada, pero al indagar encontramos seres humanos en la cadena y, fundamentalmente, desplazamientos de labor**.** Por ejemplo, a través de incorporar Captchas en sitios web para comprobar que "no soy un robot", la tarea de identificación de imágenes (calles, bicicletas, señales de tráfico, etc.) que puede ser dificultosa para las máquinas, se realiza en forma descentralizada (y gratuita) por seres humanos alrededor del mundo.

Son casos en los que la labor que antes hacía un empleado pago se traslada al propio consumidor, disfrazado -irónicamente- de una digitalización en su beneficio. Es el caso, por ejemplo, de los supermercados que instalan cajas de autocobro, de las cadenas de comida rápida con puestos digitales de autoservicio, o de los servicios de venta online, donde las aerolíneas, hoteles, cines, etc. realizan sus ventas web empleando el tiempo y el trabajo individual de sus consumidores.

En el juego de espejos y humo, otra variante de las caras ocultas de la Inteligencia Artificial es la **invisibilización de trabajadores humanos**, en donde los resultados presentados como automatizados son producto de labores concretas de trabajadores humanos en la cadena. Detrás de determinadas tareas que siguen siendo difíciles para las máquinas, como las definiciones semánticas de contexto o sentido común, se esconde la labor de intérpretes humanos que trabajan tomando microdecisiones, muchas veces con meros segundos para pasar de una tarea a otra. Por ejemplo, si una aplicación requiere que el usuario se someta a reconocimiento facial previo a acceder, los casos dudosos pueden ser girados instantáneamente a moderadores humanos que revisan la identidad de la persona contrastándola con fotos previas.

Dadas la dificultad, la contextualidad y los matices involucrados en la interpretación de la comunicación, la **moderación de contenidos online** es un campo propicio para trabajadores humanos invisibles. Recientemente, una ex moderadora de contenidos de TikTok presentó una demanda contra la plataforma, basándose en que la empresa matriz ByteDance no brinda garantías adecuadas para proteger la salud mental de los moderadores, frente a una avalancha casi constante de imágenes traumáticas. La demanda colectiva, presentada ante tribunales de California, alega que la accionante, Candie Frazier, soportaba jornadas de doce horas, moderando

68. *https://www.theinsightpartners.com/reports/online-recruitment-market*

69. *https://www.industryarc.com/Report/19231/artificial-intelligence-in-recruitment-market.html*

videos que eran subidos a TikTok, siendo subcontratada por una compañía llamada Telus International. Frazier dice que fue testigo de "...miles de actos de violencia extrema y gráfica, incluidos tiroteos masivos, violación de niños, mutilación de animales, canibalismo, homicidio en pandillas y genocidio…", y que para hacer frente al enorme volumen de contenidos que se suben a TikTok a diario, ella y sus compañeros moderadores "tenían que ver entre tres y diez vídeos simultáneamente, con nuevos vídeos cargados al menos cada 25 segundos". Los moderadores solo pueden tomarse un descanso de 15 minutos en las primeras cuatro horas de su turno, y luego descansos adicionales de 15 minutos cada dos horas. La demanda dice que ByteDance supervisa estrechamente el rendimiento y "castiga fuertemente cualquier tiempo que se quite de ver vídeos gráficos"[70].

En ***Behind the Screen: Content Moderation in the Shadows of Social Media,*** Sarah T. Roberts (2019) advierte tempranamente sobre esta tristísima realidad detrás de la circulación de contenidos en redes sociales.

Por otro lado, tareas tediosas y que insumen mucho tiempo como la rotulación de imágenes o la preparación de información para datasets, son delegadas por las empresas en trabajadores tercerizados, a disposición según requerimiento, evitando ampliar su fuerza laboral en forma permanente.

En su libro, ***Ghost work,*** Mary L. Gray y Siddharth Suri (2019) investigan las redes subterráneas de la fuerza de trabajo que alimentan la maquinaria de la Inteligencia Artificial. Son "trabajadores fantasma" cuya labor se contrata en forma tercerizada (y altamente precarizada) a través de plataformas de microtareas como Mechanical Turk de Amazon[71]. La investigación revela la realidad que viven familias en situaciones de extrema necesidad, que recurren a estas formas irregulares de empleo, en las que las condiciones se asemejan a contratos de adhesión leoninos del derecho del consumo. En condiciones que desafían en la práctica los derechos laborales conquistados históricamente, la paga, la permanencia o la calificación en la plataforma de estos trabajadores está a merced de algoritmos y suspensiones erráticas o arbitrarias, basadas en términos y condiciones ambiguos, cuya interpretación o resolución no les es comunicada, y mayormente sin recursos administrativos para solicitar su revisión.

Asimismo, no hay que pasar por alto que quienes se encuentran en poder de los medios de producción automatizados serán capaces de incrementar su riqueza a través del trabajo de estos, acentuando la polarización económica y la brecha de pobreza.

Una forma de lidiar con esto sería recurrir a un **ingreso universal general** subsidiado gubernamentalmente (Darrow, 2017), generar incentivos (por ejemplo, deducciones fiscales) (Kaplan, 2016) por donaciones para las contribuciones a fondos sociales, o imponer tasas al trabajo robótico.

En esta revolución industrial, la cuantificación de las tareas **no se limita a la vigilancia de lo mecánico, sino que apunta a medir aspectos intangibles** como la atención, la concentración o el esfuerzo, a través de métodos masificados que no responden a las particularidades de los procesos de cognición de los seres humanos como individuos.

---

70 https://www.theverge.com/2021/12/24/22852817/tiktok-content-moderation-lawsuit-candie-frazier

71 *"Amazon Mechanical Turk (MTurk) es un mercado de crowdsourcing que facilita a las personas y empresas la subcontratación de sus procesos y trabajos a una mano de obra distribuida que puede realizar estas tareas de forma virtual. Esto podría incluir cualquier cosa, desde la realización de una simple validación de datos e investigación hasta tareas más subjetivas como la participación en encuestas, la moderación de contenidos, etc."* https://www.mturk.com/

La **cuarta revolución industrial es una industrialización del intelecto,** que abarca profesiones liberales que anteriormente se sentían a salvo de sus efectos. ***The Washington Post*** publicó un informe sobre el uso de software de monitoreo de tareas en abogados que trabajan de forma remota. Algunos de los entrevistados manifestaban que "El software utiliza la cámara web del trabajador para grabar sus gestos faciales y su entorno y enviará una alerta si el abogado toma fotos de documentos confidenciales, deja de prestar atención a la pantalla o permite la entrada de personas no autorizadas al cuarto. Se espera que los abogados escaneen su rostro cada mañana, de modo que su identidad pueda ser verificada minuto a minuto para reducir posibles fraudes…". Si existe alguna falla o violación de las reglas, el sistema desconecta a los abogados, que deben reiniciar el proceso de reconocimiento facial. Los usuarios reportaron que no pueden tomar algo mientras trabajan, porque el sistema confunde la taza con una cámara; o desviar la mirada para evitar bajas de puntaje de performance. En el caso informado de una usuaria de color, el sistema "...a menudo no reconocía su cara o confundía los nudos bantúes de su pelo con dispositivos de grabación no autorizados, lo que la obligaba a volver a conectarse a veces más de 25 veces al día…"[72].

Esto es un fenómeno que se repite en distintos ámbitos de aplicación de IA, los errores del software se incrementan particularmente en personas de color, que sufren constantemente el sesgo en la construcción de los sistemas de automatización y la falta de diversidad de datos de entrenamiento, así como en los equipos de desarrollo.

Lamentablemente abundan los ejemplos de la falta de reflejo de la diversidad cultural y de la imposición sesgada de un estándar cultural hegemónico. ***Desafabo Social,*** una ONG brasileña, desarrolló una poderosa campaña contra los algoritmos de Shutterstock, Getty Images y IStock demostrando la invisibilización de la población de color en la búsqueda de términos como familia, bebé o mujer. En las imágenes solo había representación de color si se agregaba "***black***" a la palabra buscada, de lo contrario, las imágenes representaban a bebés, mujeres y familias caucásicas[73].

Además de la falta de representatividad, los sistemas de IA generalmente performan mejor en personas de piel blanca[74], lo que se traslada a fallas e inequidades que ponen a sus usuarios en diferentes posiciones. Los de piel blanca reciben diagnósticos más acertados[75], un mejor posicionamiento a la hora de un trasplante[76],

72 "Algunos abogados contratados dijeron que sentían que la carga pesaba especialmente sobre las personas de color, que ocupan una parte enorme de los puestos legales a tiempo parcial. Las personas de color representan alrededor del 15% de todos los abogados en Estados Unidos, pero cerca del 25% de los puestos de trabajo de "abogados no tradicionales", que incluyen a los abogados contratados, según estadísticas recientes de la American Bar Association y la National Association for Law Placement.

A los abogados de color también les preocupaba que el diferente rendimiento de los sistemas de reconocimiento facial con los distintos tonos de piel les dejara en desventaja desde el principio. Un abogado dijo que presentó una queja ante la Comisión de Derechos Humanos de la ciudad de Nueva York el año pasado, argumentando que se le negaba el derecho a trabajar al negarse a dar su consentimiento para ser monitoreado. Le preocupa que los escaneos de reconocimiento facial puedan poner en peligro su licencia legal o su medio de vida si se le acusa falsamente de haber puesto en peligro los datos de sus clientes…". En "Contract lawyers face a growing invasion of surveillance programs that monitor their work", en ***The Washington Post*** (2021). Disponible en: ***https://www.washingtonpost.com/technology/2021/11/11/lawyer-facial-recognition-monitoring/***

73 ***https://desabafosocial.com.br/blog/2017/06/12/desabafo-social-interfere-no-mecanismo-de-busca-do-maior-banco-de-imagem-do-mundo/***

74 . Twitter realizó una competencia para auditar su algoritmo, encontrando sesgos en la forma en que este recorta las fotos dependiendo del color de la piel de las personas. ***https://www.theverge.com/2021/8/10/22617972/twitter-photo-cropping-algorithm-ai-bias-bug-bounty-results***

75 ***https://www.nature.com/articles/d41586-019-03228-6***

76 ***https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients/***

una mejor evaluación de su performance escolar[77], una evaluación diferencial de su riesgo de reincidencia[78], mejores tasas de reconocimiento facial y menores riesgos de identificaciones erróneas[79], entre una miríada de diferencias que profundizan la ya existente **brecha sistémica de inequidad e injusticia racial**.

La ciencia detrás de dispositivos de Inteligencia Artificial que prometen medir la atención o el aprendizaje es cuanto menos dudosa y recuerda fuertemente las **promesas lombrosianas[80].** Reducir la experiencia humana a un análisis de expresiones y gestos estandarizados, sin tener en cuenta contexto ni cultura es relativista y peligroso. Promete objetivizar la complejidad de lo que significa ser humano y termina siendo un juego de asunciones automatizadas sobre lo que ocurre en la psique, basados en estadística, con todas las falencias que apuntáramos a lo largo de las dimensiones de análisis que vimos previamente. ¿Cómo probar que lo que ocurre en nuestra mente no es lo que el software de IA afirma que ocurre? El salto entre los casos de uso es una escalera resbaladiza: donde hoy se pretende inferir en los procesos mentales de aprendizaje, mañana se ampliará (siempre bajo promesas de seguridad y pretensiones de eficiencia) a indagar sobre las intenciones criminales.

En este punto el lector quizás pensará "pero… ¡eso no puede ocurrir, existen garantías procesales y penales que lo impiden!". Sin embargo, vemos diariamente cómo estos sistemas se despliegan tentativamente, rodeados de secreto, escudados detrás de la intención de eficiencia, progreso tecnológico y reducción de costos: el decálogo de la vigilancia escalable.

En 2019, se dio a conocer que la AFIP, el organismo fiscal de Argentina, había llamado a licitación para instalar en Migraciones un sistema que reconociera el "temor" de los viajeros, a fin de captar contrabandistas. En una muestra de candor poco creíble, afirmaban que "El objetivo no es buscar evasores, sino contrabando y drogas"[81].

Entre la información recopilada en la noticia, se indicaba que el pliego requería un sistema que incluyera "funcionalidades de análisis termográficos, de emociones, de datos de redes sociales de fuentes abiertas y/o públicas y la capacitación del personal de AFIP, tanto para el uso como para la administración de la Solución, instalación de todos los componentes y su posterior mantenimiento y soporte técnico". Interesantemente, no se menciona capacitación alguna sobre ética de la Inteligencia Artificial, lo que muestra el estado de situación en el que estamos ante el avance de las ofertas de soluciones de este tipo, y el apetito de las políticas públicas[82].

En cuanto al reconocimiento de emociones, se explicaba que los sistemas "deberán medir distintos tipos, de

77. *https://themarkup.org/news/2021/03/30/texas-am-drops-race-from-student-risk-algorithm-following-markup-investigation*

78. *https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/*

79. *https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/*

80. Me refiero aquí a la frenología y a Cesare Lombroso con sus teorías sobre el "delincuente nato".

81. "La idea es pasar de un control aleatorio a uno inteligente. Los parámetros a verificar son: la frecuencia de vuelos del pasajero a un mismo destino, su capacidad económica para comprar los viajes y el cruce con otros organismos de seguridad nacionales e Interpol", detallan desde la AFIP. Pretenden que en un futuro cercano no haya más "semáforos" en la Aduana, y que se pueda saber de antemano a quién se debe controlar "cómo se hace en la mayoría de las aduanas modernas del mundo". En: "La AFIP buscará sospechosos en Ezeiza con cámaras de reconocimiento facial que detectan miedo y con datos de las redes", ***Infobae***, 2019. Disponible en: *https://www.infobae.com/economia/2019/07/04/la-afip-buscara-sospechosos-en-ezeiza-con-camaras-con-reconocimiento-facial-que-detectan-miedo-y-con-datos-de-las-redes-sociales/*

82. En cada charla y clase sobre el tema, recuerdo la frase que solía repetir mi abuela ***"el camino al infierno está tapizado de buenas intenciones"***. Más vigente que nunca en temas de política pública sobre Inteligencia Artificial.

las cuales al menos una de ellas deberá responder a la emoción primaria denominada 'temor' (ansiedad, preocupación, aprehensión)". A su vez, debía detectar "expresiones del rostro, analizar posturas y comportamiento corporal: se deberá alertar ante la medición de un nivel pre configurado de 'temor' en los rostros capturados por las videocámaras instaladas en el Puesto de Control de Aduana...". Vale decir que aquellos pasajeros con temor a volar, o ansiosos de arribar a un país desconocido, o que recibieron una mala noticia al bajar del avión, podrían ser señalados por un sistema que los marcaría como puntos de interés.

Además de potencialmente defectuosas, este tipo de aplicaciones parten de una premisa falsa de eficiencia. Más allá de los avances en IA e interfaces neuronales, la experiencia humana es rica, compleja, contextual e inherentemente irreductible. La simplificación que como herramienta de marketing presentan este tipo de aplicaciones, representan riesgos irreparables de daños, y son peligrosamente invasivas en áreas muy delicadas, por lo que su despliegue debería ser prohibido, cuidadosamente regulado y, en el caso de decidirse su implementación, monitoreado continuamente.

Aunque las hipótesis que pretenden representar no pueden ser contrastadas más allá de lo indiciario en tanto se encuentran en la mente de una persona, se están realizando constantemente avances en el área. Lejos de la ciencia ficción, varias compañías como Neuralink[83] de Elon Musk están desarrollando interfaces entre cerebros y computadoras (***Brain Computer Interface***, BCI). Particularmente Facebook en el contexto de su Reality Labs y de la investigación en el **Metaverso** como síncresis de VR y dispositivos hápticos está abocada al desarrollo de dispositivos capaces de decodificar el discurso sin que sea necesario articular palabras en voz alta[84]. Por ello, en la rama de **neuro-derechos** (Yuste et al., 2021), la privacidad mental se erige como un pilar fundamental de la autodeterminación que debe ser protegido de intromisiones.

En una carrera contrarreloj por la crisis del medio ambiente, la inclusión también debe considerar **el impacto ambiental** de los costos computacionales de entrenar e implementar modelos de Inteligencia Artificial, y cómo esto impacta en comunidades ya castigadas por las consecuencias del cambio climático.

El **extractivismo no es solo de datos**, sino de los componentes necesarios para construir el hardware que sustenta estas arquitecturas: desde la minería de los componentes base, al impacto ecológico del almacenamiento masivo de información y la refrigeración de los data centers. Incluso antes de la explosión reciente de blockchain y el mercado de NFTs, el consumo energético de las industrias tecnológicas representa un enorme porcentaje de las emisiones de carbono, con estimaciones que indicaban que "el sector tecnológico aportará entre el 3,0 y el 3,6% de las emisiones mundiales de efecto invernadero en 2020, más del doble de lo que produjo el sector en 2007"[85].

83. Autodefinida como una compañía que diseña "el primer implante neural que permitirá controlar un ordenador o dispositivo móvil en cualquier lugar. Los hilos a escala micrométrica se insertan en zonas del cerebro que controlan el movimiento. Cada hilo contiene muchos electrodos y los conecta a un implante, el Link." ***https://neuralink.com/***

84. "Imagining a new interface: Hands-free communication without saying a word", 30 de marzo de 2020. Disponible en: ***https://tech.fb.com/imagining-a-new-interface-hands-free-communication-without-saying-a-word/***

85. Los investigadores Lotfi Belkhir y Ahmed Elmeligi estiman que el sector tecnológico contribuirá con un 3,0-3,6 % de las emisiones globales de gases de efecto invernadero para 2020, más del doble de lo que produjo el sector en 2007 (Belkhir y Elmeligi, 2018). La huella global estimada para 2020 es comparable a la de la industria de la aviación y mayor que la de Japón, que es el quinto mayor contaminador del mundo. Los centros de datos representarán el 45 % de esta huella (frente al 33 % en 2010) y la infraestructura de red, el 24 %.

Para tomar un par de ejemplos concretos: el informe Clicking Clean de Greenpeace de 2017 mostró que las principales compañías de transmisión -Amazon Prime, HBO y Netflix- usan menos del 22% de las energías renovables. Y Northern Virginia, que alberga la mayor concentración de centros de datos del mundo, está alimentado por una empresa de servicios públicos con solo el 1 % de su electricidad procedente de fuentes renovables (Cook et al., 2017). "AI and Climate Change: How they're connected, and what we can do about it", por Roel Dobbe (Investigador Postdoctoral) y Meredith Whittaker (Co-Fundadora), ***AI Now Institut***, 2019. Disponible en:
https://medium.com/@AINowInstitute/ai-and-climate-change-how-theyre-connected-and-what-we-can-do-about-it-6aa8d0f5b32c

En ***"Anatomy of AI",*** Kate Crawford y Vladan Joler trazan las conexiones de impacto ecológico entre tecnología y explotación minera, conectando el litio que requieren las baterías de los dispositivos móviles, a su fuente geográfica en El Salar, situado en el suroeste de Bolivia, a una altitud de 3.656 metros sobre el nivel del mar. Indican que en este "altiplano, cubierto por unos metros de corteza de sal que son excepcionalmente ricos en litio, conteniendo entre el 50% y el 70% de las reservas mundiales de litio. El Salar, junto con las regiones vecinas de Atacama en Chile y Argentina, son los principales lugares de extracción de litio". El litio, es un "metal blando y plateado (que) se utiliza actualmente para alimentar dispositivos móviles conectados, como material crucial para la producción de baterías de iones de litio. Se le conoce como "oro gris". Las baterías de los ***smartphones***, por ejemplo, suelen tener menos de ocho gramos de este material. Cada automóvil Tesla necesita aproximadamente siete kilogramos de litio para su paquete de baterías." Aún más alarmante es que "Todas estas baterías tienen una vida útil limitada y, una vez consumidas, se desechan como residuos"[86].

Los grupos vulnerables sufren una doble marginalización. Por un lado, soportan brechas de acceso, conocimiento o de aprovechamiento de Internet, encontrándose en una posición débil de no poder controlar la información (***privacy as a commodity***). Pero, además, sufren a nivel climático las consecuencias de la industrialización de las tecnologías de la información, sin la infraestructura que tienen para afrontarla los países mayormente responsables de la producción y de las emisiones de esas tecnologías. En "***El estado del clima en América Latina y el Caribe 2020***"[87], la Organización Meteorológica Mundial informa que "...En los Andes chilenos y argentinos, los glaciares han retrocedido durante las últimas décadas. La pérdida de masa de hielo se ha acelerado desde 2010, en consonancia con un aumento de las temperaturas estacionales y anuales y una reducción considerable de las precipitaciones anuales en la región. La intensa sequía en el sur de la Amazonia y el Pantanal fue la más grave de los últimos 60 años, y 2020 superó a 2019 para convertirse en el año con mayor actividad de incendios en el sur de la Amazonia. La sequía generalizada en toda la región de América Latina y el Caribe ha tenido un impacto considerable en las rutas de navegación interior, en el rendimiento de los cultivos y en la producción de alimentos, agravando la inseguridad alimentaria en muchas zonas…".

Al ya urgente debate sobre **justicia climática,** debe agregarse que este no puede darse desatendido del impacto que la Inteligencia Artificial tiene sobre el ecosistema y aquellos que son los mayores perjudicados en sus mecanismos de explotación.

En síntesis, las aristas de la Inteligencia Artificial parecen inabarcables, en tanto (como se viera) cada problema presenta una miríada de facetas que deben ser consideradas, lo que complejiza aún más las conversaciones sobre el tema.

En la complejidad de la ética como disciplina, en lo abarcativo y abstracto que pueden ser sus fronteras, se corre el riesgo de que se escondan y difuminen debates necesarios que requieren marcos de trabajo accionables en concreto, que puedan ser trasladados a la práctica cotidiana.

---

86 "The Salar, the world's largest flat surface, is located in southwest Bolivia at an altitude of 3,656 meters above sea level. It is a high plateau, covered by a few meters of salt crust which are exceptionally rich in lithium, containing 50% to 70% of the world's lithium reserves. 4 The Salar, alongside the neighboring Atacama regions in Chile and Argentina, are major sites for lithium extraction. This soft, silvery metal is currently used to power mobile connected devices, as a crucial material used for the production of lithium-Ion batteries. It is known as 'grey gold.' Smartphone batteries, for example, usually have less than eight grams of this material. 5 Each Tesla car needs approximately seven kilograms of lithium for its battery pack. 6 All these batteries have a limited lifespan, and once consumed they are thrown away as waste." (Crawford y Joler, 2018).

87 "El estado del clima en América Latina y el Caribe 2020" (OMM-Nº 1272), Organización Meteorológica Mundial, 2021, Format: Digital - ISBN 978-92-63-31272-3.

El desafío continuará siendo desarrollar modelos que promuevan la inclusión y alivien la desigualdad, frente al cambio que enfrentamos como especie con la implementación extensiva de modelos de Inteligencia Artificial. En el recorrido brindado a través de las dimensiones de análisis propuestas, se busca dotar de un marco de trabajo que permita examinar las múltiples facetas de la Inteligencia Artificial en una forma sistematizada que permita aprehender la variedad de formas en que esta se manifiesta e impacta en la sociedad.

El método presentado aspira a cubrir este segmento, brindando una herramienta práctica para presentar análisis para sopesar el impacto de la Inteligencia Artificial a la hora de la toma de decisiones.

Nada haría más feliz a la autora que poder proporcionar herramientas pragmáticas que contribuyan a simplificar e informar ese proceso.

REFERENCIAS BIBLIOGRÁFICAS

Alpaydin, Ethem (2016). ***Machine Learning: The New AI. Cambridge:*** The MIT Press.

Angwin, Julia; Larson, Jeff; Kirchner, Lauren y Mattu, Surya (2016). "Machine Bias", en ***ProPublica.*** Disponible en: ***https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing***

Belkhir y Elmeligi (2018)
L. Belkhir and A. Elmeligi, "Assessing ICT global emissions footprint: Trends to 2040 & recommendations," Journal of Cleaner Production, vol. 177, pp. 448–463, Mar. 2018.

Bender, Emily M.; Gebru, Timnit; McMillan-Major, Angelina; Shmitchell, Shmargaret (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", en ***FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,*** pp. 610–623. Disponible en: ***https://doi.org/10.1145/3442188.3445922***

Beneteau, Erin; Boone, Ashley; Wu. Yuxing; Kientz, Julie A.; Yip, Jason y Hiniker, Alexis (2020). "Parenting with Alexa: Exploring the Introduction of Smart Speakers on Family Dynamics", en 2020 CHI Conference on Human Factors in Computing Systems. Disponible en: ***https://dl.acm.org/doi/abs/10.1145/3313831.3376344***.

Brennan, Tim; Dieterich, Bill; Breitenbach, Markus; Mattson, Brian, (2016). "Commentary on NCCD 'A questions of evidence: A critique of risk assessment models used in the justice system'". Disponible en: ***http://www.northpointeinc.com/files/white-papers/Baird_Response_060409.pdf***

Cook et al., 2017
G. Cook et al., "Clicking Clean: Who is winning the race to build a green internet?," Greenpeace, Washington, DC, Jan. 2017.

Crawford, Kate y Vladan Joler (2018). "Anatomy of an AI System, The Amazon Echo as an anatomical map of human labor, data and planetary resources". Disponible en: ***https://anatomyof.ai/***

Darrow (2017)
Barb Darrow (2017) "Automation, Robots, and Job Losses Could Make Universal Income a Reality Disponible en"*https://fortune.com/2017/05/24/automation-job-loss-universal-income/* ,

Domingos, Pedro (2017). ***The Master Algorithm***. United Kingdom: Penguin Books.

Doshi-Velez, Mason; Kortz, Ryan; Budish, Chris; Bavitz, Sam; Gershman, David; O'Brien, Kate; Scott, Stuart; Schieber, James; Waldo, David; Weinberger, Adrian; Weller, Alexandra Wood (2019). "Accountability of AI Under the Law: The Role of Explanation Finale". Disponible en: *https://arxiv.org/abs/1711.01134*

Fuller, Joseph; Raman, Manjari; Sage-Gavin, Eva y Hines, Kristen (2021). "Hidden Workers: Untapped Talent", publicado por Harvard Business School Project on Managing the Future of Work y Accenture. Disponible en: *https://www.hbs.edu/managing-the-future-of-work/Documents/research/hiddenworkers09032021.pdf*

Gray, Mary L. y Siddharth Suri (2019). ***Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass***. Disponible en: *https://ghostwork.info/*

Gopinath, Divya (2021). "What does it mean to be fair? Measuring and understanding fairness", ***en Towards Data Science***. Disponible en: *https://towardsdatascience.com/what-does-it-mean-to-be-fair-measuring-and-understanding-fairness-4ab873245c4c*

Hamilton et al. (2014)
"A path to understanding the effects of algorithm awareness", CHI '14 Extended Abstracts on Human Factors in Computing Systems Disponible en *http://dx.doi.org/10.1145/2559206.2578883* y *http://www.kevinhamilton.org/share/papers/p631-hamilton.pdf*

Hiniker, Alexis, Wang, Amelia; Tran, Jonathan; Zhang, Mingrui Ray; Radesky, Jenny y Sobel, Kiley (2021). "Can Conversational Agents Change the Way Children Talk to People?", en Conference: IDC '21: Interaction Design and Children. Disponible en: al.: *https://www.researchgate.net/publication/352741658_Can_Conversational_Agents_Change_the_Way_Children_Talk_to_People*).

Hunt, Elle (2016). "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter", en The Guardian.

Kaplan (2015)
"Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence", Yale University Press

Kaplan (2016)
"Artificial Intelligence: What Everyone Needs to Know", Oxford University Press.

Kuang, Cliff (2017). "Can A.I. Be Taught to Explain Itself?", en ***The New York Times***. Disponible en: ***https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html***

Larson, Jeff y Angwin, Julia (2016). "Technical Response to Northpointe", en ***ProPublica***. Disponible en: ***https://www.propublica.org/article/technical-response-to-northpointe***

Merrick, Amy (2014). "A Death in the Database", en ***The New Yorker***. Disponible en: ***https://www.newyorker.com/business/currency/a-death-in-the-database***

Noble, Safiya U. (2018). ***Algorithms of Oppression: How Search Engines Reinforce Racism***. Nueva York: NYU Press.

O'Neil, Cathy (2016). ***Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.*** United States of America: Crown Publishers.

Pasquale, Frank (2015). ***The Black Box Society: The Secret Algorithms That Control Money and Information***. Boston: Harvard University Press.

Paul, Cari (2016). "Microsoft Had to Suspend Its AI Chatbot After It Veered into White Supremacy", en ***Motherboard***.

Powles, Julia (2017). "New York City's Bold, Flawed Attempt to Make Algorithms Accountable", en ***The New Yorker***. Disponible en: ***https://www.newyorker.com/tech/elements/new-york-citys-bold-flawed-attempt-to-make-algorithms-accountable***

Roberts, Sarah T. (2019). ***Behind the Screen: Content Moderation in the Shadows of Social Media, Behind the Screen: Content Moderation in the Shadows of Social Media***. New Haven: Yale University Press.

Rothe, Rasmus (2017). "Applying deep learning to real-world problems", en ***Medium.*** Disponible en: ***https://medium.com/merantix/applying-deep-learning-to-real-world-problems-ba2d86ac5837***

Scassa***, Teresa (2021).*** "Privacy in the Precision Economy: The Rise of AI-Enabled Workplace Surveillance during the Pandemic", Center for International Governance Innovation. Disponible en:
https://www.cigionline.org/articles/privacy-in-the-precision-economy-the-rise-of-ai-enabled-workplace-surveillance-during-the-pandemic/

Shadoan, Rachel (2014). "Why Algorithm Transparency is Vital to the Future of Thinking", en Internet Archive. Disponible en: ***https://web.archive.org/web/20180131145401/https://akashiclabs.com/why-algorithm-transparency-is-vital-to-the-future-of-thinking/***

Shewan, Dan (2017). "Robots will destroy our jobs – and we're not ready for it", en ***The Guardian***. Disponible en: ***https://www.theguardian.com/technology/2017/jan/11/robots-jobs-employees-artificial-intelligence***

Starr, Sonja B. (2013). "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination", en ***Law & Economics Working Papers***. Paper 90. Disponible en: ***https://repository.law.umich.edu/cgi/viewcontent.cgi?article=1200&context=law_econ_current***

Vertesi (2014)
"My Experiment Opting Out of Big Data Made Me Look Like a Criminal" Disponible en ***http://time.com/83200/privacy-internet-big-data-opt-out/***

Victor, Daniel (2016). "Microsoft Created a Twitter Bot to Learn from Users. It Quickly Became a Racist Jerk", en ***The New York Times***.

Watkins, Ian (2011). "The utility of Level of Service Inventory-Revised (LSI-R) assessments within NSW correctional environments". Sydney: Corporate Research Evaluation & Statistics, NSW Dept. of Corrective Services. Disponible en: http://www.correctiveservices.justice.nsw.gov.au/Documents/utility-of-level-of-service-inventory-.pdf

Yuste, Rafael, Jared, Genser y Stephanie Herrmann (2021). "It's Time for Neuro-Rights", en ***Horizons: Journal of International Relations and Sustainable Development***.