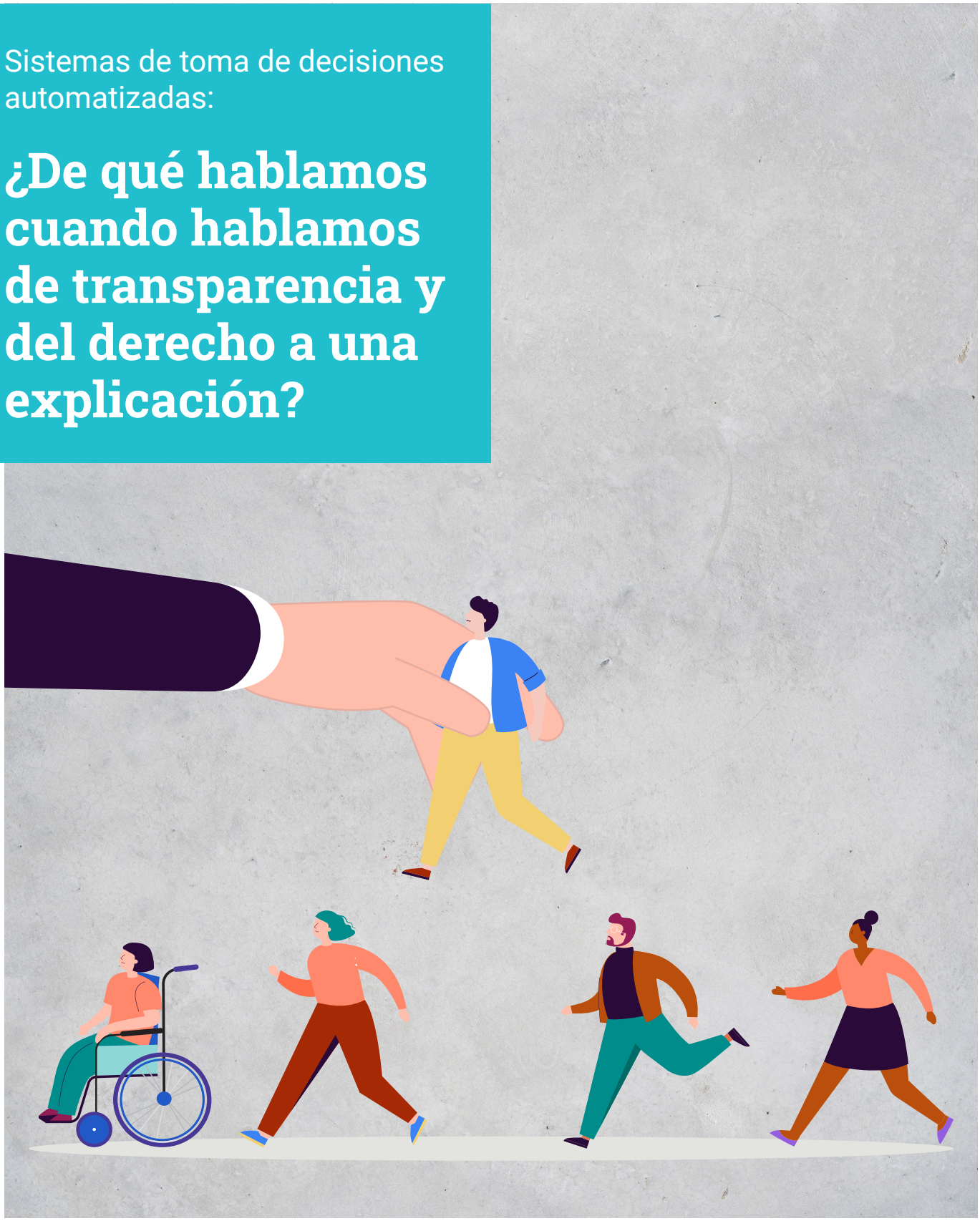
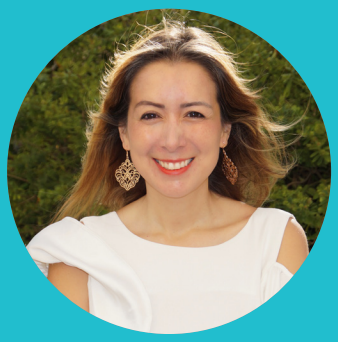
Sistemas de toma de decisiones automatizadas:

¿De qué hablamos cuando hablamos de transparencia y del derecho a una explicación?





CATHERINE MUÑOZ

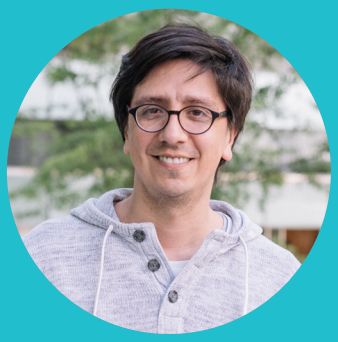
Abogada, Magíster en Derecho Internacional, Inversiones y Comercio por la Universidad de Chile y Master of Laws in International Law (LL.M.) por la Universidad de Heidelberg, especializada en propiedad intelectual y regulación de tecnologías, en particular, regulación de inteligencia artificial.

cmunozgut@gmail.com

JEANNA NEEFE MATTHEWS

Profesora de informática en Clarkson University (EE.UU.), copresidenta fundadora del Subcomité de Políticas de Tecnología de la ACM sobre Inteligencia Artificial y Responsabilidad Algorítmica, vicepresidenta del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) - Comité de Política de IA de EE. UU. y miembro del Comité de Políticas de Tecnología de la ACM (ACM TPC).

jnm@clarkson.edu



JORGE PÉREZ

Profesor Asociado del Departamento de Ciencias de la Computación de la Universidad de Chile e Investigador Asociado del Instituto Milenio Fundamentos de los Datos. Doctor en Ciencias de la Ingeniería por la Pontificia Universidad Católica de Chile. Sus intereses incluyen: datos Web, teoría de redes neuronales profundas, y el análisis de texto en medicina y política. En Twitter lo encuentras como @perez.

Introducción

A mediados de enero de 2021, en un hecho histórico, el Gobierno de los Países Bajos dimitió en bloque luego de una investigación realizada por el parlamento de dicho país que concluyó que el Jefe de Estado y sus principales ministros habían incurrido en faltas graves, evidenciando un menoscabo institucional y una discriminación sistemática contra un grupo vulnerable de la población holandesa. Esta imputación, tiene como fundamento la masiva y errónea acusación de fraude en la obtención de subsidios sociales en contra de 26.000 familias inocentes, de origen marroquí y tunecino en su gran mayoría [1].

La referida investigación constató que un sistema automatizado de toma de decisiones definía aquellos casos sospechosos de fraude en base a variables arbitrarias y abiertamente discriminatorias, como el simple hecho de tener una doble nacionalidad, evento que, por sí solo, situaba a las personas en una categoría de alto riesgo delictual. Lo anterior, unido a una mala gestión administrativa, injustamente obligó a estas familias a devolver dinero de subsidios recibidos. Muchas personas fueron llevadas a la quiebra, otras familias se desintegraron y la gran mayoría padeció estrés psicológico [2].

Lamentablemente este caso no es una excepción. Por el contrario, corresponde a una progresiva e instaurada regla general sobre el uso de sistemas automatizados de toma de decisiones que pueden afectar de manera radical la vida de las personas. Algunos ejemplos incluyen a sistemas predictivos de obtención de beneficios sociales cuya optimización se basó en reducir costos y reducir la mayor cantidad de otorgamiento de beneficios [3], sistemas calificadores de riesgos que utilizaron bases de datos, muchos de ellos con contenido de carácter sensibles, incompletos o falsos, proveídas por empresas *Data Brokers*, sin ningún estándar ético o legal [4], sistemas predictivos de justicia penal que castiga en mayor medida a grupos marginados de la población [5], sistemas de reconocimiento facial sesgados usados con fines de vigilancia y riesgosos resultados erróneos [6], y finalmente, la grave vulneración de derechos humanos, y en particular de la autonomía y privacidad de las personas, derivada del sistema automatizado de calificación de crédito social que impera en China [7].

La implementación de este tipo de sistemas en países en vías de desarrollo, como Chile, evidencian, asimismo, un creciente interés. Chile ha formulado dentro de sus políticas públicas y como

meta a corto plazo, la modernización de sus funciones y prestaciones de servicios [8], incorporando las referidas toma de decisiones automatizadas potenciadas con Inteligencia Artificial (IA). Lo anterior, bajo la consigna de eficiencia pública, administración efectiva y con la promesa de minimizar pérdidas de gastos fiscales, contribuyendo a políticas de austeridad [9].

Desde el punto de vista técnico, los sistemas de tomas de decisiones automatizadas pueden ser, o bien sistemas que apoyan determinadas decisiones teniendo la última palabra un ser humano, o sistemas que toman decisiones sin la intervención de personas [10]. Esta diferencia que pareciese ser trascendental, no es tal y en ambos casos existen similares niveles de riesgos en relación con la afectación de grupos protegidos. Por ejemplo, en el primer caso, también llamado "semiautomatizado", existe una tendencia comprobada; las personas confían más en el juicio de un algoritmo que en el propio cuando estos juicios están en contradicción [3].

Llama la atención que el entusiasmo por este tipo de tecnología no ha mermado a pesar de la abundante evidencia que alerta sobre el riesgo de aplicarlos a problemáticas sociales [11]. El denominador común en su aplicación es la naturaleza punitiva, lo que convierte a estos sistemas en una amenaza potencial de amplificación y perpetuación de injusticias sociales sobre grupos históricamente oprimidos y marginalizados, tales como pueblos originarios, afroamericanos, latinos, asiáticos, comunidades LGBTIQ+, musulmanes, personas de escasos recursos, entre otros [12].

Muchos de estos casos son evidentes e incuestionables discriminaciones, las que legalmente pueden ser acreditadas en un juicio. La información para documentar este tipo de casos toma como referencia los resultados de salida del sistema, junto con pruebas estadísticas y antecedentes relacionados con las personas involucradas en su diseño e implementación, sin necesitar información detallada del funcionamiento interno de los sistemas involucrados. Lo que se busca probar, en estos casos evidentes, es generalmente una discriminación indirecta, la cual ocurre cuando una norma, en este caso un sistema, aparentemente neutro, es aplicado a una población, perjudicando desproporcionadamente a grupos vulnerables de ésta [13]. En consecuencia, la recopilación de este tipo de información, en general, es suficiente para probar dicho "perjuicio dzesproporcionado". Éste es un tipo de "transparencia", pero no cualquiera, sino aquella estratégicamente obtenida para construir un caso judicial donde existe una evidente vulneración de derechos sobre las personas [14].

Ahora bien, ¿qué ocurre en aquellos casos donde la falta, error o injusticia son sutiles y no evidentes? Pensemos en un sistema de contratación de personal que ha rechazado una solicitud de empleo de una persona que cumplía todos los requisitos o un sistema de toma de decisiones que rechaza el ingreso de un joven a una universidad cumpliendo, asimismo, todos los requisitos para ello. Estas personas pueden albergar razonables dudas sobre si han sido injustamente excluidas o discriminadas, pero a diferencia de los casos anteriores, no es algo manifiesto. Incluso más, es posible que estos sistemas ya cuenten con auditorías que demuestren que su funcionamiento está supuestamente libre de sesgos de acuerdo con parámetros matemáticos de equidad [15]. Lamentablemente es común que estos parámetros obedezcan a una visión exclusivamente tecnocrática del problema y tengan poco sustento comparado con parámetros sociales de equidad [16, 17].

Los ejemplos más sutiles de sesgo son muy comunes, lo que va en contra de la creencia de muchas personas de que las decisiones tomadas por computadoras o sistemas automatizados son fundamentalmente lógicas e insesgadas. Y esto no es así. Las decisiones automatizadas se toman de dos formas principales: 1) según las instrucciones escritas por programadores humanos, o 2) según las reglas aprendidas automáticamente a partir de datos del pasado. Algunas personas pueden pensar que el problema principal proviene de las instrucciones escritas directamente por programadores humanos, pero de hecho, el aprendizaje automático sobre datos pasados suele crear problemas aún mayores. Aprender automáticamente desde datos del pasado es equivalente a considerar al pasado como el oráculo del futuro que queremos. En cierto sentido, aprendemos del pasado porque es todo lo que tenemos para aprender. Pero el pasado está lleno de prejuicios de muchos tipos. Si, por ejemplo, miramos quién ha sido un buen gerente en el pasado para definir quién será un buen gerente en el futuro, o quién ha sido un buen enfermero en el pasado para definir quién será un buen enfermero en el futuro, es posible que descartemos personas calificadas que no coinciden con el perfil más típico del pasado. Si codificamos estos datos del pasado en sistemas informáticos sin exigir una explicación de sus decisiones, entonces permitiremos que el pasado defina el futuro sin cuestionarlo. Estaríamos tomando la IA, que consideramos una fuerza progresista y futurista, para usarla como un oráculo y ejecutor conservador de prejuicios pasados.

Los conceptos clásicos de transparencia y participación social en la toma de decisiones, pilares fundamentales para prevenir y combatir la arbitrariedad y la discriminación, parecen quedarse cortos en el contexto actual. En particular, la transparencia puede tener diversas conceptualizaciones y se hace imprescindible distinguir en palabras sencillas transparencia, explicabilidad e interpretabilidad que son términos relacionados mas no sinónimos. ¿Qué exigimos entonces cuando exigimos transparencia y explicabilidad en las decisiones de un sistema automático?

Aprender automáticamente desde datos del pasado es equivalente a considerar al pasado como el oráculo del futuro que queremos.

No pretendemos responder cabalmente a la pregunta sino más bien aportar a la discusión desde una visión legal y computacional. Éste es el punto de partida de este artículo y nuestra motivación de escribirlo.

El concepto clásico de transparencia

Durante la última década se ha discutido sobre el nivel de transparencia que debe existir en el desarrollo y uso de sistemas de IA, en particular, en aquellos que toman decisiones automatizadas y que potencialmente pueden tener un impacto negativo sobre las personas. La transparencia ha sido instaurada como uno de los principios esenciales en esta materia y guarda relación con la capacidad de proporcionar información que permita comprender cómo se desarrolla y despliega un sistema de IA [18, 19]. Al respecto, la Iniciativa Global de IEEE sobre Ética de Sistemas Autónomos e Inteligentes ha establecido cuatro condiciones para guiar la confianza informada de los sistemas autónomos e inteligentes: 1) efectividad, 2) competencia, 3) rendición de cuentas y siendo la 4) precisamente la transparencia [20].

La necesidad de transparencia es contrastada con el hecho de que los sistemas de IA, particularmente los modelos de *deep learning* que tienen una estructura compleja, no permiten transparentar completamente su funcionamiento, siendo en muchos casos imposible explicar la construcción y decisiones de éstos, incluso para sus propios desarrolladores, la famosa caja negra. Más aún, una explicación satisfactoria [21] dependerá de la audiencia; algo que pueda ser considerado como una explicación o evidencia clara para un grupo (p.ej., código fuente de un sistema para un desarrollador de software), puede resultar opaco para otro grupo o simples detalles técnicos para un tercer grupo. A pesar de esto, diversos grupos de investigación están actualmente trabajando en proponer mecanismos para una transparencia efectiva y con sentido.

La transparencia no es sinónimo de igualdad

Comúnmente, el análisis de transparencia es *ex-ante* (antes de que el sistema se implemente), y no *ex-post* (después de que el sistema ya esté implementado y tenga un impacto en la vida de las personas). En ese sentido, se entiende que la transparencia y exigencia de información pertinente, es un requisito para la construcción de la confianza entre los ciudadanos y entidades públicas o privadas y los sistemas que éstos proveen de forma previa a su uso, de manera que las personas puedan contar con antecedentes necesarios para tomar la decisión de aceptar con cierta confianza el uso de un modelo algorítmico que puede impactarlo directamente. Pero esto es cierto sólo respecto de una parte de la población, generalmente de clases acomodadas, ya que respecto de personas vulnerables o de escasos recursos, el uso de sistemas tecnológicos en temáticas que les impactan no les es consultado y menos explicado. Hasta cierto punto, exigir y obtener transparencia es un "privilegio", un elemento más que suma e incrementa la desigualdad estructural de nuestra sociedad. En síntesis, a las personas pobres simplemente les imponen sistemas cuyas decisiones pueden afectar sus vidas a largo plazo independientemente de la transparencia.

En efecto, desde orígenes coloniales las personas de escasos recursos no han tenido control sobre su privacidad ni decisiones, en comparación con personas de clases de mayores ingresos. A lo anterior, se agrega el hecho que, debido a segregaciones y desigualdades, existe una brecha de conocimiento en las personas sobre cómo funcionan las herramientas tecnológicas y la forma en que pueden proteger sus derechos. Adicionalmente en muchos casos, la mayoría de las personas no son conscientes que están siendo parte de sistemas tecnológicos ni de los riesgos asociados [22]. Éste es un aspecto crítico que debe ser democratizado mediante mecanismos de inclusión y en consideración a la dignidad de todos los ciudadanos. Como hemos mencionado, una transparencia suficiente para una persona puede no serlo para otra, por lo que deben existir estándares de acceso a la información que consideren el entendimiento de todos los ciudadanos.

La obtención de información se complejiza, tomando en consideración que existen diferentes definiciones contrapuestas sobre conceptos relevantes como igualdad, discriminación y *fairness* [23]. Por ejemplo, dar prioridad a los derechos de los individuos, priorizar el bienestar de la sociedad en su conjunto, proteger a los grupos marginados, incluso proteger a todas las especies del planeta. *Fairness* es un concepto esencial en países de Europa o en Estados Unidos, que se opone al concepto legal de discriminación, y que posee distintas interpretaciones, dependiendo si se usa en el área computacional, social o legal [24]. Este concepto no posee un equivalente exacto en Chile ni en Latinoamérica, siendo interpretado indistintamente como imparcialidad, equidad o justicia [25] razón por la cual, en este artículo no le daremos una traducción e interpretación determinada.

Dado que las definiciones de *fairness* y ética pueden variar, es especialmente importante que todos los actores que

tienen interés en un sistema, y no sólo los desarrolladores o usuarios contratantes, reciban información que les permita discutir sus prioridades en procesos decisorios. En ese sentido, la transparencia es necesaria para que todas las partes interesadas puedan debatir en un proceso decisorio en torno a la definición de *fairness* que les parezca adecuada y no ceder esta decisión a los creadores, diseñadores y programadores de estos sistemas. En Grasso et al [21] se ha argumentado que el proceso de automatización a menudo desplaza las grandes decisiones de expertos en un dominio determinado hacia programadores sin experiencia en esta área y se discute cómo integrar los marcos de responsabilidad algorítmica con herramientas como "fichas técnicas para *datasets*" [26] y "Tarjetas modelo para informes de modelos" [27] con los códigos de ética específicos de esta materia [21].

La transparencia no es sólo técnica, sino también social

No se debe perder de vista que estamos en presencia de sistemas sociotécnicos. En ese sentido, no pueden ser entendidos sólo desde la técnica, ya que junto a ésta, toman relevancia motivaciones e intereses de las personas que poseen una relación directa en la creación e implementación de un determinado sistema. La suma de factores técnicos y sociales, inciden directamente en los impactos del despliegue de este tipo de tecnología. Como dice Shoshana Zhuboff en su libro The Age of Surveillance Capitalism, debemos preguntarnos: ¿quién sabe? ¿quién decide quién sabe? y ¿quién decide quién decide? [28].

En particular, respecto de sistemas de IA utilizados en políticas públicas, la transparencia desde el punto de vista social se traduce en parte en contar además de información técnica, con información política y social sobre los diseñadores y tomadores de decisiones, sobre la elección de determinados datos, características, modelos, qué tipo de patrones busca, por qué a unas personas sí y otras no, o por qué se dirige a determinado grupo o ámbito geográfico, etc. En definitiva, información sobre las decisiones políticas detrás de las decisiones técnicas.

Para el cumplimiento del estándar anterior, esta transparencia lleva implícita la condición que organismos públicos no adquieran sistemas de IA que estén protegidos por secretos comerciales o acuerdos de confidencialidad. En el mismo sentido, es necesario que exista una transparencia activa del Estado, con mecanismos como registros y plataformas públicas, además de procesos de licitación abiertos. La colaboración público-privada debe ser totalmente transparente, haciendo público conflictos de intereses, contratos con proveedores y cualquier información relevante, cumpliendo con las más altas exigencias de probidad y rendición de cuentas.

Asimismo, en el caso de software de uso público, los gobiernos tienen la oportunidad de establecer requisitos técnicos adicionales tanto para su propio desarrollo como para la compra de software desarrollados por terceros. Así por ejemplo, en la fase de diseño o adquisición se podrían establecer requerimientos de factores pro transparencia, como disponer de software de código abierto, acceso a artefactos de ingeniería de software, incluidos documentos de requisitos y diseño, seguimiento de errores y bitácoras de cambios en el código, planes de prueba y resultados [21].

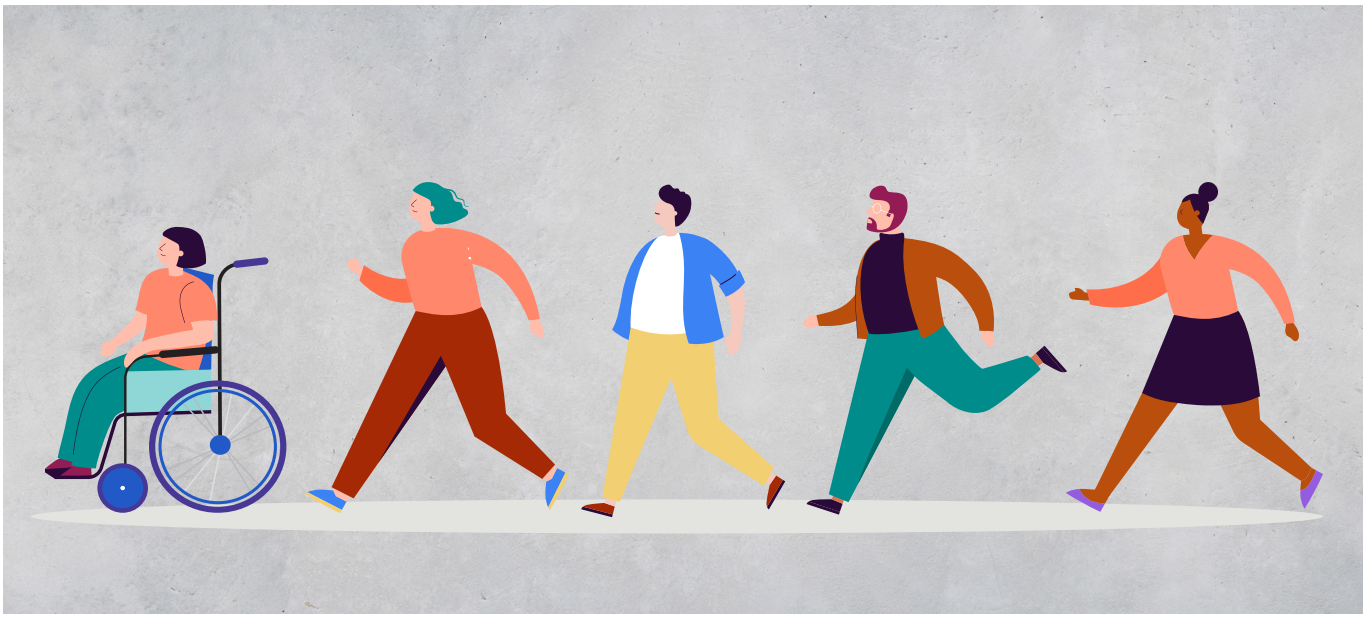
Explicabilidad e interpretabilidad

Hasta ahora nos hemos concentrado principalmente en el concepto de transparencia de los sistemas automáticos desde una perspectiva general y sobre la necesidad de contar con distintas visiones al momento de su construcción y despliegue.

Hasta cierto punto, exigir y obtener transparencia es un "privilegio", un elemento más que suma e incrementa la desigualdad estructural de nuestra sociedad.

En ese sentido, si bien la transparencia es algo deseable, en la práctica necesitamos también ser capaces de auditar el funcionamiento de los sistemas de manera dinámica, mientras están tomando las decisiones. Es aquí donde surgen dos conceptos que hemos mencionado tangencialmente pero que son de vital importancia: la *interpretabilidad* y la *explicabilidad* de un sistema de toma de decisiones automatizada.

Para una conceptualización útil de explicabilidad, podemos centrarnos en la decisión de un sistema en un caso específico, por ejemplo "una solicitud de crédito que fue rechazada". Lo que buscamos entonces, es que un humano sea capaz de entender la razón de esa decisión particular ("¿por qué fue rechazada la solicitud?"). Usualmente a esto se le llama explicación *post-hoc* y local. *Post-hoc* se refiere a que la explicación se hace considerando los veredictos del sistema después de que el sistema ya está desplegado y en funcionamiento, mientras que local se refiere a explicar una decisión particular (en oposición a explicar el sistema como un todo). Que una decisión sea explicable en un sistema, no significa que el funcionamiento en general (para todas las posibles decisiones) sea explicable también. A esta explicación global le llamamos interpretabilidad; un sistema sería interpretable entonces, si un humano es capaz de entender la manera en que el sistema toma todas sus decisiones.



De la misma manera, se debe tener presente que cualquier explicación es una simplificación del sistema completo. Larraraju et al. [29] establecen claras métricas para determinar la calidad de las explicaciones, que incluyen la fidelidad, es decir, el grado en que la explicación coincide con el sistema completo, la falta de ambigüedad o el grado en que la explicación aísla un único resultado para cada caso, y la interpretabilidad, es decir, el grado en que las personas pueden entender la explicación. La fidelidad puede medirse minimizando la cantidad de desacuerdo entre la explicación y el sistema completo. La falta de ambigüedad puede medirse minimizando la cantidad de solapamiento entre las reglas de la explicación y maximizando el número de casos cubiertos por la explicación. La interpretabilidad puede medirse minimizando el número de reglas, el número de predicados utilizados en esas reglas y la amplitud del número de casos considerados por cada nivel en el árbol de decisiones (por ejemplo, si X_1 entonces Y_1, si X_2 entonces Y_2, si X_3 entonces Y_3, sería de amplitud 3). Otras propiedades deseables de las explicaciones pueden ser que no utilicen características inaceptables (por ejemplo, utilizar la raza o el género en las decisiones de contratación) o que proporcionen una orientación predictiva (por ejemplo, si tuviera más experiencia en la categoría *X*, tendría más probabilidades de ser contratado para este trabajo en el futuro). En definitiva, en la comunidad científica se sigue trabajando en las características de las buenas explicaciones y existe una tensión natural entre diferentes características como la interpretabilidad y la fidelidad, aún no resuelta.

Un intento de formalización y la esperanza de auditabilidad

La anterior discusión se basa en que "un humano sea capaz de entender" algo, lo que es sumamente difícil de formalizar y definir de una única forma. Una manera de concretizar el problema es llevarlo a un tipo particular de explicación. Una muy usada es la del tipo contrafactual; en vez de preguntarnos el porqué de la decisión, nos preguntamos cómo cambiaría la decisión en presencia de antecedentes distintos ("¿hubiese sido rechazada la solicitud si el postulante hubiera sido una persona casada?"). Este tipo de preguntas se han usado recientemente para comparar la interpretabilidad de distintos sistemas de manera formal independiente de las características del sistema en cuestión. Más precisamente, supongamos que un sistema *M* toma cierto veredicto cuando es presentado con un conjunto *A* de antecedentes, y consideremos la siguiente pregunta: ¿cuál es el mínimo grupo de antecedentes que es necesario cambiar en *A* para cambiar también el veredicto de *M*? Podríamos definir entonces que un sistema automático es interpretable, si para cada posible conjunto de antecedentes, la anterior pregunta se puede responder en un tiempo prudente ("tiempo polinomial" en jerga computacional). Esta definición aseguraría que, por ejemplo, cada persona a la que se le haya rechazado una solicitud de crédito, podría obtener en un tiempo prudente una explicación del tipo "si cambia este grupo de antecedentes, el crédito sería aprobado".

Sin perjuicio de lo anterior, debemos notar que esta definición de interpretabilidad es sumamente acotada y posiblemente sea útil sólo en ciertos contextos. Si bien esta perspectiva es acotada, es formal, y una de las consecuencias de definir formalmente un problema de interpretabilidad, es que podemos poner a prueba de manera precisa y comparativa

Existe el riesgo de que los usuarios entendamos la explicación [acerca de la respuesta otorgada por un sistema automático] como producto de causalidades.

a distintas clases de sistemas automatizados. En efecto, con esta definición se puede demostrar formalmente la creencia popular de que sistemas basados en árboles de decisión son más interpretables que sistemas basados en redes neuronales profundas [30, 31]. Otro punto positivo de contar con una definición del tipo anterior, es que un sistema interpretable se podría auditar respecto de la existencia de sesgos en sus veredictos. Por ejemplo, si hubiese un conjunto de antecedentes protegidos (como género o raza), podríamos exigir de manera efectiva que el solo cambio de esos antecedentes protegidos no cambien el veredicto del sistema [32].

Si bien hemos mostrado posibilidades de resolver problemas de interpretabilidad de una manera un poco más precisa, la aplicación de la definición anterior (o cualquier otra que se proponga desde la técnica), no debiera obviar aspectos sociales. Por ejemplo, no debieran ser los mismos sistemas los que definan cuáles son los antecedentes protegidos. También se debe tomar en cuenta que las explicaciones serán consumidas por personas y por lo tanto se debiera evitar la jerga técnica y presentar explicaciones precisas pero simples de entender, que incluyan modelos cuantitativos, cualitativos y antropológicos, entre otros [33].

Explicaciones *post-hoc*, locales, basadas en contrafactuales y que puedan generarse en tiempo razonable (polinomial), son esencialmente conceptos técnicos y las formalizaciones han venido principalmente desde el mundo científico. En consecuencia, no debemos perder de vista que cualquier definición técnica puede tener implicancias en la forma en que las personas entenderán el proceso real para el que se usa el sistema. Por ejemplo, una explicación contrafactual ("qué habría pasado si cambiaba el antecedente *x*") no es necesariamente causal ("el antecedente *x* es el más importante en la decisión del sistema") sin embargo existe el riesgo de que los usuarios entendamos la explicación como producto de causalidades [34]. Se hace necesario entonces que la sociedad, y más precisamente la legislación, defina, al menos conceptualmente, qué tipo de explicaciones, interpretaciones y estándares deben ser exigidos a los sistemas automatizados. Visualizamos acá un círculo virtuoso: las definiciones sociales podrán guiar el desarrollo técnico, incentivando la cooperación y búsquedas de soluciones interdisciplinarias, enfocando así recursos y esfuerzos de investigación.

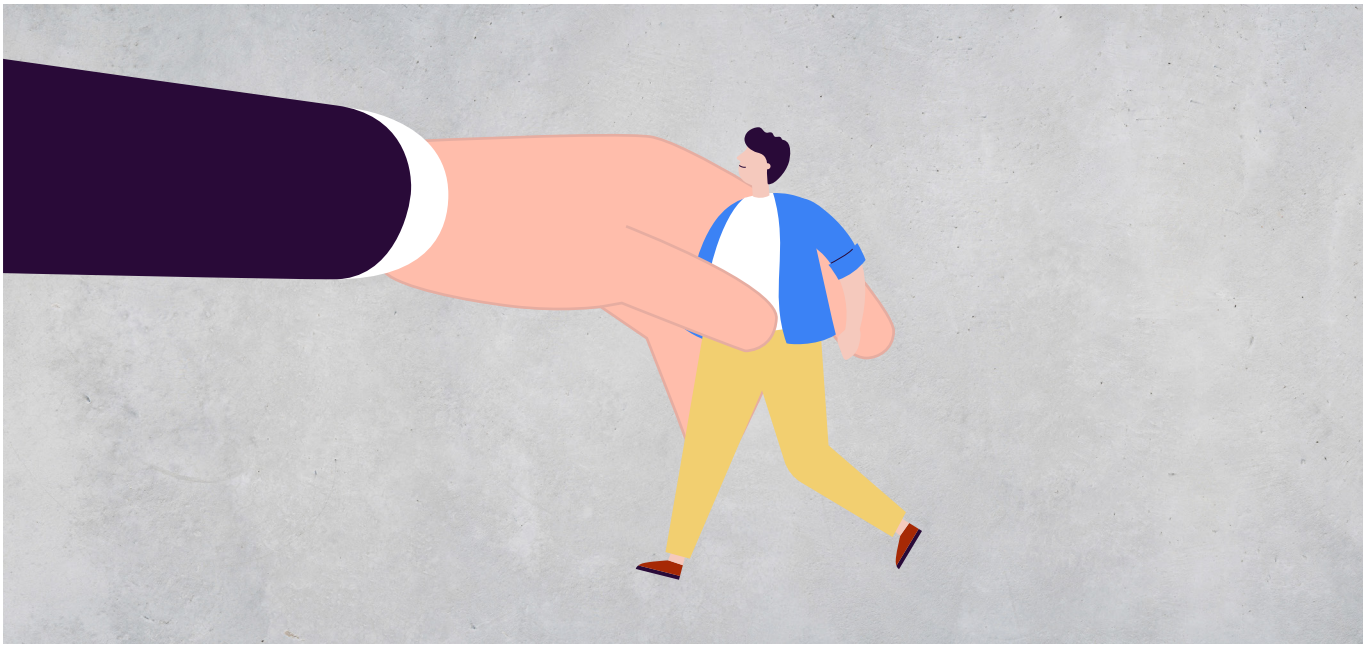
Transparencia algorítmica y el proceso constituyente

Mientras en todo el mundo los sistemas basados en IA están cambiando la forma en que se deciden aspectos importantes de la vida de las personas, Chile se encuentra en un proceso histórico de diseño de una nueva Constitución. En este contexto, Chile tiene la oportunidad de delinear el rol que los sistemas de IA tendrán en la toma de decisiones acerca de la asignación de fondos públicos, puestos de trabajo, vivienda, créditos, acceso a la salud, justicia, prevención del delito y muchos otros.

La transparencia, como un concepto general, más que un principio propiamente tal, es un medio que hace posible lograr el ejercicio de derechos fundamentales. Esto toma una relevancia adicional en relación con el uso de sistemas de IA. La Constitución, además de mantener el equilibrio de los poderes del Estado, consagra derechos fundamentales. De estos derechos, los que más riesgo de vulnerabilidad corren a la luz del uso de sistemas de tomas de decisiones automatizadas poco transparentes o no explicables, corresponden principalmente a los derechos de igualdad, privacidad y protección de datos, debido proceso y acceso a un juicio justo, seguridad, autonomía, así como, acceso a información y libertad de expresión.

Respecto del derecho de igualdad consideramos que es una oportunidad histórica consagrar expresamente a la igualdad no como "no discriminación" sino como un principio de antisubordinación. El propósito del principio de igualdad desde esta perspectiva (que muchos autores llaman igualdad real) tiene por finalidad eliminar las estructuras sociales históricamente discriminatorias y excluyentes [35]. Lo anterior tiene una importante consecuencia sobre la regulación de sistemas de toma de decisiones automatizadas, ya que se traduce en que cualquier resultado de éstos, que reproduzca y perpetúe condiciones estructurantes de injusticia social, no serán tolerados por la legislación y serán sancionados, sin considerar otros elementos como la intención de provocar daños. Este punto es importante cuando no podemos contar con toda la transparencia requerida frente a potenciales efectos negativos en el uso de sistemas de toma de decisiones automatizadas.

Por su parte, sobre la protección de la privacidad y la protección de datos personales, la transparencia, y la interpretabilidad, cumplen un rol fundamental. Notable es el caso de los artículos 13° y 15° del Reglamento General de Protección de Datos (GDPR, por sus siglas en inglés) en Europa, que proveen el derecho a una "explicación significativa de



la lógica involucrada" en las decisiones automáticas. Selbst y Powles [36] consideran que esto traza un fundamento claro hacia el "derecho a la explicación", que son complementadas con los artículos 22° y 35° del mismo cuerpo legal. Chile tiene una oportunidad histórica de consagrar de manera no ambigua en su nueva Constitución el "derecho a la explicación" respecto de sistemas de IA, en particular, de toma de decisiones automatizadas.

Considerando lo descrito en puntos anteriores, específicamente sobre los límites y riesgos de explicaciones descontextualizadas o no entendidas, creemos que tomando todas las prevenciones del caso, es fundamental el establecimiento de un "Derecho a la transparencia y suministro de información sobre sistemas de toma de decisiones automatizadas", consagrados en la nueva Constitución dentro de un "Derecho a la transparencia e información" de carácter más general, el cual para garantizarlo, debe ser complementado con la promulgación de normas de rango legal en donde se detallen los mecanismos y estándares para su cumplimiento. Al respecto, la reciente publicación de la Propuesta de Reglamento del Parlamento Europeo y del Consejo Europeo que establece normas armonizadas sobre la inteligencia artificial (Ley de Inteligencia Artificial, publicada con fecha 21 de abril de 2021 [EU Council 2021]), es un excelente ejemplo del contenido mínimo que debieran tener estas futuras normas legales, además de las ya referidas al GDPR, para el debido ejercicio de este nuevo derecho constitucional.

La Propuesta de Reglamento del Parlamento Europeo sobre la inteligencia artificial establece estándares de transparencia, registro y explicabilidad, respecto de sistemas considerados por este cuerpo legal como de alto riesgo, y que pueden ser resumidos en los siguientes puntos:

a. Deben contener instrucciones de uso con información concisa, pertinente, accesible y comprensible, sobre datos de proveedor, características, capacidades y limitaciones de funcionamiento, finalidad prevista, rendimiento, especificaciones de los datos de entrada, las medidas de supervisión humana, incluidas las medidas técnicas establecidas para facilitar la interpretación de los resultados de los sistemas de IA por parte de los usuarios; entre otros.

b. Deben contener documentación técnica sobre finalidad prevista, desarrolladores, la interacción del sistema con hardware o software que no forma parte del mismo, los métodos y pasos realizados para el desarrollo del sistema, incluido, el uso de sistemas preentrenados o de herramientas proporcionadas por terceros, lógica general del sistema y de los algoritmos, las opciones clave de diseño, las personas o grupos de personas con los que se pretende utilizar el sistema, opciones de clasificación, entre otras.

c. Información detallada sobre el seguimiento, el funcionamiento y el control de sistemas de IA, en particular, respecto a sus capacidades y limitaciones, incluidos los grados de precisión para grupos de personas específicos en los que se prevé utilizar y el nivel general de precisión esperado en relación con su finalidad prevista. A este último punto se debe complementar el requisito que el nivel de precisión debe estar avalado por metodologías con bases científicas robustas e independientes.

A lo anterior, se debiese agregar la obligación de efectuar una evaluación de impacto en relación con la afectación de derechos humanos. Las evaluaciones dejan documentado el proceso de acuerdo con la letra (b) precedente y permiten prever riesgos antes de su implementación y posibles mejoras o derechamente decidir sobre su no uso.

Conclusiones

La transparencia y el acceso a la información es una idea que ha ocupado un lugar destacado en la agenda política de las sociedades democráticas occidentales durante muchos años. Ha sido cultivada, propagada y, a veces, mal utilizada por los medios de comunicación en forma interesada.

En este artículo intentamos contribuir a la discusión, considerando la importancia de distinguir las distintas funciones de la transparencia y de contar con explicaciones e interpretaciones sobre las decisiones que toman los sistemas automáticos de manera que todas las partes interesadas y posiblemente afectadas puedan entender y responder a ellas.

En particular, consideramos que se debe promover un acceso equitativo sobre transparencia social y aspectos técnicos, teniendo presente que estamos frente a sistemas sociotécnicos, así como promover el acceso a información interpretable que pueda ser usada por profesionales especializados. Para ello nos encontramos en una oportunidad histórica de plasmarlo en nuestra nueva Constitución como un derecho consagrado para todos los chilenos.

Lo anterior en ningún caso se debe interpretar como que estas propuestas conllevan una carga sobre las personas respecto de la decisión de determinar si un sistema de IA es confiable o no. Sería una carga injusta para lo cual no estamos capacitados, por lo que siempre será una obligación del Estado asegurar que estos sistemas sean confiables y cumplan con todos los estándares necesarios para la protección de los ciudadanos y en particular de aquellos más vulnerables.

Finalmente, tanto o más importante que decidir qué rol esperamos que cumpla la IA, es el determinar qué rol esperamos que no cumpla y para ello el análisis en el uso de sistema de toma de decisiones automatizadas no puede ser abordado netamente desde una perspectiva económica de costos versus beneficios, sino que se debe considerar si corresponde desplegar este tipo de sistemas en consideración a los derechos y dignidad de las personas. Para asegurarnos de que esto se cumpla, requerimos, nuevamente, transparencia e información.

Como profesionales del área legal y de las ciencias de la computación, sabemos que los sistemas computacionales complejos cometen errores, y a veces muchos errores. Por eso estamos en contra de un mundo regido por el principio de que "el computador sabe más que nadie" o la creencia de que, a diferencia de los humanos, los sistemas automáticos "pueden tomar decisiones sin sesgos". Soluciones simplistas, o que sólo vengan del mundo técnico podrían, más que ayudar, crear más daño. Éste es uno de esos problemas en donde basados en ciencia y evidencia, pero sobre todo basados en el bien común, debemos buscar una solución como sociedad. ▪

REFERENCIAS

[1] G. Geiger, «How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud», ene. 01, 2021. https://www.vice.com/en/article/jgq35d/how-a-discriminatory-algorithm-wrongly-accused-thousands-of-families-of-fraud (accedido abr. 28, 2021).

[2] T. K. der Staten-Generaal, «Parlementaire ondervraging kinderopvangtoeslag; Brief Presidium; Brief van het Presidium over een voorstel voor een parlementaire ondervraging kinderopvangtoeslag», jul. 01, 2020. https://zoek.officielebekendmakingen.nl/kst-35510-1 (accedido abr. 28, 2021).

[3] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[4] H. Fry, *Hello world: Being human in the age of algorithms*. WW Norton & Company, 2018.

[5] J. N. Matthews *et al.*, «When Trusted Black Boxes Don't Agree: Incentivizing Iterative Improvement and Accountability in Critical Software Systems», 2020, pp. 102-108.

[6] K. Hill, «What Happens When Our Faces Are Tracked Everywhere We Go?», *The New York Times*, mar. 18, 2021.

[7] S. Engelmann, M. Chen, F. Fischer, C.-Y. Kao, y J. Grossklags, «Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior», ene. 2019, pp. 69-78, doi: 10.1145/3287560.3287585.

[8] https://digital.gob.cl, «Ley de Transformación Digital», *Ley de Transformación Digital*. http://digital.gob.cl/transformacion-digital/ley-de-transformacion-digital/ (accedido abr. 28, 2021).

[9] J. Hughes, «Algorithms and posthuman governance», *J. Posthuman Stud.*, vol. 1, n.º 2, pp. 166-184, 2018.

[10] C. Orwat, «Risks of Discrimination through the Use of Algorithms. A study compiled with a grant from the Federal Anti-Discrimination Agency», 2020.

[11] F. Chiusi *et al.*, «Automating Society Report 2020», *Automating Society Report 2020*. https://automatingsociety.algorithmwatch.org (accedido abr. 28, 2021).

[12] R. Benjamin, «Race after technology: Abolitionist tools for the new jim code», *Soc. Forces*, 2019.

[13] T. Khaitan, *A theory of discrimination law*. OUP Oxford, 2015.

[14] S. Wachter, B. Mittelstadt, y C. Russell, «Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI», *ArXiv Prepr. ArXiv200505906*, 2020.

[15] K. Creel y D. Hellman, «The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems», *Va. Public Law Leg. Theory Res. Pap.*, n.o 2021-13, 2021.

[16] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, y J. Vertesi, «Fairness and abstraction in sociotechnical systems», 2019, pp. 59-68.

[17] M. Srivastava, H. Heidari, y A. Krause, «Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning», 2019, pp. 2459-2468.

[18] S. Garfinkel, J. Matthews, S. S. Shapiro, y J. M. Smith, «Toward algorithmic transparency and accountability», 2017.

[19] A. Now, «The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems», *https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/*, 2018.

[20] K. Shahriari y M. Shahriari, «IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems», 2017, pp. 197-201.

[21] I. Grasso, D. Russell, A. Matthews, J. Matthews, y N. R. Record, «Applying Algorithmic Accountability Frameworks with Domain-specific Codes of Ethics: A Case Study in Ecosystem Forecasting for Shellfish Toxicity in the Gulf of Maine», 2020, pp. 83-91.

[22] M. Madden, M. Gilman, K. Levy, y A. Marwick, «Privacy, poverty, and big data: A matrix of vulnerabilities for poor Americans», *Wash UL Rev*, vol. 95, p. 53, 2017.

[23] A. Narayanan, «Translation tutorial: 21 fairness definitions and their politics», 2018, vol. 2, n.o 3, pp. 6-2.

[24] A. Xiang y I. D. Raji, «On the legal compatibility of fairness definitions», *ArXiv Prepr. ArXiv191200761*, 2019.

[25] J. Rawls, «Justice as fairness», *Philos. Rev.*, vol. 67, n.o 2, pp. 164-194, 1958.

[26] T. Gebru *et al.*, «Datasheets for datasets», *ArXiv Prepr. ArXiv180309010*, 2018.

[27] M. Mitchell *et al.*, «Model cards for model reporting», 2019, pp. 220-229.

[28] S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019*. Profile Books, 2019.

[29] H. Lakkaraju, E. Kamar, R. Caruana, y J. Leskovec, «Faithful and customizable explanations of black box models», 2019, pp. 131-138.

[30] P. Barceló, M. Monet, J. Pérez, y B. Subercaseaux, «Model Interpretability through the lens of Computational Complexity», *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 15487-15498, 2020.

[31] Z. C. Lipton, «The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.», *Queue*, vol. 16, n.o 3, pp. 31-57, jun. 2018, doi: 10.1145/3236386.3241340.

[32] P. Barceló, J. Pérez, y B. Subercaseaux, «Foundations of Languages for Interpretability and Bias Detection». *Algorithmic Fairness through the Lens of Causality and Interpretability Workshop at NeurIPS 2020*

[33] M. M. Malik, «A Hierarchy of Limitations in Machine Learning», *ArXiv Prepr. ArXiv200205193*, 2020.

[34] R. Moraffah, M. Karami, R. Guo, A. Raglin, y H. Liu, «Causal interpretability for machine learning-problems, methods and evaluation», *ACM SIGKDD Explor. Newsl.*, vol. 22, n.o 1, pp. 18-33, 2020.

[35] R. B. Siegel, «Equality talk: Antisubordination and anticlassification values in constitutional struggles over Brown», *Harv Rev*, vol. 117, p. 1470, 2003.

[36] A. D. Selbst y J. Powles, «Meaningful Information and the Right to Explanation», Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3039125, nov. 2017. Accedido: abr. 28, 2021. [En línea]. Disponible en: https://papers.ssrn.com/abstract=3039125.