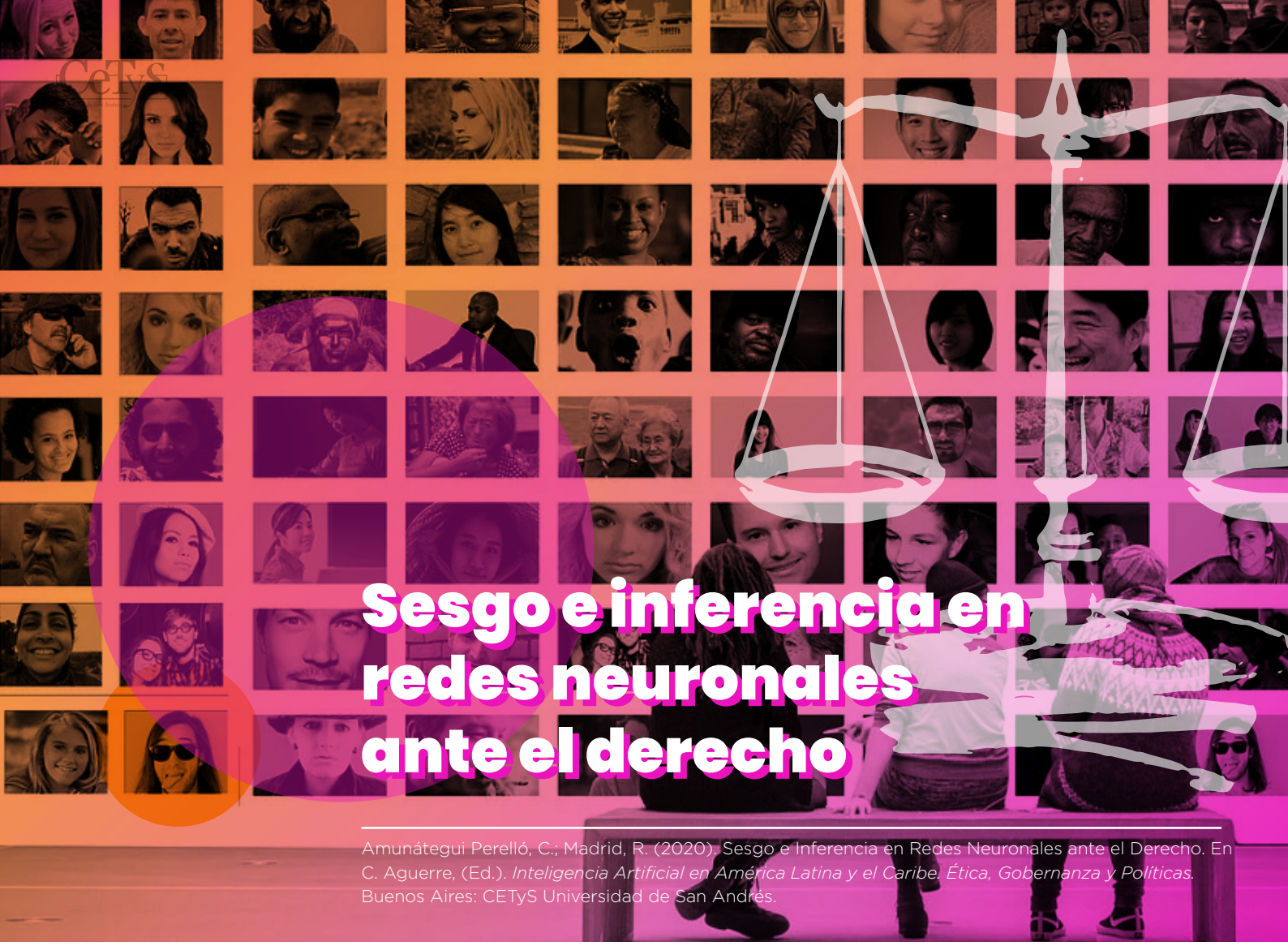


Sesgo e inferencia en redes neuronales ante el derecho

Amunátegui Perelló, C.; Madrid, R. (2020). Sesgo e Inferencia en Redes Neuronales ante el Derecho. En C. Aguerre, (Ed.). *Inteligencia Artificial en América Latina y el Caribe. Ética, Gobernanza y Políticas.* Buenos Aires: CETyS Universidad de San Andrés.

Carlos Amunátegui Perelló [1*]
Raúl Madrid [2**]

Resumen

El presente artículo intenta realizar una aproximación al fenómeno del sesgo (bias) generado a través de redes neuronales, sea en su entrenamiento, sea en el diseño de su función de éxito (objective function) y analizar algunas de sus posibles implicancias jurídicas.

[1*] *Profesor de Inteligencia Artificial, Pontificia Universidad Católica de Chile. Investigador del Programa de Ciencia y Tecnología de la Facultad de Derecho.*

[2**] *Director del Programa de Derecho, Ciencia y Tecnología de la Facultad de Derecho de la Pontificia Universidad Católica de Chile.*

En colaboración con: fAIr LAC | BID

1 Introducción

A partir de las primeras redes neuronales profundas (*deep neural networks*), desarrolladas en 1987 (Rumelhart et al., 1986, 533-536.), el desarrollo de sistemas de inteligencia artificial que se basan en ellas se ha hecho cada vez más dependiente de la acumulación masiva de los datos con los que se alimentan los algoritmos que se diseñan (Huang et al., 2013), a tal punto que hoy es más relevante la copiosa alimentación de datos que su elegancia o eficiencia (Lee, 2018, 14).

Por causa de esta creciente dependencia, se ha agudizado lo que llamamos el problema del "sesgo". Esta voz española ha sido utilizada para traducir a nuestra lengua el concepto anglosajón de "*bias*" aunque, en realidad, en español, se ajustaría mejor el término "prejuicio". En efecto, la palabra "*bias*" es definida por el *Cambridge English Dictionary* como "la acción de apoyar u oponerse a una persona o cosa en particular de manera injusta, debido a que permite que las opiniones personales influyan en su juicio"[3]. El sentido negativo es evidente si se compara con la definición que la Real Academia de la Lengua Española ofrece para la palabra "prejuzgar": "juzgar una cosa o a una persona antes del tiempo oportuno, o sin tener de ella cabal conocimiento"[4]. En ambos casos la acción aparece como un acto repudiable, ya sea porque se actúa con dolo, con un (aparente) defecto del pensamiento, o simplemente con ausencia de responsabilidad. El término "sesgo", en cambio, resulta más sutil al expresar la misma idea de un modo general, por cuanto en su forma adjetiva apunta a algo torcido, cortado o situado oblicuamente (Barcia, 1889). Esta idea de lo oblicuo también está presente en la voz anglosajona "*bias*", aplicada a la idea del error en el juicio sobre algo, pero precedida del término "prejuicio" (Roget y Davidson, 2003, 204).

3 *Cambridge English Dictionary, voz "bias". Disponible en: https://dictionary.cambridge.org/-dictionary/english/bias La traducción es nuestra.*

4 *Diccionario de la Real Academia de la Lengua Española, voz "prejuzgar".*

En colaboración con: fAIr LAC | BID

Desde el punto de vista que aquí nos ocupa, el concepto de "sesgo" indica precisamente la posibilidad de existencia de un prejuicio en el sentido apuntado antes, toda vez que los datos acumulados para la alimentación de algoritmos pueden estar afectos a toda clase de presunciones, prevenciones o preconcepciones, sean estas conscientes o inconscientes. Estas ideas preconcebidas pueden derivar del proceso de recolección o de acumulación de información o del diseño del algoritmo, específicamente de su función de éxito (*objective function*), lo cual puede derivar en casos de discriminación arbitraria realizada por el algoritmo, incluso sin que quienes lo han diseñado sean conscientes de esta parcialidad.

En realidad, desde una perspectiva jurídica, el problema es más profundo, toda vez que el Derecho no solo limita la capacidad de los operadores jurídicos para tomar decisiones basadas en ciertos criterios (como la raza, por ejemplo)[5], sino que incluso, en ocasiones, obliga a tomar decisiones basadas en criterios explícitamente formulados, como la peligrosidad o como la posibilidad de fuga en la libertad provisional[6].

Adicionalmente, existen problemas e incompatibilidades que se generan entre las redes neuronales y el sistema jurídico en general como cuerpo de reglas. En este sentido, intentaremos dilucidar las particularidades del sesgo en las redes neuronales y sus posibles causas para luego establecer los problemas y las incompatibilidades que se generan entre estas y el sistema jurídico en general como cuerpo de reglas.

5 Así, la Constitución Política de la República (CPR), en su artículo 19 Nº 2, establece el derecho a la no discriminación arbitraria. Este derecho ha sido desarrollado en diversos aspectos por la ley 20.609, que en su artículo segundo establece como definición de discriminación arbitraria lo siguiente: "Artículo 2º.- Definición de discriminación arbitraria. Para los efectos de esta ley, se entiende por discriminación arbitraria toda distinción, exclusión o restricción que carezca de justificación razonable, efectuada por agentes del Estado o particulares, y que cause privación, perturbación o amenaza en el ejercicio legítimo de los derechos fundamentales establecidos en la Constitución Política de la República o en los tratados internacionales sobre derechos humanos ratificados por Chile y que se encuentren vigentes, en particular cuando se funden en motivos tales como la raza o etnia, la nacionalidad, la situación socioeconómica, el idioma, la ideología u opinión política, la religión o creencia, la sindicación o participación en organizaciones gremiales o la falta de ellas, el sexo, la maternidad, la lactancia materna, el amamantamiento, la orientación sexual, la identidad de género, el estado civil, la edad, la filiación, la apariencia personal y la enfermedad o discapacidad."
Tal criterio ha sido afirmado por numerosas sentencias de la Corte Suprema (CS), entre las que podemos citar la reciente de 23-10-2017 en autos "Huerta con Sociedad Plaza", Rol 2847-2017, considerando sexto:
"[L]a vida en sociedad implica la existencia de diferenciación, en tanto supone elecciones que se determinan en la cotidianeidad; sin embargo, lo que la ley proscribe, es la ***discriminación*** *arbitraria, esto es, aquella distinción carente de racionalidad, que no tiene otra justificación que no sea el mero capricho de quien la ejecuta.*
Es en este contexto que el artículo 2º antes trascrito entrega criterios orientadores que permiten asentar una ***discriminación*** *arbitraria, entre los que se encuentra la discapacidad. En efecto, cualquier distinción, exclusión o restricción de derechos que se realice entre las personas sobre la base de la discapacidad, que no tenga sustento en la razón, constituye el acto que la ley sanciona."*

6 (CPR 19, Nº7, e)) "La libertad del imputado procederá a menos que la detención o prisión preventiva sea considerada por el juez como necesaria para las investigaciones o para la seguridad del ofendido o de la sociedad. La ley establecerá los requisitos y modalidades para obtenerla."

En colaboración con: fAIr LAC | BID

2. Redes neuronales y correlaciones

De un modo general, una red neuronal artificial es una red de elementos simples (nodos) dotados de organización jerárquica interconectada, masivamente y en paralelo, que busca interactuar con los objetos del mundo real del mismo modo que lo hace funcionalmente el sistema nervioso biológico (Gurney, 1997, 13). En principio, las redes neuronales son configuradas por los programadores, quienes determinan la arquitectura del modelo. Seguidamente, estas redes se entrenan en un conjunto de datos, a través de los cuales el modelo determina los pesos que asignará a las diversas conexiones entre sus capas, y, en buena medida, configura sus capas ocultas. De este modo, para funcionar eficientemente y detectar las correlaciones entre los datos con que se entrena, una red neuronal requiere de una gran cantidad de ejemplos. Así, los algoritmos formulados sobre la base de redes neuronales tienen una alta dependencia de los datos que los alimentan. Una red neuronal, para decirlo en términos simples, se entrena con un número de ejemplos tomados de una base de datos, a partir de los cuales corregirá los pesos (*weights*) de sus axiones mediante retropropagación, siguiendo usualmente un descenso estocástico de la gradiente inversa de la integral[7]. Sea cual sea el método que se adopte, que puede ser el clásico aprendizaje supervisado (*supervised learning*) o uno no supervisado, como las redes adversariales generativas (GANs)[8], las redes neuronales dependerán siempre de la calidad de los datos con que son alimentadas, puesto que a partir de estos datos los modelos generados desarrollan eficiencia y eficacia.

Si las redes neuronales no funcionaron correctamente hasta la década de 2010, esto se debió principalmente a la pobreza relativa de datos a su disposición, que solo fue suplida por las enormes bases de información generadas a través de Internet (Kai-Fu, 2018). Así, las tecnologías actuales de inteligencia artificial son completamente dependientes de los datos y su eficacia se encuentra significativamente determinada por ellos. De hecho, la abundancia de datos que ha generado Internet es el factor clave en la efectividad de los actuales modelos de redes neuronales, ya que gracias a la existencia de cientos de miles de registros de fotografías, videos, escrituras y búsquedas, se entrenan los actuales algoritmos y se configuran los pesos que vinculan sus diversas capas. Nada es tan eficiente como contar con cada vez más datos, de manera que incluso un algoritmo de diseño pobre, entrenado en una base de datos enorme, funciona mejor, es decir, tiene mayor capacidad predictiva que uno más refinado pero entrenado con una base de datos más modesta. Contar con cientos de miles de millones de datos de toda índole ha permitido que las grandes compañías de Internet[9] se conviertan en centros de desarrollo y de puesta en práctica de esta tecnología.

7 *Vid al respecto: Hinton y Salakhutdinov, 2006, 504-507.*

8 *Vid al respecto: Radford et al., 2016.*

9 *Nos referimos a las siete mayores compañías que dominan la red y que tienen las mayores inversiones en inteligencia artificial, a saber, Alphabet (matriz de Google), Alibaba, Amazon, Apple, Facebook, Baidu y Tencent, que tienen una posición dominante en la acumulación de datos y en el tráfico de la red.*

CeTyS Universidad de San Andrés
En colaboración con: fAIr LAC | BID Banco Interamericano de Desarrollo

La expansión de la *Internet de las cosas* ha generado además un auge de los datos disponibles, por cuanto hoy en día la generación de datos se extiende a muchos más elementos que antes, abarcando las imágenes emitidas por las cámaras de automóviles, la geo localización vía GPS y todos los datos que emiten los dispositivos dotados de conectividad a Internet en general, cualquiera sea su naturaleza. Puesto que la abundancia de datos optimiza los algoritmos, las compañías que dispongan de la mayor cantidad de datos, tendrán los mejores algoritmos, en el sentido de los más eficientes, y por ende, los mejores productos con funciones de inteligencia artificial[10]. La concentración de los datos ha generado tensiones, puesto que implica el establecimiento de un oligopolio natural en la tecnología entre las grandes compañías de Internet.

Evidentemente, la dependencia de los datos plantea un problema mayor a la hora de evaluar los resultados que tales modelos presentan, en la medida en que la calidad de sus resultados parece depender de los datos con los que las mencionadas redes son entrenadas. Así, si una red es entrenada con datos que presentan algún tipo de sesgo (racial, sexual, o de cualquier otro tipo), los resultados generados por la red expresarán el mismo sesgo, no solo reproduciéndolo, sino incluso ampliándolo[11] e institucionalizándolo en su propio ámbito de aplicación.

Los sesgos que un modelo puede adquirir en relación con los datos con que es entrenado suelen clasificarse en, al menos, tres tipos: sesgo de interacción (*interaction bias*), sesgo latente (*latent bias*) y sesgo de selección (*selection bias*). El primero consiste en que el propio usuario o programador inadvertidamente introduce un sesgo en el modelo por la manera en que interactúa con él. Esto se da, por ejemplo, al definir la función de éxito (objective function) del modelo, o el objeto que se busca.

10 *Esto es lo que se ha denominado Matthew effect, puesto que aquellos que más tengan recibirán más. Vid: Pasquale (2015, 82).*

11 *V id: O'Neil (2016, 23).*

CeTyS Universidad de San Andrés
En colaboración con:
fAIr LAC | BID Banco Interamericano de Desarrollo

El sesgo latente, en cambio, tiene lugar cuando el modelo realiza correlaciones inapropiadas, generalmente al establecer falsos nexos entre puntos de datos. Así, por ejemplo, si las personas de poca solvencia no pagan sus créditos, y la falta de solvencia puede correlacionarse con la probreza y esta con la segregación espacial en la ciudad, un algoritmo puede tomar la residencia de una persona en un punto de la ciudad segregado como indicador de su alto riesgo crediticio, lo cual puede ser perfectamente falso. En cuanto al sesgo de selección, este tiene lugar cuando la base de datos no es suficientemente representativa de la diversidad existente en el medio social. Así, por ejemplo, si un algoritmo se entrena para determinar elementos que hagan predecir habilidades en una población para ser buenos jugadores de fútbol con los datos médicos de todos los jugadores de la primera división argentina, probablemente dicho algoritmo sea inútil para realizar esa predicción en Japón, debido a su baja representatividad de población asiática.

A la hora de evaluar jurídicamente estas aplicaciones, el problema del sesgo (*bias*), que acabamos de describir, resulta muy relevante. Como ya se ha insinuado, un algoritmo tiene la calidad que le otorgan los datos con que se entrena; lo único que hacen las redes neuronales es establecer correlaciones entre datos, pero las conclusiones a las que lleguen dependerán siempre de la naturaleza de los datos que se le hayan entregado. Por ejemplo, si un algoritmo debe entrenarse para detectar diversas razas de perros, y los canes domésticos aparecen siempre en contextos hogareños, mientras que los lobos figuran en contextos naturales, puede que el simple hecho de figurar un perro en un contexto silvestre (un bosque nevado), implicará que la red neuronal lo asocie a los lobos, como de hecho ha ocurrido[12]. Existen muchos casos en que un algoritmo comete errores que están implícitos en los datos desde donde se entrena para realizar sus correlaciones.

Al respecto pueden mencionarse algunos casos que son francamente preocupantes. Un ejemplo de esto ocurrió con el programa de reconocimiento facial de Facebook. Joy Buolamwini (Buolamwini y Gebru, 2018) determinó que el algoritmo para el reconocimiento de rostros de esta compañía era deficiente a la hora de reconocer la cara de una persona negra, puesto que simplemente no detectaba su rostro. Incluso, un poco antes, en 2015, el programa de reconocimiento facial de Google había identificado a dos hombres negros como gorilas (Zhang, 2015), por lo que la compañía debió pedir disculpas públicamente. Estos casos parecen haberse generado porque las bases de datos a partir de las cuales Google y Facebook entrenan a sus algoritmos no contenían suficientes datos de personas negras, lo que provocó un serio problema a la hora de aplicarlo.

No obstante lo anterior, existen problemas mucho mayores relativos a los posibles sesgos que un algoritmo puede contener. Un ejemplo de esto es el caso del algoritmo predictivo de reincidencia COMPAS utilizado en algunos lugares de Estados Unidos. Puesto que la discriminación arbitraria es un problema agudo en el mundo entero, en algunos estados del país del norte se comenzó a utilizar un algoritmo que aconseja al juez a la hora de elaborar sus sentencias y de otorgar libertades provisionales, todo ello de acuerdo a los riesgos de reincidencia que los acusados o condenados presentaban en cada caso. El objetivo era generar decisiones basadas en los datos y, por tanto, ajenas a los riesgos relativos a los prejuicios humanos. Sin embargo, el resultado fue francamente inquietante. El algoritmo, tomando en cuenta cuestionarios psicológicos rellenados por los reclusos, sistemáticamente recomendaba penas más largas y preveía

12 *El asunto ocurrió en 2017 y muestra la fragilidad de las correlaciones realizadas por las redes neuronales. Véase un artículo de prensa donde se analiza el caso: http://innovation.uci.edu/2017/08/husky-or-wolf-using-a-black-box-learning-model-to-avoid-adoption-errors/*

CeTyS Universidad de San Andrés
En colaboración con: fAIr LAC | BID Banco Interamericano de Desarrollo

en contra de otorgar la libertad provisional a personas de color, aunque hubiesen cometido el mismo delito y tuviesen los mismos antecedentes que personas blancas. En otras palabras, distinguía en cuanto a otorgar la libertad en razón de la raza, un criterio universalmente prohibido[13] y considerado discriminatorio y arbitrario[14]. Es muy interesante que el resultado obtenido resulte exactamente el opuesto al buscado, puesto que en lugar de llegar a una justicia ciega, el velo que cubre los ojos de la diosa pareció caer por completo y hacerse parcial. El problema, como ya hemos dicho, parece encontrarse en el hecho de que los algoritmos tomen la información y establezcan correlaciones a partir de los datos con que son alimentados, por lo que si estos datos contienen un sesgo, que puede provenir en varias ocasiones de condicionamientos históricos, dicho sesgo será reproducido por el modelo, e incluso amplificado, proyectándolo hacia el futuro[15].

Otro caso interesante se da en la selección de currículos. Puesto que la práctica de seleccionar currículos es bastante ardua, especialmente en áreas donde los postulantes pueden ser muchos, es posible automatizar esta labor a fin de confeccionar una lista corta de candidatos respecto a los cuales se tome una decisión. Un criterio comúnmente utilizado para diseñar estos algoritmos es considerar como buenos empleados a aquellos que han permanecido en la compañía un número determinado de años y que han obtenido una promoción. Los currículos recibidos se comparan con estos modelos. Los resultados han sido fallidos en varias oportunidades, ya que en muchas compañías, y especialmente en las tecnológicas, los hombres predominan frente a las mujeres, por lo que los algoritmos tienden a desechar los currículos de las mujeres que postulen a dichos cargos solo por el hecho de ser mujeres[16].

13 Así, la Declaración Universal de los Derechos Humanos de 1948 establece en su artículo 2:
"Toda persona tiene todos los derechos y libertades proclamados en esta Declaración, sin distinción alguna de raza, color, sexo, idioma, religión, opinión política o de cualquier otra índole, origen nacional o social, posición económica, nacimiento o cualquier otra condición. Además, no se hará distinción alguna fundada en la condición política, jurídica o internacional del país o territorio de cuya jurisdicción dependa una persona, tanto si se trata de un país independiente, como de un territorio bajo administración fiduciaria, no autónomo o sometido a cualquier otra limitación de soberanía."

14 La investigación respecto al caso fue destapada por Propublica, una organización sin fines de lucro que analiza los usos de la inteligencia artificial, entre otras cosas. Disponible en: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
Consultado el 12 de junio de 2019.

15 El asunto resultó en una demanda y fue determinada la existencia de sesgo en el algoritmo. Véase: State v. Loomis, 881 N.W.2d 749 (Wis. 2016). El caso se encuentra analizado en: Recent Cases, en Harvard Law Review, 130, 2017, pp. 1530-1537.

16 Los casos en que esto ha ocurrido son muchos. Como muestra, véase el reportaje de Reuters respecto a Amazon, disponible en: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
Consultado el 12 de junio de 2019.

CeTyS Universidad de San Andrés
En colaboración con: fAIr LAC | BID Banco Interamericano de Desarrollo

Esto nos lleva a otros problemas relativos a las correlaciones que los algoritmos establecen, puesto que incluso con grupos de datos que no parecen contener sesgos, el resultado puede resultar socialmente dañino. En este sentido, se puede formular fácilmente un caso con los datos de los supervivientes del Titanic[17]. Se trata de un ejercicio simple de formulación de un algoritmo neuronal que suele recomendarse para principiantes en programación. Sobre la base de una lista de sobrevivientes del Titanic que contiene algunos datos tomados de los registros originales, se diseña un modelo que predice las probabilidades de sobrevivir a una catástrofe marítima. Su resultado es que los factores más relevantes para sobrevivir al choque contra el témpano de hielo son: ser mujer y viajar en primera clase. Por tanto, si diseñamos un algoritmo predictivo sobre la base de ese modelo para, por ejemplo, una compañía de seguros, los hombres pobres pagarían mucho más por representar un riesgo más alto en el transporte de pasajeros náutico que las mujeres ricas. Este tipo de algoritmos son relativamente simples de construir y están siendo utilizados sin nuestro conocimiento en áreas como los seguros de salud y los créditos bancarios, lo cual resulta inquietante, por su fundamental arbitrariedadc[18]. ¿Deberán los pobres pagar más intereses por el solo hecho de ser pobre, no obstante disponer de un historial crediticio impecable?

Un problema de fondo que implican los sistemas de redes neuronales es que sus resultados son siempre extensos juegos de correlaciones. Incluso si las bases de datos utilizadas son apropiadas y no implican, en sí mismas, la existencia de sesgos importantes, toda vez que las redes neuronales establecen una correlación entre puntos de datos y resultados, puede que la correlación establecida contenga una confusión de elementos estadísticos (*confounding*). Esto puede deberse a múltiples motivos, como por ejemplo que un mismo factor provoque varios efectos no relacionados entre sí, o que existan colisionadores (*colliders*) estadísticos, ante lo cual la red puede interpretar que uno de los efectos es dependiente estadísticamente del otro, sin tomar en consideración que ambos dependen de un terceroc[19]. Así, en un ambiente de alta segregación, determinados nombres comunes en ciertas minorías pueden estar relacionados con la pobreza, mientras que la miseria podría estar relacionada con la falta de poder adquisitivo, y esta última con un riesgo mayor de insolvencia. Al no tener un razonamiento causal y no ser capaz de distinguir los procesos sociales adecuadamente, la red neuronal podría correlacionar determinados nombres propios con la insolvencia y denegar el acceso al crédito a las personas que presenten esos nombres, con independencia de su poder adquisitivo, historial crediticio o riesgo real de insolvencia. En pocas palabras, la existencia de sesgo en los sistemas de redes neuronales no es solamente un problema de calidad de datos, sino que también puede deberse a un riesgo estructural de una arquitectura no diseñada para la detección de causas y efectos.

17 *El ejercicio fue diseñado y analizado por Meredith Broussard. Vid: Broussard (2018, l. 2150).*

18 *Frank Pasquale (2015, 38) pinta oscuros colores sobre la materia: "Reputation systems are creating new (and largely invisible) minorities, disfavored due to error or unfairness. Algorithms are not immune from the fundamental problem of discrimination in which negative and baseless assumptions congeal into prejudice".*

19 *Sobre el particular, vid: Pearl y Mackenzie (2018, 135 y ss.).*

CeTyS Universidad de San Andrés
En colaboración con: fAIr LAC | BID Banco Interamericano de Desarrollo

3. El Derecho en Chile y las correlaciones

Esto nos lleva a un problema importante, que consiste en que el Derecho, como tal, es un sistema de inferencias reglado. Así, puede asumirse que un vendedor debe cumplir con su obligación de entregar la cosa vendida sobre la base de ciertas reglas pre-establecidas en el Código Civil, como son la existencia de un contrato de compraventa, el grado de cumplimiento del comprador de su propia obligación de pagar el precio, la existencia de vicios y demás. Un algoritmo puede bien inferir la existencia de la obligación de entregar la cosa (o la falta de ella), de acuerdo con cualquier otro factor incluido en los datos con que fue entrenado como, por ejemplo, el nombre de uno de los contratantes, a lo cual el algoritmo puede dar una importancia superlativa dados los datos con que cuenta. Aunque sería extraño que lo hiciese, es perfectamente posible[20]. Si este algoritmo está diseñado para recomendar al juez una sentencia o directamente para decidir un conflicto, el resultado puede ser perfectamente ilegal y arbitrario. En principio esto podría prevenirse con mejores datos o directamente desechando las recomendaciones realizadas por algoritmos que no se basen en correlaciones lícitas, pero esto no es fácil de determinar, ya que los modelos de redes neuronales tienden a ser opacos y poco explicables.

20 Para explicitar mejor este ejemplo, podemos imaginar una compañía que en varias oportunidades ha sido demandada por haber incumplido en su obligación de entregar la cosa vendida, por ejemplo una multitienda, de lo cual el algoritmo correlacione el nombre de la multitienda demandada con la condena a entregar la cosa vendida, aunque en el caso particular no se den los supuestos jurídicamente necesarios para ello.

En colaboración con:
CeTyS Universidad de San Andrés
fAIr LAC
BID Banco Interamericano de Desarrollo

En principio, las redes neuronales no son capaces de proveer por sí mismas explicaciones de sus predicciones. Es decir, hasta el momento resultan incapaces de dar razón de ellas, de explicar de modo suficiente el fundamento de su decisión. A diferencia de los viejos sistemas expertos basados en principios lógicos[21], no existe una cadena de asunciones perfectamente trazables y comprensibles, sin perjuicio de que los sistemas jurídicos expertos tampoco disponían de las herramientas para justificar conclusiones, en un sentido no meramente genealógico. Para individualizar las razones detrás de una determinación usando redes neuronales, es necesario que un analista de datos establezca a qué conexiones neuronales se está dando importancia y, dada la complejidad de algunos de estos, que pueden llegar a tener miles de capas ocultas, esto puede ser perfectamente imposible. En otros términos, el factor capaz de formular la justificación del acto sigue siendo un individuo humano, con todo el componente cognitivo y hermenéutico que le es propio a un sujeto concreto.

Esto nos lleva a una nueva cuestión, relacionada con lo anterior: la opacidad de los sistemas de *deep learning* basados en redes neuronales. De momento, no existe un mecanismo simple que permita determinar con certeza las correlaciones que un algoritmo realiza y la fuerza que cada elemento que puede emerger en una capa de una red tendrá finalmente. Para esto se requiere el trabajo de analistas de datos y, aun en ese caso, puede resultar difícil o prácticamente imposible. Usualmente, para determinar si un modelo tiene algún sesgo importante, lo más simple es aplicarlo en un caso en el que tal sesgo no exista y en otro en que sí y ver si ambos casos son tratados de la misma manera, aunque hay un grado importante de incerteza en este procedimiento. Esto se complejiza aún más si tenemos en cuenta que las compañías que diseñan o utilizan algoritmos no suelen compartirlos, puesto que son el secreto comercial que permite que sus asociaciones desarrollen un tráfico mercantil importante y, en caso de hacerse públicos, las compañías perderían rápidamente valor. En pocas palabras, tales algoritmos no solo resultan opacos en la toma de decisiones, sino que además su funcionamiento suele ser un secreto comercial.

La Unión Europea, en su Guía Ética para una Inteligencia Artificial Confiable, publicada el 9 de abril de 2019, exige que el diseño de sistemas de inteligencia artificial se guíe por el principio de explicabilidad (Unión Europea, 2019, 13), esto es, que en la medida de lo posible, los algoritmos sean transparentes en cuanto a las decisiones que toman. No obstante, no es claro que esto sea técnicamente posible. Actualmente se trabaja en dicho problema y hay varias alternativas al respecto, aunque no se puede determinar con certeza hasta dónde se puede llegar en este camino.

Algunos autores estiman que la inteligencia artificial se convertirá en un importante problema de derechos humanos durante este siglo (Noble, 2018, 1). Este aspecto cobra mucha mayor importancia cuando el campo en que se aplica la inteligencia artificial no es simplemente un asunto

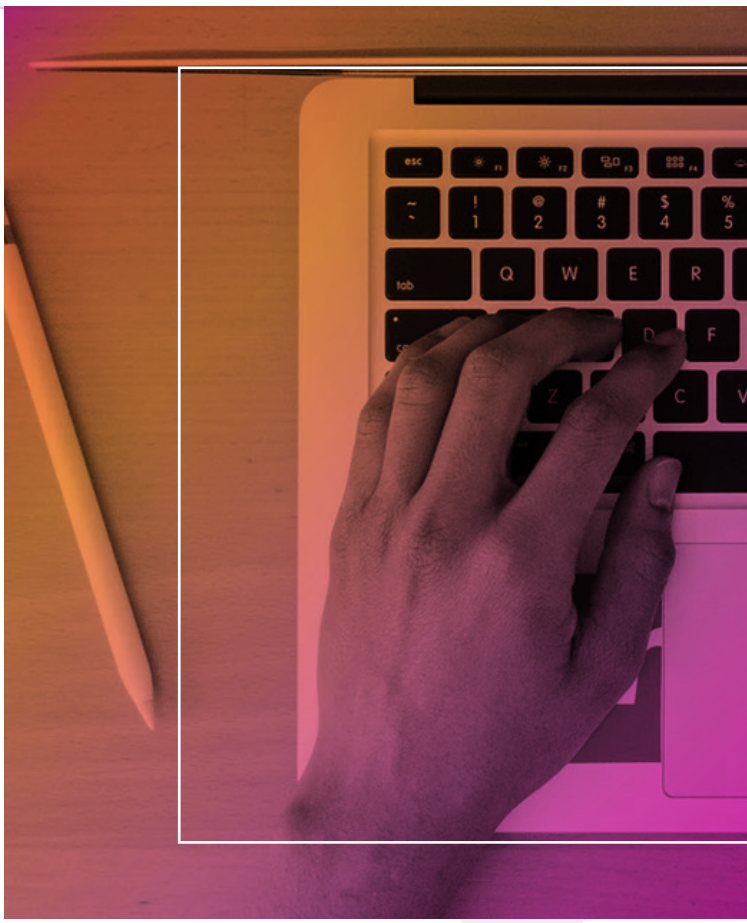
21 Nos referimos a los sistemas conectivistas que primaron desde la década de 1970 hasta el desarrollo de las modernas redes neuronales. Dichos modelos reducen los problemas de decisión a un conjunto de inferencias que se realizan a partir de reglas heurísticas programadas manualmente. Dichos sistemas aún se encuentran en uso, aunque tuvieron su hegemonía en la década de 1980 cuando se desarrollaron los llamados "sistemas expertos". También de los conoce como GOFAI, por sus siglas en inglés (good old fashion AI).

En colaboración con:
fAIr LAC | BID

contractual, sino que se emplea en el desarrollo de una función de carácter público, como puede ser la administración de justicia o la provisión de servicios sociales. En principio, el actuar del aparato público debe basarse en reglas pre-establecidas y objetivas, que trate a todas las personas que están en una misma condición de manera equivalente. En esto consiste la igualdad ante la ley, que verdaderamente es un deber que la Constitución impone a la actuación del Estado (19 Nº2 CPR). Esta igualdad implica también un trato no discriminatorio no solo del aparato público, sino también de los particulares en sus relaciones entre sí, de manera que el establecimiento de diferencias injustificiables en motivos racionales son contrarias al orden jurídico político. ¿Un mecanismo fundamentalmente inexplicable en sus decisiones puede cumplir con estos deberes implicados en la noción de igualdad ante la ley y prohibición de la discriminación? Es más, ¿puede un algoritmo que fundamenta sus decisiones en correlaciones de datos oscuras para el administrado y para el propio administrador, tomar decisiones que cumplan con los requisitos básicos de fundamentación y de transparencia que exige nuestro orden jurídico para el actuar de órganos públicos?

Un algoritmo no es más inteligente que los datos con que se alimenta[22] y, si estos contienen un sesgo, entonces el algoritmo adquirirá el sesgo de los mismos datos con que fue entrenado. Este es un problema de sobregeneralización (*overgeneralization* u *overfitting*, en terminos estadísticos), en que el algoritmo detecta patrones en los datos –en ocaciones invisibles para el propio programador– y luego los reproduce y amplifica (Surden, 2014, 106).

Un segundo elemento llamativo es que las decisiones del algoritmo están determinadas por su definición de éxito. Y como esta se encuentra determinada a priori por los



diseñadores del modelo, si no incluye algún tipo de corrección explícita de los sesgos del pasado, el modelo simplemente los reproducirá. Para corregir los problemas de discriminación que un conjunto de datos puede presentar, es necesario que los programadores sean conscientes de su existencia y activos en su corrección, de lo contrario la pretendida objetividad que pretenden lograr no solo no se dará, sino que es perfectamente posible que el modelo profundice los sesgos. El agente se limitará a reproducir los sesgos que encuentre en los datos, produciendo resultados moralmente cuestionables y jurídicamente inaceptables.

Un elemento inquietante se relaciona con la política de precios y condiciones de contratación que tales algoritmos pueden generar. Estabeciendo perfiles para los distintos usuarios, pueden establecerse distintos precios para los diversos demandantes de un servicio. Amazon admitió, ya en el año 2000, utilizar esta política (Broussard, 2018, l. 2128) y parece ser relativamente común en el ciberespacio. Un

22 *O puesto de otra manera, el ser humano es más listo que los datos. Vid: Pearl y Mackenzie (2018, 21).*

CeTyS Universidad de San Andrés
En colaboración con: fAIr LAC | BID Banco Interamericano de Desarrollo

efecto de ese mecanismo es que las personas con mayor poder adquisitivo tienden a recibir ofertas de productos a precios más bajos que aquellas que cuentan con menos recursos (Broussard, 2018, l. 2128). A partir de las búsquedas que una persona realiza, de las opiniones manifestadas en las redes sociales o de cualquier otro conjunto de datos, puede diseñarse un algoritmo que detecte las posibles necesidades de un individuo en concreto y enviarle ofertas personalizadas, en condiciones distintas a las ofertadas para los demás internautas. El servicio AdSense de Google Mail detecta patrones en el contenido de los correos electrónicos que cada usuario recibe para mostrarle publicidad personalizada. Al margen de los problemas que tal servicio puede implicar por violación del secreto de la correspondencia (Chopra y White, 2011, 110), dicho mecanismo genera, a partir de un conjunto de datos, un perfil del usuario y ofrece bienes y servicios específicos para él.

La pregunta en esta instancia es si esas ofertas pueden estar afectas a algún sesgo, y si la política de precios diferenciados es acorde con la protección al consumidor[23]. Respecto al primer problema, la posible existencia de un sesgo en la generación de una política de precios dependerá verdaderamente del caso y no podemos dar una respuesta a priori. Puesto que las decisiones algorítmicas son verdaderas cajas negras, no podemos estar seguros de los motivos de tal o cual modelo en particular para segregar su política de precios. Puede que la base de datos contenga un sesgo o que la función de éxito esté incorrectamente diseñada, en cuyo caso el resultado será discriminatorio. Lo importante es analizar el caso en concreto que se juzga sospechoso, comparar los resultados que arroja el algoritmo si se incorpora o se excluye el elemento que pueda causar discriminación, así como también atender a las grandes tendencias que emergen al estudiar los precios ofrecidos. Evidentemente, si los resultados son explicables por diferencias arbitrarias, como el sexo, la raza u otro factor similar, nos encontraremos ante un caso de discriminación algorítmica y se activará la posibilidad de ejercer acciones constitucionales como el recurso de protección en Chile, o de protección al consumidor[24], pero por la naturaleza oscura de este tipo de algoritmos, esto requiere de un análisis especializado. Debe recordarse que un algoritmo puede resultar discriminatorio aun cuando no tome entre sus puntos de datos el factor que origina la discriminación. En el caso del modelo que seleccionaba currículos de Amazon[25], el sexo no estaba entre los elementos que el algoritmo consideraba. Ahora bien, el resultado de su aplicación era la exclusión de las mujeres toda vez que el algoritmo consideraba positivamente actividades que los hombres suelen desarrollar (como integrar un equipo de fútbol americano) y neutrales o negativas aquellas que suelen realizar mujeres (ser *chearleader*, por ejemplo). De ahí que, aunque el agente no tuviese una consideración explícita del sexo, el resultado era la discriminación por sexo.

Ahora bien, suponiendo que la política de precios algorítmica se encuentra fundada en criterios no discriminatorios a priori, la pregunta que queda abierta es si es aceptable que diversos consumidores reciban distintos

23 Este problema fue abordado por Zuiderveen Borgesius (2018) en un informe para el Consejo de Europa.

24 Se infringiría el artículo 3 letra c de la Ley 19496 que fija el derecho del consumidor a no ser discriminado arbitrariamente.

25 Este ejemplo también es recuperado en el artículo de Gómez Mont, Constanza; May Del Pozo, C.; Martín del Campo, Ana Victoria Economía de datos e inteligencia artificial en América Latina de este volumen.

En colaboración con:
CeTyS Universidad de San Andrés
fAIr LAC
BID Banco Interamericano de Desarrollo

precios. Teóricamente existe un precio general (más alto) que se ofrecería a la mayor parte de los consumidores, pero en la práctica, puesto que muchísimos consumidores al buscar diversos productos reciben precios distintos, parece que el precio general no existe o existe solo para una minoría. Esto parece encontrarse reñido con lo señalado por la normativa de consumidores vigente hoy en Chile, que indica la obligatoriedad[26] de mantener visibles los precios ofertados. Ahora bien, puesto que la práctica común consiste en tener un precio de lista y ofrecer al consumidor personalmente uno más bajo a título de oferta, la dificultad parece subsanarse, toda vez que habría un precio general cognoscible por todo consumidor y uno especial ofrecido en concreto a algunos de ellos, pero queda en el aire la pregunta relativa a por qué dicho consumidor recibe un precio especial y los demás no. ¿Resulta esto discriminatorio, en los términos del artículo 3 letra c de la Ley 19496? Toda vez que los motivos del algoritmo para establecer en un caso concreto un precio distinto para un determinado consumidor –y no para otro– son correlaciones de datos de incongnoscible significado, no podemos saberlo, y sería necesario hacer un estudio acerca de su tendencia. En cualquier caso, lo generalizado de la práctica y la dificultad de encontrar motivos objetivos para cada caso, hace necesario discutirlo.

Más complejo se torna el caso de las ofertas de servicios financieros, materia en que comúnmente se aplican algoritmos de este tipo para realizar ofertas y establecer concidiones crediticias. Son muchos los puntos de datos que podrían eventualmente tomarse en consideración para establecer que una persona es solvente o no, entre los cuales está su historial de incumplimientos, su solvencia, y otros similares. Ahora bien, no sabemos qué elementos toma en cuenta el algoritmo que determina el riesgo de una persona en concreto pero, si ha sido construido sobre la base del *deep learning* y de bases de datos abiertas, puede estar afecto a criterios discriminatorios, como el lugar donde vive, características del nombre u otros factores igualmente problemáticos. En materia financiera, por lo demás, el artículo 3º letra a de la Ley 19496 establece que, en caso de rechazarse la contratación de un servicio financiero, el consumidor debe "ser informado por escrito de las razones del rechazo a la contratación del servicio financiero, las que deberán fundarse en condiciones objetivas." En este caso, vale la pena preguntarse si la determinación de un alto riesgo por parte de un algoritmo es razón suficiente, o si debe también informarse de los motivos que inducen al modelo a establecer ese nivel de riesgo en concreto.

26 *L. 19496, Artículo 30: "Los proveedores deberán dar conocimiento al público de los precios de los bienes que expendan o de los servicios que ofrezcan, con excepción de los que por sus características deban regularse convencionalmente.*
El precio deberá indicarse de un modo claramente visible que permita al consumidor, de manera efectiva, el ejercicio de su derecho a elección, antes de formalizar o perfeccionar el acto de consumo. Igualmente se enunciarán las tarifas de los establecimientos de prestación de servicios.
Cuando se exhiban los bienes en vitrinas, anaqueles o estanterías, se deberá indicar allí sus respectivos precios. La misma información, además de las características y prestaciones esenciales de los productos o servicios, deberá ser indicada en los sitios de Internet en que los proveedores exhiban los bienes o servicios que ofrezcan y que cumplan con las condiciones que determine el reglamento.
El monto del precio deberá comprender el valor total del bien o servicio, incluidos los impuestos correspondientes.
Cuando el consumidor no pueda conocer por sí mismo el precio de los productos que desea adquirir, los establecimientos comerciales deberán mantener una lista de sus precios a disposición del público, de manera permanente y visible."

En colaboración con: fAIr LAC | BID

De tomarse la segunda opción, que parece más acorde con el espíritu de la legislación, es difícil imaginar cómo podría fundamentarse la decisión del algoritmo. En efecto, los modelos crediticios han sido cuestionados, especialmente porque su determinación de riesgo –y, por tanto, del interés de un crédito– es ciertamente oscura para el consumidor[27].

Es importante señalar que en una sociedad donde históricamente la movilidad social fue baja y la pobreza y, por tanto, la falta de solvencia, han tenido históricamente componentes étnicos, como la nuestra, es posible que un modelo tome factores correlacionados con la pobreza, como el nombre[28], el lugar de nacimiento, el establecimiento educacional donde se cursó la primaria u otros similares, para construir un modelo que acreciente el riesgo de las personas con tales antecedentes y lo aminore respecto de otros que detentan factores usualmente correlacionados con el éxito[29]. En pocas palabras, un algoritmo podría fácilmente tomar factores que históricamente han fundado el prejuicio social y potenciarlos de manera completamente inaceptable[30]. Los programadores y los diseñadores del modelo pueden estar perfectamente inconscientes de este hecho, e incluso pueden estar buscando fines complemente contrarios, como promover a través de diseños matemáticos una sociedad más igualitaria y menos segregada y, no obstante, el agente producto de sus esfuerzos puede tender a la inmovilidad social, a la re-etnificación de la pobreza y al cierre de los grupos sociales con más medios económicos. Esto es lo que se ha denominado "Armas de Destrucción Matemática" (*Weapons of Math Destruction*)[31].

En ocasiones se ha señalado que, aunque los algoritmos estén abiertos a contener sesgos de diversa especie, como también lo están los seres humanos, estos sesgos serán siempre menores que aquellos que aplican las personas (Casey y Niblett, 2016, 437), por lo que sería importante no exagerar su riesgo. Discrepamos de tal opinión, ya que en lo que respecta a los seres humanos, podemos preguntarles sus motivos y, en ocasiones, exigirles jurídicamente que los manifiesten, mientras que no contamos con medios similares respecto a los agentes artificiales, que no son más que entidades inconscientes que manipulan mecánicamente símbolos a los que no son capaces de asignar significado.

27 Frank Pasquale (2015, 4) señala al respecto:
"A bad credit socre may cost a borrower hundreds of thousands of dollars, but he will never understand exactly how it was calculated." Véase también: Marron (2007, 111).

28 En efecto, ha habido casos en que se ha detectado sesgo en el algoritmo de predicción de resultados de Google. Vid: Pasquale (2015, 40).

29 Vid: Ramirez (2019).

30 La situación ha sido descrita en los siguientes términos:
"Mounting evidence shows that automated decision-making systems are disproportionately harmful to the most vulnerable and the least powerful, who have little ability to intervene in them- from misrepresentation to prison sentencing to accessing credit and other life-impacting formulas". Noble (2018, 49).

31 Cathy O'Neil (2016, 3) acuñó el término, y describe el problema en oscuros colores:
"Nevertheless, many of these models encoded human prejudice, misunderstananding, and bias into the software systems that increasingly managed our lives. Like gods, these mathematical models were opaque, their workings invisible to all but the hihest priests in their domain: mathematicians and computer scientists... they tended to punish the poor and the oppressed in our society, while making the rich richer".

En colaboración con:
CeTyS Universidad de San Andrés
fAIr LAC | BID Banco Interamericano de Desarrollo

Ahora bien, si el riesgo de un sesgo hace delicado el uso de agentes artificiales en el campo privado, en el mundo público esto es aún más difícil. Ha habido diversos esfuerzos tendientes a servirse de agentes artificiales a fin de automatizar acciones del gobierno, de manera que estos modelos realicen actos que usualmente se dejaban al criterio de algún ser humano, como un asistente social o un juez. Si tales mecanismos cuentan con un sesgo obtenido, sea de los datos, sea del diseño de su función de éxito, el resultado será una amplificación de la discriminación que experimentan los miembros más vulnerables de nuestra sociedad[32], esta vez distribuida de una manera automatizada por parte de la propia autoridad estatal que, teóricamente, está justamente encargada de combatirla.

Un ejemplo que merece nuestra atención consiste en un algoritmo de predicción de sentencias diseñado ya hace algún tiempo con el que experimentó el conocido grupo de investigación *Lex ex Machina*, entonces radicado en la Universidad de Standford (Surdeanu et al., 2019, 116-120). Este fue uno de los primeros modelos que se servían de *machine learning* para predecir los resultados de litigios relativos a marcas y a propiedad intelectual. Una vez realizado el análisis de sus correlaciones y de sus puntos de datos, se encontró que los factores más relevantes para predecir el resultado eran la identidad del juez y el nombre del equipo jurídico de las partes (Ashley, 2017, 124). El resultado es interesante porque da cuenta del valor que tiene el hecho de contar con un buen juez y con un equipo de abogados solvente. Pero, si se quisiera transformar este algoritmo en un sistema no predictivo, sino sentenciador, terminaría decidiendo los casos no en razón de sus méritos, sino de la identidad de los abogados, lo cual resultaría evidentemente ajurídico.

En relación con la evaluación del riesgo que representa el ofensor para la sociedad, es interesante señalar que parece apuntar a las caractirísticas de la persona, y no a los hechos que pueda haber cometido. La aserción de riesgo que realiza el modelo se fundamenta en quién es la persona, en cuáles son sus amistades, cuál es su crianza y su nivel socieconómico, antes que en la naturaleza de los hechos imputados y que en su conducta anterior. En esto, parece apuntar al Derecho penal de autor, que juzga a las personas por quiénes son y no por lo que hacen, lo cual es impropio en una sociedad democrática. Si se juzga que puede predecirse el riesgo de comisión de delitos con certeza, no es de extrañar que se depliegue una suerte de sistema preventivo en contra de las personas sospechosas de cometer delitos en el futuro. Por increíble

32 *Virginia Eubanks (2018, 11) describe la situación en Estados Unidos en los siguientes términos:*
"Across the country, poor and working-class people are targeted by new tools of digital poverty management and face life-threateing consequences as a result. Automated eligibility systems discourage them from claiming public resources that they need to survive and thrive."
adquirir, los establecimientos comerciales deberán mantener una lista de sus precios a disposición del público, de manera permanente y visible."

CeTyS Universidad de San Andrés
En colaboración con: fAIr LAC | BID Banco Interamericano de Desarrollo

que parezca, esto es exactamente lo que ha ocurrido (O'Neil, 2016, 102). Robert McDaniel fue visitado por la policía en 2013 sin haber cometido jamás un delito, porque el departamento de policía de la ciudad de Chicago decidió llevar adelante un programa de prevención algorítmica del delito y él fue seleccionado como una de las cuatrocientas personas que más probablemente cometería uno en el futuro cercano[33]. En China, de hecho, ya existe una política de prevención de futuros delincuentes[34]. Cuando, basado en determinados antecedentes evaluados algorítmicamente, se determina la posibilidad alta de futura comisión de un delito por parte de un ciudadano, este es detenido y enviado a un campo de "re-educación".

El punto más importante es que los jueces, al igual que los demás órganos del Estado, deben someter su acción a la Ley y a la Constitución (art. 6º CPR), por lo que sus actos deben estar enmarcados dentro de las normas jurídicas que nuestro ordenamiento establece, en la forma que dichos preceptos establecen. Esto implica que en cada caso concreto deberán revisar si la situación se adapta o no al tipo legal y, una vez verificado esto, aplicarán la sanción o consecuente jurídico pre-establecido por la norma al hecho evaluado. Si actúan guiados por correlaciones arbitrarias, sus actos serán reflejos de estas correlaciones y no constituirán, en sí mismos, aplicaciones de normas. Los algoritmos de *deep learning* no aplican normas, de hecho, los mecanismos de tipo conectivista no son capaces de entenderlas ni de manipular símbolos de manera de realizar una subsunción, por lo que serán siempre, esencialmente, ajurídicos. En este sentido, al servirse de ellos, se renuncia a la aplicación de reglas y, por tanto, al Derecho como tal. Esto no es exclusivo de los jueces, sino común a toda la administración del Estado en cuanto que el órgano ejecutivo debe aplicar el Derecho vigente. La existencia de un sistema jurídico implica la aplicación de normas, y una mera correlación matemática no es tal. La pregunta es dónde queremos vivir como sociedad, si en un lugar donde el Estado de Derecho impere o donde no lo haga. Desde la Antigüedad, ya Aristóteles (3.8-19, 1286a.) defiende el gobierno de las leyes frente a incluso el mejor de los hombres. El gobierno de las leyes es el fundamento de la democracia moderna y, en su esencia, la libertad consiste en poder hacer lo que las leyes permiten[35], de manera que renunciar a un sistema normativo que regule la conducta de los individuos en sociedad implica alejarse del concepto de libertad que ha servido de base a nuestra tradición jurídico política.

33 Véase el reporte de prensa en: https://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist Consultado el 21 de Octubre de 2019.

34 Véase información de prensa al respecto en The Washington Times disponible en: https://www.washingtontimes.com/news/2019/jun/25/chinas-chilling-pre-crime-prison-indoctrination-sy/ Consultado el 21 de Octubre de 2019.

35 Libertas in legibus consistit. Cic. De l. Agr. 2.102.

CeTyS Universidad de San Andrés
En colaboración con: fAIr LAC | BID Banco Interamericano de Desarrollo

Una cuestión final que presenta esta sección consiste en plantear un problema específico respecto al sesgo, que consiste en su corrección. En principio, una vez que se detecta que un algoritmo funciona con un sesgo, o incluso, al momento de diseñarlo y de establecer su función de éxito, pueden introducirse factores que lo mitiguen o, incluso, que lo hagan desaparecer. Por ejemplo, si el algoritmo de selección de currículos descarta sistemáticamente aquellos pertenecientes a mujeres, puede incorporarse dentro de la función de éxito el tender hacia una selección equilibrada entre hombres y mujeres. El principal problema de esta solución es que requiere identificar los sesgos que puede contener nuestra base de datos y ser cuidadosos a la hora de diseñar la función de éxito, toda vez que es necesario hacerlo conscientemente. En esto existen dos peligros, el primero es la posible pérdida de eficacia del sistema, y el segundo es la selección del conjunto de valores que quieren incorporarse a él.

Por último está el delicado problema de la corrección de los sesgos. Efectivamente, cuando se detecta un sesgo, en ocasiones (no siempre) es posible eliminarlo. Para ello usualmente debe examinarse la función de éxito del algoritmo. ¿Cuál es su finalidad? Si su finalidad es contratar empleados de características similares a los que ya trabajan en la empresa (en cuyo caso reproducirá los sesgos que contenga la plantilla), o modificar tal plantilla en algún sentido (aumentando su diversidad, disminuyendo el ausentismo laboral, o lo que sea). Si se quiere variar sobre la base de datos que alimenta al agente, será menester considerarlo expresamente y programar este tipo de elementos dentro de la función de éxito, pero es necesario poder visualizarlos para corregirlos. Si se prohibe, por ejemplo, indicar el sexo en el currículo, pero el algoritmo de selección de antecedentes, de manera inesperada, segrega en razón de sexo (utilizando sustitutos estadísticos de este) a las mujeres, como ocurrió en el caso de Amazon, es díficil corregir el sesgo sin tener un indicador explícito sobre la calidad de hombre o mujer del candidato. Amazon debió desechar su algoritmo simplemente porque la ausencia de mención sobre el sexo en los currículos impedía corregirlo de manera aceptable. Extrañamente, la corrección de sesgo implica la visualización de las inequidades que un grupo humano presenta, y en este sentido, es un ejercicio de sinceridad.

Ahora bien, una pregunta evidente en este problema es qué valores deben incorporarse dentro de un modelo a fin de que corrija los sesgos que genere. Esta es una decisión eminentemente política y moral. En un sistema constitucional axiológico, donde se incorporan valores al contenido normativo de la Constitución, la respuesta ha de ser, a lo menos para el aparato público, el sistema ético que respalda a la Constitución, espcialmente en su título primero, “Bases de la Institucionalidad”, y en el artículo 19, sobre los derechos de las personas. Ahora bien, ante el silencio, no es claro cómo proceder.

En cuanto al sector privado, las dudas son tal vez más relevantes. ¿Puede una empresa proyectar sus valores a la sociedad, suprimiendo términos y funciones que sean incompatibles con estos? ¿Hay un deber de neutralidad axiológica?

Desde al menos 2017, Google tiene una política expresa de justicia social en sus algoritmos a fin de corregir los sesgos que estos puedan contener, que se conoce como Machine Learning Fairness

(MLF)[36]. Esta incorpora expresamente políticas de no discriminación en la construcción de sus algoritmos, a fin de que den una visión más equilibrada de la sociedad en los resultados de las búsquedas que los usuarios practican. La idea es que no se refleje en ellas solo lo más popular, sino también las creencias y prácticas de los grupos menos favorecidos y minoritarios. Esta política ha sido puesta en cuestión en el último tiempo, toda vez que resulta una imposición a personas que no necesariamente comparten los valores de la compañía y, hasta cierto punto, distorciona la realidad de la red (Murray, 2019, l. 2102 y ss.). No obstante, evitaremos entrar en este problema por la extensión que requeriría su tratamiento detallado.

4 Conclusiones

A partir de nuestro análisis podemos concluir que los sistemas de inteligencia artificial basados en redes neuronales presentan variados riesgos a la hora de ser aplicados como mecanismos predictivos al mundo del Derecho.

El llamado "problema del sesgo" es lo que en el pensamiento continental se denomina "prejuicio" (la existencia de un juicio anterior a otro, que afecta o modifica el segundo), y remite a un debate de mucha mayor entidad teórica y antropológica: la pregunta sobre si el conocimiento que posee el ser humano se origina en una evidencia o es fruto de un método, es decir, de un proceso de demostración. La respuesta del mundo clásico, romano y medieval fue –cada uno a su modo– que el punto de partida del conocimiento era una cierta realidad exterior, constituida sustancialmente en sí misma, que se presentaba de modo fenoménico a la conciencia humana. Fruto de este encuentro con la existencia misma de las cosas exteriores, el intelecto podía luego pensar la realidad de modo consciente. En este contexto, el pre-juicio, es decir, el juicio anterior al pensamiento formal, no solo no constituía un defecto, sino que apuntaba a la condición misma de la estructura cognitiva humana. Es la Modernidad, a través de su giro metodológico y de su intención de aplicar al conocimiento humano las condiciones de "pureza" de las ciencias lógico-empíricas, la que convierte al prejuicio en un (supuesto) error del pensar. Así, la reflexión, para ser pura,

36 Véase el documento Responsible AI Practices, disponible en: https://ai.google/responsibilities/responsible-ai-practices/?category=fairness Consultado el 22 de Octubre de 2019.

En colaboración con: fAIr LAC | BID

adecuada o eficiente, debía carecer de contenidos previos al encuentro entre el intérprete y lo interpretado. En este contexto, surge el prurito de la eliminación de todas las preconcepciones, con objeto de alcanzar la pureza metodológica de la decisión. Sin embargo, ello es imposible, porque todo juicio humano es un acto de la conciencia y la conciencia es fundamento de identidad: nada puede separar la conciencia de sus preconcepciones, lo que no significa que el agente, desde su individualidad, no pueda ni deba hacer un esfuerzo por separar las preconcepciones que no sean atingentes a la materialidad del juicio que formula. En este sentido, todo juicio y toda decisión humana están configurados por las preconcepciones y por sus pre-juicios. El modo correcto de actuar no es suprimirlos (porque no se puede), sino regularlos de acuerdo con criterios éticos y con criterios de justicia.

Desde el punto de vista de la inteligencia artificial, se agrega un problema más al ya trazado, y es la dificultad para que el discernimiento mecánico pueda apreciar los contextos, generando de este modo proposiciones que no interpretan correctamente la realidad, como se explicaba antes. La razón de esta aporía radica en que el modo de razonar de las estructuras informáticas no es finalista sino lineal, es decir, no puede comprender la naturaleza de los fines ni el carácter abstracto de su contemplación. La inteligencia artificial puede conocer el fin integrado por quien lo programa, pero carece de la capacidad de realizar una acción contemplativa, sintética o dúctil de dicho fin o de los medios para lograrlo. En otros términos, la razón de la máquina no parece ser ni especulativa ni práctica, sino puramente lógica, en sentido lineal.

Aunque el riesgo más evidente es el de reproducir y ampliar sesgos y discriminaciones arbitrarias existentes en los datos con que se alimente a la red, este no es el único. Existe también un riesgo evidente en el hecho de que los modelos de deep learning simplemente realicen correlaciones y determinen resultados a través de análisis lineares que no son del todo compatibles con la estructura del Derecho. El Derecho ha sido concebido esencialmente como un sistema de reglas que regula situaciones sociales asignando derechos (subjetivos) a los diversos operadores jurídicos, según la base de lo que les está atribuido por la naturaleza o por la ley (derecho objetivo). Este modelo es básicamente incompatible con un sistema de correlaciones a partir de datos, toda vez que la operatividad de ambos es diversa. Un modelo tiene como punto de partida un conjunto de reglas, desde las cuales se realizan subsunciones para llegar a resultados concretos, mientras que el otro correlaciona datos con resultados, sin atender ni comprender las reglas explícitamente formuladas. Si bien los viejos sistemas expertos de la década de 1980 eran capaces de comprender normas, sus limitaciones fueron tales que, a la hora de configurarse más allá de áreas muy específicas, tendían a colapsar sin alcanzar los resultados esperados. La aproximación actual a las redes neuronales, sin embargo, implica una renuncia implícita a la presencia de reglas, por lo que su operación es, al menos en principio, incompatible con un modelo jurídico basado en ellas.

Las soluciones a esta dificultad no son obvias ni sencillas. Si bien es cierto que un algoritmo adecuadamente entrenado podría determinar predictivamente soluciones jurídicas similares a las obtenidas a través de un sistema de reglas, el hecho de que sus resultados no estén determinados por dichas normas implica que, al margen del buen resultado, el modelo no aplica el Derecho objetivo para la solución del caso y esto constituye un problema de mayor importancia desde el punto de vista de los operadores del Derecho. En cierto sentido, la cuestión presenta dificultades similares a las sugeridas por Kantorowicz (1906, 10-17) cuando postulaba la irrelevancia de la dogmática a la hora de resolver un conflicto, aunque dejaba en manos del sentir del juez la verdadera resolución y no en las de un algoritmo basado en correlaciones.

Referencias bibliográficas

Ashley, D. K., *Artificial Intelligence and Legal Analytics. New Tools for Law Practice in the Digital Age*, Cambridge University Press, Cambridge, 2017.
Aristóteles, Política, 3.8-19.

Barcia, R., *Diccionario Etimológico de la Lengua Castellana*, José María Faquineto Editor, Madrid, 1889, voz "sesgo".

Broussard, M., *Artificial Unintelligence. How Computers Misunderstand the World*, MIT Press-kindle, Cambridge MA-Londres, 2018.

Buolamwini, J. y Gebru, T., Gender Shades: Intersectional Accuracy Disparities, en *Commercial Gender Classification in Proceedings of Machine Learning Research*, 81, 2018. Disponible en:
http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline
Consultado el 12 de junio de 2019.

Casey, A. J. y Niblett, A., Self-driving Laws, en *University of Toronto Law Journal*, 66-4, 2016, pp. 429-442.

Chopra, S. y White, L. F., *A Legal Theory for Autonomous Artificial Agents*, University of Michigan Press-Kindle, Michigan, 2011.

Eubanks, V., A*utomating Inequality. How High-Tech Tools profile, police, and punish the poor*, St. Martin's Press, Nueva York, 2018.

Gurney, K., *An Introduction to Neural Networks*, UCL Press Limited, Londres, 1997.

Hinton, G., Salakhutdinov, R., Reducing the Dimensionality of Data with Neural Networks, en *Science*, 313, 2006.

Huang, X., Baker, J. y Reddy, R., A Historical Perspective of Speech Recognition, en *Communications of the ACM*, 57-1, 2013, pp. 93-103.

Kai-Fu, L., *AI Superpowers: China, Silicon Valley and the New World Order*, Houghton Mifflin Hancourt, Boston, 2018.

Kantorowicz, H., *Der Kampf um der Rechtswissenschaft*, Carl Winter, Heidelberg, 1906.

Lee, K. F., *AI Super-Powers. China, Silicon Valley, and the New World Order*, Houghton Mifflin Harcourt, Boston-Nueva York, 2018.

Marron, D., Lending by Numbers': Credit Scoring and the Constitution of Risk within American Consumer Credit, en *Economics and Society*, 35, 2007.

Murray, D., *The Madness of Crowds. Gender, Race and Identity, Bloomsbury Continuum-Kindle Edition*, Londres-Oxford-Nueva York- Nueva Delhi-Sidney, 2019.

Noble, S. U., *Algorithms of Oppression. How Search Engines Reinforce Racism*, New Yor University Press-Kindle, Nueva York, 2018.

O'Neil, C., *Weapons of math Destruction. How Big Data Increases Inequality and Threatens Democracy*, Crown Publishers, Nueva York, 2016.

Pasquale, F., *The Black Box Society. The Secret Algorithms that Control Money and Information*, Harvard University Press, Cambridge MA-Londres, 2015.

Pearl, J. y Mackenzie, D., *The Book of Why. The New Science of Cause and Effect*, Basic Books, Nueva York, 2018.

Radford, A., Metz, L. y Chintala, S., *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks in ICLR*, 2016. Disponible en:
https://arxiv.org/pdf/1511.06434.pdfï%C2%BC‰ã€‘
Consultado el 8 de agosto de 2019.

Ramirez, E., *Privacy Challenges in the Era of Big Data: The View from the Lifeguard's Chair.* Disponible en:
https://www.ftc.gov/sites/default/files/documents/public_statements/privacy-challenges-big-data-view-lifeguard's-chair/130819bigdataaspen.pdf
Consultado el 15 de Octubre de 2019.
Roget, P. M., y Davidson, G. W., *Roget's Thesaurus of English Words and Phrases*, voz "bias", Penguin, Londres, 2003.

Rumelhart, D., Hinton, G. y Williams, R., Learning Representations by Back-Propagating Errors, en *Nature*, 323, 1986, pp. 533-536.

Surdeanu, M., Nallapi, R, Gregory, G., Walker, J. y Manning, C., *Risk Analysis for Intellectual Property Litigation in Proceedings of the 13th International Conference on Artificial Intelligence and Law*, ACM, Nueva York, 2019, pp. 116-120. Disponible en: https://nlp.stanford.edu/pubs/icail11.pdf
Consultado el 21 de Octubre de 2019.

Surden, H., Machine Learning and Law, en *Washington Law Review*, 89, 1, 2014.

Unión Europea, *High Level Expert Group on Artificial Intelligence, Ethic Guidelines for Trustworthy AI*. Disponible en:
https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
Consultado el 13 de Junio de 2019.

Zhang, M., Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software, en *Forbes*, 1 de julio de 2015. Disponible en:
https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/#1530ee48713d
Consultado el 12 de junio de 2019.

Zuiderveen Borgesius, F., *Discrimination, Artificial Intelligence and Algorithmic Decision-Making (Directorate General of Democracy, 2018, Estrasburgo)*. Informe para el Consejo de Europa. Disponible en:
https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73
Consultado el 10 de Octubre de 2019.

Descargo de responsabilidad. Las opiniones expresadas en la publicación incumben únicamente a los/as autores/as. No tienen intención de reflejar las opiniones o perspectivas del CETyS ni de ninguna otra organización involucrada en el proyecto.

En colaboración con: fAIr LAC | BID