{ Cómo una máquina aprende y falla – Una gramática del error para la Inteligencia Artificial }

Matteo Pasquinelli[1]

El presente artículo es una traducción[2] de Pasquinelli, Matteo (2019). How a Machine Learns and Fails – A Grammar of Error for Artificial Intelligence. *spheres: Journal of Digital Cultures 5 (Spectres of AI)*. Disponible en: https://spheres-journal.org/contribution/how-a-machine-learns-and-fails-a-grammar-of-error-for-artificial-intelligence/

Cómo citar: Pasquinelli, Matteo (2022). Cómo una máquina aprende y falla. Una gramática del error para la Inteligencia Artificial (Traducción de Emilio Cafassi, Carolina Monti, Hernán Peckaitis y Graciana Zarauza), *Revista Hipertextos,* 10(17), pp. 13-29. https://doi.org/10.24215/23143924e054

Resumen. Trabajando en la convergencia entre las humanidades y las ciencias de la computación, este texto pretende esbozar una gramática general del aprendizaje automático y proporcionar sistemáticamente una visión general de sus límites, aproximaciones, sesgos, errores, falacias y vulnerabilidades. Se conserva el término convencional de Inteligencia Artificial aunque técnicamente hablando, sería más preciso llamarla aprendizaje automático o estadística computacional, pero estos términos no serían atractivos para las empresas, las universidades y el mercado del arte. Se hace una revisión de las limitaciones que afectan a la IA como técnica matemática y cultural, destacando el papel del **error** en la definición de la inteligencia en general. Se describe al aprendizaje automático como compuesto por tres partes: conjunto de datos de entrenamiento, algoritmo estadístico y aplicación del modelo (como clasificación o predicción) y se distinguen tres tipos de **sesgos**: del mundo, de los datos y del algoritmo. Se sostiene que los **límites** lógicos de los modelos estadísticos producen o amplifican el sesgo (que a menudo ya está presente en los conjuntos de datos de entrenamiento) y provoca errores de clasificación y predicción. Por otro lado, el grado de compresión de la información por parte de los modelos estadísticos utilizados en el aprendizaje automático provoca una **pérdida de información** que se traduce en una pérdida de diversidad social y cultural. En definitiva, el principal efecto del aprendizaje automático en el conjunto de la sociedad es la **normalización** cultural y social. Existe un grado de mitificación y sesgo social en torno a sus construcciones matemáticas, donde la Inteligencia Artificial ha inaugurado la era de la *ciencia ficción estadística*.

Palabras clave: inteligencia artificial, aprendizaje automático, sesgo algorítmico, error estadístico, datos de entrenamiento

[1] Matteo Pasquinelli (PhD) es profesor de filosofía de los medios de comunicación en la Universidad de Artes y Diseño de Karlsruhe (Alemania), donde coordina el grupo de investigación sobre inteligencia artificial y filosofía de los medios de comunicación KIM. Su investigación se centra en la intersección de las ciencias cognitivas, la economía digital y la inteligencia artificial. Ha editado la antología *Alleys of Your Mind: Augmented Intelligence and Its Traumas (Meson Press)* y, con Vladan Joler, el ensayo visual "The Nooscope Manifested: AI as Instrument of Knowledge Extractivism" (nooscope.ai). Actualmente está por publicar una monografía sobre la historia de la IA titulada *The Eye of the Master: A Labour Theory of Artificial Intelligence*.

[2] La traducción fue autorizada por el autor y realizada por parte del Equipo Editorial de Revista Hipertextos: Emilio Cafassi, Carolina Monti, Hernán Peckaitis y Graciana Zarauza.

How a Machine Learns and Fails – A Grammar of Error for Artificial Intelligence

Abstract. Working at the convergence between the humanities and computer science, this text aims to outline a general grammar of machine learning and systematically provide an overview of its limits, approaches, biases, errors, fallacies and vulnerabilities. The conventional term Artificial Intelligence is retained although technically speaking, it would be more accurate to call it machine learning or computational statistics, but these terms would not be attractive to companies, universities and the art market. A review is made of the limitations affecting AI as a mathematical and cultural technique, highlighting the role of **error** in the definition of intelligence in general. Machine learning is described as consisting of three parts: training data set, statistical algorithm and model application (as classification or prediction) and three types of **biases** are distinguished: world, data and algorithm. It is argued that the logical **limits** of statistical models produce or amplify bias (which is often already present in the training data sets) and cause classification and prediction errors. On the other hand, the degree of information compression by the statistical models used in machine learning causes a **loss of information** that results in a loss of social and cultural diversity. In short, the main effect of machine learning on society as a whole is cultural and social **normalization**. There is a degree of mythologizing and social bias around its mathematical constructs, where Artificial Intelligence has inaugurated the era of *statistical science fiction*.

Keywords: artificial intelligence, machine learning, algorithmic bias, statistic error, training data

Como uma máquina aprende e falha – Uma Gramática de Erro para Inteligência Artificial

Resumo. Trabalhando na convergência entre as ciências humanas e a informática, este texto visa delinear uma gramática geral de aprendizagem de máquinas e fornecer sistematicamente uma visão geral de seus limites, aproximações, enviesamentos, erros, falácias e vulnerabilidades. O termo convencional Inteligência Artificial é mantido, embora tecnicamente falando, seria mais preciso chamá-lo de aprendizagem mecânica ou estatística computacional, mas estes termos não seriam atraentes para as empresas, universidades e o mercado de arte. É feita uma revisão das limitações que afetam a IA como uma técnica matemática e cultural, destacando o papel do **erro** na definição da inteligência em geral. O aprendizado da máquina é descrito como consistindo de três partes: conjunto de dados de treinamento, algoritmo estatístico e aplicação do modelo (como classificação ou previsão) e três tipos de **vieses** são distinguidos: mundo, dados e algoritmo. Argumenta-se que os **limites** lógicos dos modelos estatísticos produzem ou amplificam o viés (que freqüentemente já está presente nos conjuntos de dados de treinamento) e levam a erros de classificação e previsão. Por outro lado, o grau de compressão da informação por modelos estatísticos utilizados na aprendizagem de máquinas causa uma **perda de informação** que resulta em uma perda de diversidade social e cultural. Em última análise, o principal efeito da aprendizagem mecânica na sociedade como um todo é a **normalização** cultural e social. Há um certo grau de mitologia e preconceito social em torno de suas construções matemáticas, onde a Inteligência Artificial deu início à era da *ficção científica estatística*.

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

Palavras-chave: inteligência artificial, aprendizagem de máquinas, viés algorítmico, erro estatístico, dados de treinamento

> *"Una vez que los números característicos hayan sido establecidos para la mayoría de los conceptos, la humanidad poseerá un nuevo instrumento que mejorará las capacidades de la mente en mucha mayor medida que los instrumentos ópticos fortalecen nuestros ojos, y reemplazará al microscopio y al telescopio en la misma medida en que la razón es superior a la vista"— Gottfried Wilhelm Leibniz.*

> *"La Ilustración no fue [...] una cuestión de consenso, no fue una cuestión de unidad sistemática y no fue una cuestión de despliegue de la razón instrumental: lo que se desarrolló en la Ilustración fue una idea moderna de la verdad definida por el error, una idea moderna del conocimiento definida por el fracaso, el conflicto y el riesgo, pero también por la esperanza." — David Bates.*

> *"No hay inteligencia en la Inteligencia Artificial, ni realmente aprende, aunque su nombre técnico sea aprendizaje automático, es simplemente una minimización matemática." —Dan McQuillan.*

> *"Cuando estás recaudando fondos, es Inteligencia Artificial. Cuando estás contratando, es Aprendizaje Automático. Cuando estás implementando, es regresión logística" — Joe Davidson.*

¿Qué significa para la inteligencia y, en particular, para la Inteligencia Artificial (IA) fallar, equivocarse, romper una regla? Reflexionando sobre una era anterior a la racionalidad moderna, el epistemólogo David Bates ha argumentado que la novedad de la Ilustración, en la búsqueda de conocimiento, fue una nueva metodología del error en lugar de la razón instrumental dogmática.[3] Por el contrario, el proyecto de IA (esto es, casi siempre, la IA corporativa), independientemente y quizás debido a sus sueños de cognición superhumana, se queda corto a la hora de reconocer y discutir los límites, las aproximaciones, los sesgos, los errores, las falacias y las vulnerabilidades que son nativas de su paradigma. Un paradigma de racionalidad que fracasa a la hora de proporcionar una metodología del error está destinado a terminar, presumiblemente, convirtiéndose en una caricatura de ferias de títeres, como es el caso de la idea alardeada de AGI (Inteligencia General Artificial).[4]

El aprendizaje automático se basa técnicamente en fórmulas de corrección de errores, pero la naturaleza, escala e implicaciones del error es rara vez discutida en la comunidad de programadores. Los programadores de aprendizaje automático poseen y siguen ampliando un vasto armamento de trucos de corrección de errores; sin embargo, se empeñan en una inquieta "optimización del código" sin reconocer el impacto social de sus aproximaciones lógicas. Por la complejidad de las matemáticas involucradas, el debate público sobre la IA es incapaz de considerar las limitaciones lógicas en ella, quedando polarizada entre posiciones *integradas* y *apocalípticas*, entre la tecnofilia y la tecnofobia (Eco, 2000). La posición integrada sigue los pasos de Ray Kurzweil en su feliz viaje hacia la Singularidad creyendo que las matemáticas resolverán todos los problemas, y que la automatización masiva se desarrollará sin perturbaciones para el

[3] "Fue [en la Ilustración], quizás por primera vez en el pensamiento moderno, que el error asumió un papel significativo no sólo en la definición del conocimiento sino en la propia búsqueda del conocimiento" (Bates, 2002)

[4] Una referencia al robot Sophia, construido en 2016 por Hanson Robotics. Ben Goertzel, histriónico mecenas del llamado paradigma de la Inteligencia General Artificial, supervisó el proyecto.

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

orden social. En una postura apocalíptica especulativa, malinterpretando el efecto de caja negra en el aprendizaje de las máquinas, autores como Nick Bostrom entre otros, advierten sobre una próxima era oscura de la razón en la que máquinas enceguecidas se salen de control (Bostrom, 2014). Esta última postura comparte regiones con la teoría conspirativa por la que los sistemas de IA no pueden ser estudiados, conocidos ni controlados. Incluso esta posición apocalíptica se queda en el nivel de la especulación ("*¿qué pasaría si la IA...?*") y no aclara la lógica interna del aprendizaje automático ("*¿qué es la IA?*").

Por suerte, lentamente va surgiendo una visión crítica de la IA. Gracias a libros populares como Weapons of Math Destruction, de Cathy O'Neil, entre otros, está quedando claro que el problema de la IA no tiene nada que ver con la inteligencia en sí misma, sino con la forma en que se aplica a la gobernanza de la sociedad y al trabajo a través de modelos estadísticos -que deberían ser transparentes y expuestos al escrutinio público (O'Neil, 2016; Noble, 2018; Eubanks, 2018). Como ha señalado Yarden Katz, la IA no es más que una operación de marketing utilizada para renombrar lo que hace una década se conocía al negocio de análisis de datos y centros de datos.(Katz, 2017) Profundizando en los elementos centrales de los sesgos algorítmicos, Kate Crawford ha subrayado las amplias implicaciones éticas de la clasificación y las taxonomías del aprendizaje automático, recordando que "el aprendizaje automático es el mayor experimento de clasificación de la historia de la humanidad" (Crawford, 2017). El ensayo de Kate Crawford y Vladan Joler "Anatomía de un sistema de IA" es otro ejemplo de investigación incisiva de la caja negra de la IA, en el que deconstruyen el dispositivo Amazon Echo remapeando cada uno de sus componentes en la ecología y la economía globales. Los tiempos parecen propicios para una crítica radical de la inteligencia de las máquinas: Dan McQuillan, por ejemplo, aboga por el surgimiento de una contracultura que se posicione en contra del opaco aparato normativo del aprendizaje automático (McQuillan, 2018).

En términos generales, se puede estudiar la IA como una construcción técnica o como una construcción social. Sin embargo, el debate sobre los límites de la IA puede ser inexacto si se separan los límites técnicos de los sociales, y viceversa. Las observaciones de Deleuze y Guattari sobre el reloj pueden aplicarse a la IA de forma útil: el reloj puede verse como un engranaje mecánico que proyecta el tiempo universal, o como una disciplina abstracta que controla el tiempo colectivo.[5] Estas dos perspectivas están, por supuesto, imbricadas y se estimulan mutuamente. Sin embargo, es el conjunto social el que indica la verdad sobre la técnica y la hace posible y poderosa en la historia. Parafraseando lo que Guattari dijo una vez sobre las máquinas en general, la inteligencia de las máquinas está, en definitiva, constituida por "ciertos aspectos de la subjetividad humana en formas hiperdesarrolladas e hiperconcentradas" (Guattari, 2013, p. 2). Trabajando en la convergencia de las humanidades y las ciencias de la computación, este texto pretende esbozar una gramática general del aprendizaje automático, y proporcionar sistemáticamente una visión general de sus límites, aproximaciones, sesgos, errores, falacias y vulnerabilidades. Se conserva el término convencional de Inteligencia Artificial para señalar la recepción pública y la espectacularización del aprendizaje automático y el negocio de la analítica de datos (Big Data). Técnicamente hablando, sería más preciso llamar a la Inteligencia Artificial aprendizaje automático o estadística computacional, pero estos términos tendrían cero atractivo de marketing para las empresas, las universidades y el mercado del arte. Dado el grado de

[5] "Una misma máquina puede ser a la vez técnica y social, pero sólo cuando se contempla desde perspectivas diferentes: por ejemplo, el reloj como máquina técnica para medir el tiempo uniforme, y como máquina social para reproducir las horas canónicas y para asegurar el orden en la ciudad" (Deleuze y Guattari, 1983, p. 141).

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

mitificación y sesgo social en torno a sus construcciones matemáticas, la Inteligencia Artificial ha inaugurado la era de la *ciencia ficción estadística*.

1. Presentación del Nooscopio: Un diagrama general del aprendizaje automático

El padrino de las redes neuronales convolucionales, Yann LeCun, sostiene que los sistemas de IA actuales no son versiones sofisticadas de cognición, sino de percepción (LeCun, 2018). A finales de los años 50, el aprendizaje automático surgió como una forma de reconocimiento de patrones visuales que luego se extendió al análisis de datos no visuales. En el caso de los automóviles auto-conducidos, los patrones reconocibles son las características visuales más comunes de un escenario vial y, en el caso de la traducción automática, los patrones son las secuencias de palabras más comunes entre dos idiomas. Sin embargo, lo que el aprendizaje automático calcula no es un patrón exacto, sino la **distribución estadística de un patrón**. Sólo con ver la superficie del marketing de la IA, uno se encuentra con una construcción estadística compleja de examinar. ¿Cómo se construyen estos modelos estadísticos? ¿Qué grado de precisión y fiabilidad tienen? ¿Cuál es la relación entre los modelos estadísticos y la inteligencia humana? De hecho, convendría reformular la ingenua pregunta "¿puede pensar una máquina?" en la teóricamente más sólida "¿puede pensar un modelo estadístico?".

La Inteligencia Artificial no es "inteligente" en absoluto. Sería más preciso enmarcarla como un instrumento de conocimiento o ampliación lógica que *percibe* patrones que están fuera del alcance de la mente humana. Leibniz, abordando esta modalidad de IA, utiliza el telescopio y el microscopio como metáforas de su *calculus ratiocinator*.[6] De manera similar, un sistema de aprendizaje automático puede compararse con un **nooscopio**, un dispositivo que mapea y percibe patrones complejos a través de grandes espacios de datos (lo que las humanidades digitales denominan **lectura a distancia**) (Moretti, 2013; Vogl, 2007). Sin embargo, cada instrumento de medición y percepción viene con aberraciones incorporadas y contingentes. Del mismo modo que las lentes de los microscopios y telescopios nunca son perfectamente curvilíneas y lisas, las *lentes lógicas* de los sistemas de IA tienen sus propios defectos y aberraciones. Estudiar el impacto de la IA es estudiar el grado en que los flujos de información son difractados, distorsionados y perdidos por ella. Para entender la naturaleza de esa pérdida de información, hay que estudiar la anatomía algorítmica de los modelos estadísticos que subyacen al aprendizaje automático.

En términos matemáticos, el aprendizaje automático se utiliza para predecir un valor *output y* a partir de un valor *input x*. Los algoritmos dibujan una función que relaciona *x* con *y* aprendiendo de datos pasados en los que tanto *x* como *y* se conocen: y = f(x). Construyendo esa función, el algoritmo podrá predecir y basándose en futuras configuraciones de *x*. Por ejemplo, dadas unas fotos de animales (x), el algoritmo aprende su asociación con las categorías "gato" o "perro" (y) y luego intenta clasificar las nuevas fotos en función de ellas. En este caso, el número *input x* es una imagen digital, y el número *output y* es un porcentaje relacionado con una etiqueta semántica (97 % "gato", 3 % "perro"). Se trata de un proceso de clasificación que se distingue de la regresión en la que el *output* es un número continuo. Un ejemplo de esto último sería un algoritmo que aprende a predecir la puntuación de una evaluación, el *output y*, para cualquier edad

[6] Ver la cita inicial de Leibniz.

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

de un grupo de estudiantes, el *input x*. Tanto la clasificación como la regresión son casos de aprendizaje supervisado, en los que el algoritmo toma datos en los que se conoce la relación entre el *input x* y el *output y* e intenta adivinar la salida *y* para futuros *inputs* desconocidos *x*. Se dice que un algoritmo de aprendizaje automático aproxima la función que relaciona *y* con *x*.

Un sistema de aprendizaje automático se presenta ante un usuario u operador compuesto por tres elementos o etapas: datos de entrenamiento, algoritmo de aprendizaje y aplicación del modelo.

1. **Datos de entrenamiento**: El conjunto de datos de entrenamiento contiene datos que deben analizarse para extraer conocimiento e "inteligencia", es decir, patrones de asociación entre sus elementos. En el aprendizaje supervisado, el conjunto de datos de entrenamiento se compone de dos elementos: el *input x* (por ejemplo, imágenes en crudo, edades de los estudiantes) y el *output y* (etiquetas que describen esas imágenes, puntuaciones de evaluación). En el aprendizaje no supervisado o autosupervisado, sólo se da el *input x*, a partir del cual hay que descubrir un patrón desconocido *y*.
2. **Algoritmo de aprendizaje**: El algoritmo de aprendizaje extrae patrones de los datos de entrenamiento leyendo la asociación entre el *input x* y el *output y* y construyendo una descripción estadística de esta asociación. El modelo estadístico es el núcleo del aprendizaje automático, depositario de la "inteligencia" extraída de los datos de entrenamiento. Sin embargo, nunca es 100 % preciso y no existe un método científico para evaluarlo: el proceso de entrenamiento se detiene cuando un operador humano decide que *se ha alcanzado una tasa de error aceptable para un conjunto de datos de prueba*.
3. **Aplicación del modelo**: Cuando el modelo estadístico se considera suficientemente entrenado y "se ajusta" a los datos de entrenamiento, puede aplicarse a diferentes tareas, como la clasificación y la predicción. En la clasificación (o *reconocimiento*), un nuevo valor *x* se asocia a una etiqueta *y*, si *x* se ajusta a la distribución del modelo estadístico. En la predicción (o *generación*), un nuevo valor *x* se utiliza para generar y predecir su correspondiente valor *y* utilizando el mismo modelo estadístico (la generación de patrones es, lógicamente, lo mismo que la predicción).

El montaje de estos tres elementos (Datos + Algoritmo + Modelo) se propone como un diagrama general del aprendizaje automático. Siguiendo con la metáfora de los medios ópticos como los telescopios y microscopios, puede decirse que el flujo de información que atraviesa dicho instrumento de conocimiento (aquí denominado *nooscopio*) se comporta como un haz de luz que es proyectado por los datos de entrenamiento, difractado por el algoritmo y su modelo estadístico y reflejado de vuelta al mundo con una distorsión incorporada. Los siguientes pasajes describen cada componente individual centrándose en particular en la naturaleza del modelo estadístico que se encuentra en el núcleo del aprendizaje automático.

2. Datos de entrenamiento, o la fuente colectiva de la inteligencia de las máquinas

La digitalización masiva, que comenzó después de la Segunda Guerra Mundial con la

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

comercialización de los *mainframes* industriales y alcanzó su punto álgido en los 2000s con los centros de datos globales, sentó las bases de un régimen de *extractivismo de inteligencia*. La inteligencia de las máquinas se entrena con extensos conjuntos de datos que no son acumulados ni de forma técnicamente neutral ni socialmente imparcial. Los datos neutrales no existen, ya que dependen del trabajo individual, de los datos personales y de los comportamientos sociales que se acumulan durante largos periodos de tiempo, a partir de extensas redes y diversas taxonomías culturales (Gitelman, 2013).

Los datos de entrenamiento probablemente son el factor más importante en la calidad de la "inteligencia" que extraen los algoritmos del aprendizaje automático. El conjunto de datos de entrenamiento suele estar compuesto de datos *input* y datos *output* ideales: las imágenes digitales en bruto, por ejemplo, pueden estar asociadas con etiquetas (que es la manera en que los humanos suelen categorizar esas imágenes con sus significados). Tal y cómo se describe anteriormente en términos matemáticos, el aprendizaje automático se calcula a partir de la relación entre la imagen inicial (*input*) con sus etiquetas (*output*) con el fin de predecir las etiquetas (*output*) de futuras imágenes similares (*input*). La elaboración, el formateo y la edición del conjunto de datos de entrenamiento es una tarea laboriosa y delicada, que probablemente sea más importante que los parámetros técnicos que controlan el algoritmo de aprendizaje.[7] En la preparación de los conjuntos de datos de entrenamiento pueden reconocerse cuatro etapas:

1. **Producción**: el trabajo o fenómeno individual que produce información.
2. **Captura**: la captación de información mediante un instrumento que la convierte en datos.
3. **Formateo**: la codificación de información en un formato de datos específico.
4. **Etiquetado**: la aplicación de categorías de una taxonomía determinada al conjunto de datos.

Los conjuntos de datos de entrenamiento más populares usados para aprendizaje automático (NMIST, ImageNet, Labelled Faces in the Wild, etc.) se originaron en empresas, universidades y agencias militares del Norte Global (aunque si se mira con más cuidado, se descubre una profunda división del trabajo que inerva al Sur Global). Los datos de entrenamiento pueden provenir de comportamientos espontáneos en línea (a través de las redes sociales, la cobertura de noticias, la geolocalización de los teléfonos móviles, etc.) o del trabajo en pantalla que se realiza a partir de colaboración colectiva (a través de Amazon Mechanical Turk, por ejemplo). En los dos casos, se lleva a cabo trabajo invisibilizado y poco reconocido. Los datos personales, en particular, son enterrados y desaparecen en conjuntos de datos privatizados sin saberlo y sin transparencia (Murgia, 2019).[8] Por este motivo, tales conjunto de datos desencadenan cuestiones de soberanía de datos, privacidad y derechos civiles de las que los organismos políticos y la ley están tomando conciencia poco a poco (véase el reglamento de privacidad de datos GDPR que fue aprobado en mayo de 2018 por el Parlamento Europeo).

[7] Por ejemplo, se necesitaron nueve años de trabajo manual para etiquetar los 14 millones de imágenes del conjunto de datos de entrenamiento ImageNet, patrocinado por las universidades de Google, Amazon, Princeton y Stanford.

[8] Ver el proyecto Megapixel de Adam Harvey (megapixels.cc).

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

4. Las modalidades del aprendizaje automático: entrenamiento, clasificación y predicción

Cuando los datos de entrenamiento están listos para ser analizados, se presentan al algoritmo de aprendizaje, el cual es elegido, entre muchas opciones, por un operador humano a partir de parámetros específicos. Por ejemplo, las redes neuronales convolucionales requieren la especificación de una topología muy compleja y un conjunto de hiper-parámetros (número de capas, neuronas, tipo de conexión, comportamiento de cada capa y neurona, etc.). Aunque las redes neuronales surgieron inicialmente como una técnica de reconocimiento de patrones, los informáticos prefieren hoy en día la expresión más abstracta y precisa de **mapeo de *input-output*** para evitar la anticuada comparación con los sistemas biológicos y la percepción visual. Sin embargo, la construcción de la relación entre un *input* x y un *output* y sigue siendo la búsqueda de un patrón. Un ejemplo primordial del reconocimiento de patrones básicos es el Perceptrón de Frank Rosenblatt, el cual fue creado en 1957 y fue la primera red neuronal operativa. Dada una matriz visual de 20x20 fotorreceptores, esta máquina podía aprender a reconocer una simple letra. Hoy, dado un *input* mucho más complejo como la grabación en vídeo de una calle concurrida, se pide a la red neuronal de un automóvil auto-conducido que controle los engranajes mecánicos y tome decisiones éticas cuando se produzcan situaciones de peligro, lo que exige un mapeo de *input-output*.

Sin importar su complejidad, para la perspectiva numérica del aprendizaje automático, nociones como imagen, movimiento, forma, estilo o decisión pueden describirse como distribuciones estadísticas de un patrón. Desde el punto de vista del modelo estadístico, se dan tres modalidades de funcionamiento del aprendizaje automático: 1) entrenamiento, 2) clasificación y 3) predicción. En términos más intuitivos, pueden definirse como: abstracción de patrones, reconocimiento de patrones y generación de patrones.

1. En la modalidad de **entrenamiento** (*abstracción de patrones*), el algoritmo "aprende" la asociación de un *input* x con un *output* y (por ejemplo, su etiqueta). Como se ha mencionado, el algoritmo teje una distribución estadística de los patrones subyacentes y los extrae de su fondo. El modelo estadístico se considerará entrenado cuando se alcance una tasa de error aceptable en un conjunto de datos de prueba (hasta la fecha, no existe ningún método científico para determinar cuándo un modelo está suficientemente entrenado, es decir, cuándo una IA parece ser "inteligente").
2. En la modalidad de **clasificación** (*reconocimiento de patrones*), los nuevos datos de *input* x se comparan con el modelo estadístico para determinar si entran o no en su distribución estadística. En caso afirmativo, se les asigna la correspondiente etiqueta de *output* y. Hoy en día existen clasificadores de objetos que pueden detectar todos los objetos más comunes en un escenario vial y aplicar etiquetas como persona, coche, camión, bicicleta o semáforo en cuestión de milisegundos -por supuesto, con un margen de error-.
3. En la modalidad de **predicción** (*generación de patrones*), los nuevos datos de *input* x se utilizan para predecir su valor de *output*. En esta modalidad, se puede decir que el modelo estadístico se ejecuta hacia atrás para generar nuevos patrones en lugar de registrarlos. La expresión "arte creado por la IA" significa en realidad que un operador humano aplica la modalidad generativa de las redes neuronales después de entrenarlas con un determinado

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

conjunto de datos. Por ejemplo, tras ser entrenada con el conjunto de datos MIDI de un compositor musical, una red neuronal puede generar una nueva melodía que se asemeje al estilo del compositor. La modalidad generativa es útil, como una especie de "control de la realidad" algorítmica, ya que muestra lo que el modelo aprendió, es decir, cómo el modelo "ve el mundo".

4. Tres tipos de sesgo

El bucle de información entre la IA y la sociedad -es decir, entre el aprendizaje automático y los datos de entrenamiento- no es virtuoso, sino que está corrompido por un sesgo técnico. Cada conjunto de datos de entrenamiento - independientemente de lo preciso que pueda parecer- es un muestreo estadístico y, por lo tanto, una visión parcial del mundo. Además, el grado de comprensión de la información de los algoritmos del aprendizaje automático afecta las proporciones originales de los datos de entrenamiento, lo que a su vez amplifica el sesgo. El sesgo es la temática más debatida y conocida del aprendizaje automático ya que tiene implicaciones sociales directas y es una buena forma de empezar a ilustrar las limitaciones lógicas de estos modelos estadísticos. En el aprendizaje automático, es necesario distinguir entre el sesgo del mundo, de los datos y de los algoritmos.

El **sesgo del mundo** ya es evidente en la sociedad antes de la intervención tecnológica, pero los conjuntos de datos refuerzan las desigualdades de raza, género y clase, normalizando aún más los estereotipos ya operables. La naturalización del sesgo por parte del aprendizaje automático, es decir, la integración de la desigualdad en un algoritmo como "datos aparentemente imparciales", puede ser perjudicial por sí misma (Eubanks, 2018). Para precisar las categorías de sesgo, Kate Crawford ha distinguido entre un daño de asignación de recursos (por ejemplo, cuando un algoritmo niega hipotecas a un grupo minoritario) y un daño de representación social (como la denigración, la infrarrepresentación o la determinación injusta de la raza, el género y la clase) (Crawford, 2017).

El **sesgo de los datos**, por otro lado, es introducido a través de la captura, formateo y etiquetado de datos mediante el conjunto de datos de entrenamiento. El acto de capturar y formatear los datos tiene la potencialidad de afectar la resolución y precisión de la información, pero la parte más delicada del proceso es el etiquetado. Universidades, empresas y organismos militares construyen conjuntos de datos de entrenamiento con mano de obra tosca y barata. A menudo, utilizan **taxonomías** antiguas y conservadoras provocando una visión del mundo distorsionada de las culturas y diversidades. Como ya había dilucidado Foucault, esas taxonomías suelen reproducir jerarquías sociales y son expresiones del poder normativo (Foucault, 2005). Hoy en día, las taxonomías culturales y científicas son integradas en el aprendizaje automático y formalizadas por él: su poder normativo no es más institucional sino computacional.

El **sesgo algorítmico** (también conocido como "sesgo de la máquina", "sesgo estadístico" o "sesgo del modelo") es la amplificación del sesgo del mundo y de los datos causada por los errores computacionales, la compresión de la información y las técnicas de aproximación de los algoritmos de aprendizaje automático. Debido a sus ratios en la comprensión de la información, los algoritmos del aprendizaje automático *difractan* y *distorsionan* los sesgos del mundo y de los datos, produciendo que las desigualdades sean aún más desiguales. La difracción y la amplificación se representan en la ilusión de la perspectiva anamórfica utilizada en la pintura y el

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

diseño gráfico. Asimismo, la visión del mundo del aprendizaje automático también es *anamórfica*: a pesar de que se respete la forma o topología del mundo, se distorsiona sus proporciones.

5. Los límites lógicos del modelo estadístico

En el núcleo de los actuales sistemas de IA hay un algoritmo de aprendizaje cuyo propósito es calcular un modelo estadístico de los datos de entrenamiento. Los informáticos lo llaman simplemente "**el modelo**". El modelo es la representación estadística de un conjunto de datos de entrenamiento amplio y diverso en un solo archivo. Desde los tiempos del Perceptrón de Rosenblatt, la primera red neuronal operativa, el objetivo clave del aprendizaje automático ha sido almacenar un pequeño modelo estadístico, en lugar de memorizar, por ejemplo, mil fotos del mismo objeto desde diferentes ángulos. El modelo se calcula mediante diferentes técnicas (por ejemplo, redes neuronales, máquinas de vectores de apoyo, redes bayesianas) que siempre adoptan la forma de una **inferencia estadística** cuyos resultados tomarán, en consecuencia, la forma de una **distribución estadística**. Técnicamente se dice que el modelo aprende la distribución estadística de los datos de entrenamiento mapeando las correlaciones (también conocidas como patrones o dependencias) entre el *input* y el *output* deseado. En última instancia, el modelo estadístico construye una función *f* que, cuando es eficaz, describe los datos de entrenamiento de forma adecuada y predice el *output* de un *input* futuro.

Tomemos un ejemplo clásico de aprendizaje automático: LeNet, desarrollada por Yann LeCun en 1988, es una red neuronal convolucional para el reconocimiento óptico de números en códigos postales y cheques bancarios. Los datos de entrenamiento proceden de la base de datos MNIST, que contiene 60.000 números escritos a mano (recogidos únicamente entre dos grupos sociales: estudiantes de secundaria de EE.UU. y empleados de la Oficina del Censo). El modelo interno de LeNet registra la asociación estadística de imágenes de números escritos a mano con su etiqueta correcta, que en este caso es un número (LeCun et al, 1989). Después de ser entrenado, el modelo estadístico de LeNet reconocerá las futuras ocurrencias de números escritos a mano con un margen de error.

Se dice que un modelo estadístico ha sido entrenado con éxito cuando puede **generalizar** los patrones del conjunto de datos de entrenamiento a nuevos datos "naturales", **ajustándose** elegantemente a los datos de entrenamiento con el menor margen de **error** posible (*siempre* hay un margen de error en el aprendizaje automático). Si un modelo aprende demasiado bien los datos de entrenamiento, sólo será capaz de reconocer las coincidencias exactas y pasará por alto los patrones con gran similitud. En este caso, se dice que el modelo está **sobre-ajustado** (*overfitting*), ya que no es capaz de distinguir los patrones del fondo, es decir, ha aprendido meticulosamente todo, *incluido* el ruido. Por otro lado, el modelo está **infra-ajustado** (*underfitting*) cuando no es capaz de formular patrones a partir de los datos de entrenamiento. En el sobreajuste no hay compresión de información, mientras que en el infra-ajuste el modelo ha perdido la mayor parte de la información valiosa.[9]

Es habitual describir la IA como la medida estadística de una correlación entre puntos de datos. De hecho, el aprendizaje automático no *aprende* nada en el sentido propio de la palabra; sólo mapea un *input x* con un *output y*, dibujando una función que describe *aproximadamente* su

[9] Un tercer caso puede darse cuando un modelo aprende una asociación de patrones errónea. Si la apofenia es la tendencia humana a percibir patrones significativos en datos aleatorios, el infra-ajuste es una especie de apofenia maquínica. La apofenia maquínica se produce si un modelo estadístico ve un patrón que no existe, es decir, si lee el ruido como similar a un patrón existente.

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

tendencia, aplicando luego dicha función a futuros *inputs* para predecir sus *outputs*. Esta función también es una aproximación, en el sentido de que adivina las "partes que faltan" del gráfico de datos: ya sea mediante **interpolación**, que es la proyección y predicción de un *output y* que cae dentro del intervalo conocido del *input x* en el conjunto de datos de entrenamiento, o mediante **extrapolación**, que es la proyección y predicción del *output y* más allá de los límites de *x*, a menudo con altos riesgos de inexactitud.

El aprendizaje automático es increíblemente eficiente como algoritmo para analizar datos y aproximar una función matemática que los describa. De hecho, los informáticos se sienten más cómodos con la definición de la IA como técnica de **compresión de información** que con la concepción popular de que es una manifestación de cognición sobrehumana.[10] Desde la antigüedad, los algoritmos han sido procedimientos de naturaleza económica, diseñados para lograr un resultado en el menor número de pasos y consumiendo la menor cantidad de recursos, como espacio, tiempo, energía, etc. La actual carrera armamentística entre las empresas de IA sigue consistiendo en encontrar los algoritmos más rápidos para calcular modelos estadísticos. La compresión de la información, por tanto, mide la proporción de ganancias de estas empresas, pero también, la proporción de **pérdida de información** – y dicha pérdida suele significar una pérdida de la diversidad cultural del mundo.

La analogía de los medios ópticos ilumina las características de la IA mejor que la analogía del cerebro humano. Dejando de lado, por el momento, el hecho de que la primera red neuronal operativa, el Perceptrón, era una *máquina de visión* (Virilio, 1994), existen similitudes epistémicas entre el aprendizaje automático y los medios ópticos como en el caso de, utilizando las sugerencias de Leibniz, el microscopio y el telescopio.[11] El aprendizaje automático, al igual que estos dispositivos, presenta problemas tanto de **resolución de la información** como de **difracción de la información**, y los modelos estadísticos desempeñan una función correctora similar a la de las lentes en los medios ópticos. En términos de **ofuscación de la información**, un problema bien conocido del aprendizaje automático es probablemente el **efecto de caja negra**, presente en las grandes redes neuronales (Aprendizaje Profundo). "Caja negra" es un término popular que se utiliza para describir cómo la compresión de la información borra una gran cantidad de información aparentemente inútil, dando lugar a una condición de ofuscación que es irreversible.[12] Esto ocurre a medida que cada capa de neuronas descarga la mayor parte de los datos recibidos de la anterior, olvidando en el proceso algunos eslabones de la cadena de "razonamiento". Fuera de la informática, "caja negra" se ha convertido en una metáfora genérica para indicar la aparente complejidad de los sistemas de IA, ya que pueden parecer inescrutables y opacos, cuando no ajenos y fuera de control. Proyectos como *Explainable Artificial Intelligence*, *Interpretable Deep Learning* y *Heatmapping*, entre otros, han demostrado, sin embargo, que es posible entrar en la "caja negra" y hacer que su oscura cadena de cálculo sea interpretable para los usuarios.[13]

[10] Los informáticos argumentarían que la IA pertenece realmente a un subcampo del procesamiento de señales, es decir, la compresión de datos.

[11] Como ya se ha mencionado, el aprendizaje automático es una especie de cine estadístico, proyectando el nuevo género de la ciencia ficción estadística.

[12] También hay problemas de propagación de errores en los que algunas características del hardware de la GPU pueden generar una cadena de errores que llega a las capas superiores de abstracción de funciones (Li, Guanpeng et al., 2017).

[13] No obstante, la plena interpretabilidad y explicabilidad de los modelos estadísticos de aprendizaje automático sigue siendo también un mito (Lipton, 2016).

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

Debido al grado de compresión y pérdida de información que se produce en sus modelos estadísticos, el aprendizaje automático requiere una reducción de las etiquetas y categorías que están inicialmente presentes en los conjuntos de datos de entrenamiento. En una técnica denominada **reducción de dimensionalidad**, por ejemplo, las categorías que muestran una *varianza baja* (es decir, cuyos valores fluctúan poco) se agregan y eliminan para reducir los costes de cálculo. La reducción de la dimensionalidad, por tanto, conduce a algo que puede llamarse **reducción de categorías**, la cual es también una reducción de taxonomías culturales. Eventualmente, el efecto del aprendizaje automático sobre la diversidad del mundo es la **normalización**, es decir, la equiparación de las anomalías a una norma media. El término técnico de regresión se refiere en realidad al fenómeno de regresión hacia la media que Francis Galton observó al medir la altura de las personas. Las redes neuronales de reconocimiento facial, por ejemplo, muestran una tendencia a favorecer las imágenes de personas de color de piel claro. La **regresión hacia la media** no es entonces sólo una técnica matemática de aprendizaje automático, sino que tiene claras consecuencias sociales e implicaciones políticas.

6. Técnicas de aproximación y los peligros de la correlación

Como bien dice Dan McQuillian: "No hay inteligencia en la Inteligencia Artificial, ni aprende realmente, aunque su nombre técnico sea aprendizaje automático, es simplemente minimización matemática"(McQuillan, 2018a). Es importante recordar que la "inteligencia" del aprendizaje automático no se basa en la aplicación de fórmulas exactas de análisis matemático, sino en algoritmos de **aproximación**, es decir, en procedimientos heurísticos. La forma de la función de correlación entre el *input* x y el *output* y se calcula algorítmicamente, paso a paso, a través de tediosos procesos mecánicos de ajuste gradual. Es el mismo procedimiento que se utiliza en la **geometría diferencial** o en el cálculo, en el que se utilizan pequeños bloques cuadrados para aproximar un área irregular en lugar de dibujar una forma curvilínea exacta. Se dice que las redes neuronales están entre los algoritmos más eficientes para el aprendizaje porque estos métodos diferenciales de aproximación permiten *adivinar* cualquier función dadas suficientes capas de neuronas y tiempo de computación (como demuestra el llamado Teorema de Aproximación Universal). Cuando se dice que "las redes neuronales pueden resolver cualquier problema", se quiere decir que pueden *aproximar* la forma de cualquier curva (cualquier función no lineal) en un espacio multidimensional de datos.[14] La aproximación gradual de una función mediante fuerza bruta es la característica principal de la IA actual, y sólo desde esta perspectiva se pueden entender sus potencialidades y limitaciones.

Otra problemática del aprendizaje automático es cómo se utiliza la **correlación estadística** entre dos elementos para explicar la **causalidad lógica** de uno a otro. En la gramática de los errores de la IA, esto no es un error atribuido a la máquina, sino una falacia humana. Se entiende comúnmente que *la correlación no implica causalidad*, lo que significa que una correlación estadística por sí sola no es suficiente para demostrar la causalidad. Esta falacia lógica se convierte fácilmente en una falacia política. La ilusión de la causalidad puede utilizarse, por ejemplo, para respaldar algoritmos policiales predictivos. Cuando el aprendizaje automático se aplica a la

[14] Nota bene: en estos pasajes, la línea divisoria entre los puntos de datos de *input* y *output* se ha descrito como una curva. En realidad, el aprendizaje automático calcula esas aproximaciones diferenciales en espacios n-dimensionales dibujando, entonces, hiperplanos (en lugar de una curva en una matriz bidimensional).

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

sociedad de este modo, las correlaciones predictivas se transforman en un aparato político de **prevención**. Dan McQuillan señala: "La naturaleza predictiva del aprendizaje automático promueve la prevención, es decir, la acción que intenta anticipar o prevenir el resultado previsto" (McQuillan, 2018b). La prevención, como automatización de la toma de decisiones, contribuye a la exclusión de la participación colectiva en las instituciones sociales y políticas. El aprendizaje automático puede incluso apoyar correlaciones arbitrarias y sin sentido (por ejemplo, entre el consumo diario de queso, la etnia y el puntaje de crédito siempre se puede encontrar una correlación estadística). Esto es lo que se llama **apofenia algorítmica**, la consolidación ilusoria de correlaciones o relaciones causales que no existen en el mundo material, sino sólo en la mente de la IA.[15]

7. La imprevisión de lo nuevo

Otro límite lógico encontrado en el núcleo del aprendizaje automático es la incapacidad de predecir y reconocer una nueva **anomalía única**, es decir, una anomalía que sólo aparece una vez, como una nueva metáfora en una poesía, un nuevo chiste en el lenguaje cotidiano o un objeto misterioso en medio de la ruta. Por ejemplo, los sistemas de IA con algoritmos de reconocimiento del habla tienen problemas cuando se enfrentan a los dialectos locales. Y lo que es peor, las minorías sociales a menudo quedan fuera del radar de la logística de la IA y son excluidas (por ejemplo, las personas que hablan con acento escocés a Amazon Alexa o las comunidades negras a las que no llega el servicio de entrega de Amazon) (Ingold y Soper, 2016). La no detección de lo nuevo (algo que es inesperado, es decir, que nunca antes ha sido "visto" por una máquina y, por lo tanto, no está clasificado en una categoría conocida) es un problema especialmente peligroso para los coches autoconducidos, que ya han causado fatalidades por este motivo. Los **ataques adversarios** explotan estos puntos ciegos en el aprendizaje de las máquinas, utilizando patrones insólitos que obstruyen la lectura visual del entorno por parte de la máquina: estos patrones son a veces diseñados por una mente humana a sabiendas de que una "mente" de IA nunca los ha visto.

En el aprendizaje automático, el problema de la **predicción de lo nuevo** está lógicamente relacionado con el problema de la **generación de lo nuevo**. Curiosamente, la definición lógica de un problema de seguridad también describe el límite lógico de la creatividad en el aprendizaje automático. La trillada pregunta "¿Puede la IA crear arte?" debería reformularse en términos técnicos: ¿Puede la IA crear obras que no sean imitaciones del pasado? ¿Es capaz la IA de extrapolar más allá de los límites estilísticos de los datos de entrenamiento? La respuesta es: no realmente. La "creatividad" del aprendizaje automático se limita a la **detección de los estilos antiguos** a partir de los datos de entrenamiento y a la posterior improvisación aleatoria a lo largo de dichos estilos. En otras palabras, el aprendizaje automático sólo puede explorar e improvisar dentro de los límites de las categorías establecidas por los datos de entrenamiento. Las obras de arte del Obvious Collective (*nomen est omen*), un proyecto de colaboración que crea cuadros utilizando la IA, ofrecen pruebas visuales de estas limitaciones. El estilo de sus retratos está muy normalizado y es estéticamente predecible (Vincent, 2018). Por tanto, sería más preciso denominar el arte de la IA como *arte estadístico*.

[15]Ver también Illusory Correlation. (Marzo, 2019). Sobre la apofenia, ver Pasquinelli (2015).

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

En cuanto al procesamiento del lenguaje natural, cabe preguntarse si la IA es capaz de inventar nuevas metáforas de forma consistente y no aleatoria. En una época anterior al aprendizaje automático, cuando se le preguntó si una metáfora podía ser inventada por un algoritmo, Umberto Eco respondió: "No existe ningún algoritmo para la metáfora, ni se puede producir una metáfora mediante las instrucciones precisas de una computadora, sea cual sea el volumen de información organizada que se introduzca" (Eco, 1986, p.127).

Cualquier metáfora nueva es la ruptura de una regla y la invención de otra nueva, argumentó Eco. ¿Se puede programar un algoritmo para que rompa las reglas (patrones) de sus datos de entrenamiento de forma creativa? El aprendizaje automático nunca podrá detectar o generar el famoso verso de Rimbaud "Yo es otro" tras realizar un análisis estadístico de un millón de periódicos. El aprendizaje automático nunca inventa códigos y mundos, sino que dibuja espacios vectoriales que reproducen frecuencias estadísticas de datos antiguos. En la estadística computacional, una nueva metáfora es un **nuevo vector** sin similitudes de frecuencia con vectores antiguos, algo que desaparecería fácilmente en el siguiente pasaje computacional. Además, una metáfora no es la correlación estadística de dos significados, sino la construcción de un *nuevo modelo de mundo* en el que esta nueva expresión adquiriría un sentido lógico (una *causalidad*) que no tenía en el antiguo modelo de mundo. Una nueva metáfora es la invención de un paradigma constituyente. A menudo las metáforas son banales, pero a veces pueden ser brillantes y *abiertas*, como cuando dejan espacio para una interpretación infinita, un proceso clave para las humanidades aunque no sólo. Uno se pregunta quién sueña con mecanizar la hermenéutica, el arte de la interpretación y el juicio estético, procesos que deben permanecer sin ataduras. Aunque el arte de la interpretación puede enriquecerse y ampliarse, por supuesto, con nuevos instrumentos de ampliación lógica y exploración de patrones.

8. Conclusión

Toda **anomalía** (también social y política) es la invención de un nuevo código o norma. Por otro lado, el poder suele basarse en la normalización de códigos y reglas, que buscan minimizar la aparición de lo anómalo. El aprendizaje automático no es una excepción cuando se aplica a la medida y la gobernanza de la sociedad. Véase, por ejemplo, el experimento de incrustación de palabras de Bolukbasi et al., que utilizó Word2vec como modelo estadístico pre-entrenado para analizar los posteos de Google News como datos de entrenamiento. Cuando se pidió al algoritmo que resolviera la ecuación "el hombre es al programador informático lo que la mujer es a x", respondió problemáticamente con x = 'ama de casa', lo que demuestra el efecto de la IA en el refuerzo de los estereotipos (Bolukbasi et al., 2016). Las diversidades sociales y culturales desaparecen fácilmente en el aprendizaje automático, ya que los algoritmos no pueden expresar la profundidad semántica a menos que se vuelvan lentos e ineficaces.[16] La IA está representando un mundo cada vez más estandarizado, en el que las normas institucionales y sociales tradicionales se traducen y amplifican en nuevas normas estadísticas y computacionales (Pasquinelli, 2017).

[16] El impacto de la IA en la sociedad ya se registra en los comportamientos cotidianos, cuando las personas se adaptan al algoritmo y no al revés. Cada vez es más común, por ejemplo, ajustar la pronunciación y neutralizar la entonación para asegurarse de que el software de reconocimiento de voz de un centro de llamadas o de un teléfono inteligente recoja las palabras correctamente. Este comportamiento autocorrectivo es una integración y absorción inconsciente de los sesgos del aprendizaje automático por parte de la propia sociedad.

DOI: https://doi.org/10.24215/23143924e054, https://revistas.unlp.edu.ar/hipertextos

Este ensayo ha intentado revisar las limitaciones que afectan a la IA como técnica matemática y cultural, destacando el papel del error en la definición de la inteligencia en general. Construyó un índice tentativo de límites, aproximaciones, sesgos, errores, falacias y vulnerabilidades del aprendizaje automático. Describió el aprendizaje automático como compuesto por tres partes: conjunto de datos de entrenamiento, algoritmo estadístico y aplicación del modelo (como clasificación o predicción). Luego, distinguió tres tipos de sesgos: del mundo, de los datos y del algoritmo. Sostuvo que los **límites** lógicos de los modelos estadísticos producen o amplifican el **sesgo** (que a menudo ya está presente en los conjuntos de datos de entrenamiento) y provoca **errores** de clasificación y predicción. Sin embargo, no se trata de una cuestión de máquinas, sino de una **falacia** política, cuando una correlación estadística entre números dentro de un conjunto de datos se recibe y acepta como causalidad entre entidades reales del mundo. El grado de compresión de la información por parte de los modelos estadísticos utilizados en el aprendizaje automático provoca una **pérdida de información** también con respecto a la granularidad de las categorías y taxonomías, lo que se traduce en una pérdida de diversidad social y cultural. El límite último de los modelos de IA se encuentra en la incapacidad de detectar y predecir una **anomalía única**, como una metáfora en el lenguaje natural. Por la misma razón, los sistemas de IA también son vulnerables a los **ataques de adversarios** que puede lanzar un operador externo conociendo las regiones débiles de un modelo estadístico. En definitiva, el principal efecto del aprendizaje automático en el conjunto de la sociedad es la **normalización** cultural y social. La IA corporativa no hace sino extender el poder normativo de las antiguas instituciones del conocimiento a los nuevos aparatos computacionales. La normatividad distorsionada de la IA procede de las limitaciones lógicas del modelado estadístico, una técnica que se venera, vergonzosamente, como tótem animista de la cognición sobrehumana.

Referencias

Bates, D.W. (2002). *Enlightenment Aberrations: Error and Revolution in France*. Ithaca, NY: Cornell University Press.

Bolukbasi, T. et al. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv.org*. Disponible en: arxiv.org/abs/1607.06520

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

Crawford, K. (diciembre, 2017). The Trouble with Bias. *Conferencia magistral en la Annual Conference on Neural Information Processing Systems (NIPS)*. Video disponible en: https://www.youtube.com/watch?v=fMym_BKWQzk

Davison, J. (Junio, 2018). No, Machine Learning is not just glorified Statistics. *Medium*. Recuperado el 21 de marzo de 2019 de https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3

Deleuze, G. and Guattari, F. (1983). *Anti-Oedipus*. Minneapolis: University of Minnesota Press.

Eco, U. (1986). *Semiotics and the Philosophy of Language*. Bloomington: Indiana University Press.

Eco, U. (2000). *Apocalypse Postponed*, Bloomington, IN: Indiana University Press.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Foucault, M. (2005). *The Order of Things: An Archaeology of the Human Sciences*. (2da Ed. reimpresa. London: Routledge.

Gitelman, L. (ed.) (2013). *Raw Data is an Oxymoron*. Cambridge, MA: MIT Press.

Guattari, F. (2013). *Schizoanalytic Cartographies*. London: Continuum.

Illusory Correlation. (Marzo, 2019). En *Wikipedia:* http://en.wikipedia.org/wiki/Illusory_correlation [Consultado el 21 de marzo de 2019]

Ingold, D. and Soper, S. (Abril, 2016). Amazon Doesn't Consider the Race of Its Customers. Should It? *Bloomberg*. Recuperado el 21 de marzo de 2019 de: www.bloomberg.com/graphics/2016-amazon-same-day

Katz, Y. (2017). Manufacturing an artificial intelligence revolution (SSRN Scholarly Paper ID 3078224). Social Science Research Network. http://dx.doi.org/10.2139/ssrn.3078224

LeCun, Y. et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1* (4), pp. 541–551. https://doi.org/10.1162/neco.1989.1.4.541

LeCun, Y. (Julio, 2018). Learning World Models: the Next Step towards AI. *Conferencia magistral en la International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm, Sweden.

Leibniz, G. W. (1951). *Preface to the General Science. 1677*. Wiener, Leibniz: Selections.

Li, Guanpeng et al. (2017). Understanding error propagation in deep learning neural network (DNN) accelerators and applications. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, ACM.

Lipton, Z.C. (2016). The Mythos of Model Interpretability. *arXiv preprint*. Disponible en: https://arxiv.org/abs/1606.03490

McQuillan, D. (Junio, 2018). Manifesto on Algorithmic Humanitarianism. Presentado en el *Simposio Reimagining Digital Humanitarianism*, Goldsmiths, University of London. Disponible en: https://osf.io/preprints/socarxiv/ypd2s/download

McQuillan, D. (2018). People's Councils for Ethical Machine Learning. S*ocial Media and Society, 4* (2). https://doi.org/10.1177%2F2056305118768303

Moretti, F. (2013). *Distant Reading*. London: Verso Books.

Murgia, M. (Abril, 2019). "Who's using your face? The ugly truth about facial recognition", *Financial Times*. Disponible en: https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e

O'Neil, C. (2016). *Weapons of Math Destruction*, New York: Broadway Books.

Noble, S. *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York: NYU Press.

Pasquinelli, M. (2015). Anomaly Detection: The Mathematization of the Abnormal in the Metadata Society. Paper presentado en Transmediale. Disponible en: www.academia.edu/10369819. [Consultado el 21 de marzo de 2019].

Pasquinelli, M. (2017). Arcana Mathematica Imperii: The Evolution of Western Computational Norms. En: Maria Hlavajova et al. (eds.), *Former West* Cambridge, MA: MIT Press, pp. 281–293.

Vincent, J. (Octubre, 2018). Christie's sells its first AI portrait for $432,500, beating estimates of $10,000. *The Verge*. Recuperado el 21 de marzo de 2019 de: dewww.theverge.com/2018/10/25/18023266

Virilio, P. (1994). *The Vision Machine*. London/Bloomington: British Film Institute/Indiana University Press.

Vogl, J. (2007). Becoming Media: Galileo's Telescope. *Grey Room, 29*, pp. 14–25. https://doi.org/10.1162/grey.2007.1.29.14