Una breve discusión sobre la Inteligencia Artificial y sus usos

Bruno Constanzo, Valentina Fernández, Lucía Algieri, Ana Di Iorio
Laboratorio de Investigación y Desarrollo de Tecnología en Informática Forense
InFo-Lab (Universidad FASTA, Ministerio Público de la Provincia de Buenos Aires)
Mar del Plata, Buenos Aires

Abstract
En los últimos años se escuchan cada vez más y más discusiones acerca del uso de la Inteligencia Artificial en distintos ámbitos de la sociedad, y en particular mucho se dice de su uso por el Estado, y específicamente la Justicia. Se ven dos posturas marcadas: por un lado, los desarrolladores de herramientas forenses ponen a disponibilidad modelos de IA en sus productos, prometiendo la rápida, y casi mágica, resolución de todos los casos. En un extremo opuesto, llamados expertos en IA nos previenen sobre los sesgos de los algoritmos y los peligros a los que se expone la sociedad por el uso de esta tecnología. ¿Hay un punto medio entre estas posturas? ¿Es realmente la Inteligencia Artificial una bala de plata para resolver todos nuestros problemas, o incorporar estas herramientas es un paso que eventualmente nos llevará a nulidades?

Palabras Clave
Inteligencia Artificial, Evidencia Digital, Ética y Responsabilidad Social.

Introducción

Es innegable que la humanidad en su conjunto se ha sumido en lo que llamamos Sociedad de la Información: sistemas informáticos que son pervasivos a todas nuestras actividades, gestionan nuestros datos e información desde cuestiones básicas como pueden ser el uso de transporte público, hasta aspecto más complejos y críticos, como la gestión y procesamiento de datos médicos. En nuestro país, 87% de la población accede a internet, y 88% cuenta con un teléfono celular[1].

En esta informatización de nuestras vidas, cuanto más rápidos, eficientes, y precisos sean los algoritmos que se utilizan, tanto mejores serán los resultados: si procesar un dato tarda un minuto, será aplicable a un determinado grupo de interés. En cambio, si el mismo análisis pudiera realizarse en microsegundos, será posible analizar toda la población de una región, cambiando dramáticamente la escala y el impacto de este procesamiento de datos.

Por supuesto, para realizar esos análisis es necesario contar con datos útiles y confiables. En ese aspecto, los fabricantes de tecnología han respondido: hoy la mayoría de la población cuenta con dispositivos que reúnen información de manera constante y sutil. Los *smartphones* y los *smartwatches* son evidentes, pero también sitios web, el sistema operativo de un teléfono o computadora, y cualquier plataforma electrónica tienen la capacidad de recopilar datos (y muchas veces la ejercen).

La Inteligencia Artificial emerge en este ámbito como una herramienta capaz de extraer relaciones entre los datos, información y "conocimiento" del enorme volumen de datos, para ponerlo en manos de las personas, empresas, o instituciones y empoderarlas.

Parte del problema al que nos enfrentamos radica en que algunos de los proponentes de utilizar Inteligencia Artificial, aunque se presentan como expertos en la temática, carecen de conocimiento en las técnicas, tecnologías, y metodologías utilizadas para el entrenamiento de modelos de IA, la creación de soluciones, y a veces incluso el desarrollo de software. Y aún con esta falta de conocimiento, se prometen soluciones mágicas a problemas que se enfrentan las personas, las empresas y la sociedad, con una promesa de que como si producto o servicio "utiliza IA", entonces ha sido conferido con propiedades mágicas que no hace falta explicar, y sólo por eso es mejor que otras alternativas.

Ante este "engaño", y reforzados por escándalos de público conocimiento[1], surge un polo opuesto que es muy vocal en contra del uso de la Inteligencia Artificial, y nos intenta prevenir de catástrofes y males que caerán sobre la sociedad si cedemos ante su uso. En este grupo también será difícil encontrar gente que realmente conozca del funcionamiento de este tipo de sistemas.
Para abordar realmente las problemáticas del uso de la IA en la Justicia, se deberá entonces entender primero qué es la IA, qué límites tiene, y qué beneficios nos puede brindar. Sólo en ese marco se puede dar una discusión real y útil sobre su uso, las implicancias éticas, y si como sociedad debemos aceptar o rechazar su uso. Y por supuesto, la respuesta a la que se llegará es que una respuesta binaria y simple no es posible, que hay matices sutiles a considerar, y que en el fondo el buen o mal uso de una herramienta no está en la herramienta en sí, sino en quién la usa y cómo la usa.

Marco Teórico

Se explican a continuación algunos conceptos básicos de las computadoras, programación, e inteligencia artificial, a fin de que sea posible luego la discusión alrededor del problema que se plantea en este trabajo.

Computadoras y Programación

Las computadoras[2] son en primer lugar máquinas de calcular: operan sobre datos que se representan por medio de bits en una memoria, e implementan operaciones lógicas y matemáticas que permiten transformar estos datos[7].
La unidad más pequeña de información con la que una computadora puede trabajar es el *bit*, que puede tomar únicamente dos valores: 0 o 1, prendido o apagado, verdadero o falso.
Del agrupamiento de varios bits en unidades lógicas que respetan algunas reglas, surgen los tipos de datos básicos que maneja una computadora: números enteros, números de punto flotante, y cadenas de texto. De la combinación de estos, en registros o estructuras, surgen luego tipos de datos más complejos como lo son los arreglos en memoria, pilas, colas, listas, tablas de hashes y muchas otras estructuras de datos [8]. También es posible definir estructuras de datos propias, específicas para resolver un problema determinado.
Sobre cada tipo de datos pueden realizarse operaciones básicas, que tienen sentido en el contexto de los mismos:

- Dos números enteros pueden sumarse, restarse, multiplicarse y dividirse entre sí.
- Los números de punto flotante permiten manejar cierto nivel de precisión después del punto decimal, que dependerá de la magnitud del número con el que se está trabajando.
- Las cadenas de texto representan cada carácter de un alfabeto como un número en memoria, y podremos concatenar dos cadenas de texto para unirlas en una sola de mayor longitud, buscar o reemplazar una cadena de texto dentro de otra, o recortar una cadena de acuerdo a posiciones de inicio y fin para obtener una sub-cadena de menor longitud, entre otras operaciones.
- Las estructuras más complejas y tipos de datos abstractos definen también operaciones que tienen sentido en su contexto.
- Un arreglo en memoria es el agrupamiento de muchos datos de un mismo tipo, de manera que se organicen secuencialmente en la memoria, de acuerdo a una disposición en una, dos o más dimensiones. Un arreglo de una sola dimensión se suele llamar vector, de

[1] Por nombrar sólo algunos casos: el hombre que se enteró que iba a ser abuelo por las ofertas del supermercado[2, 3], las aspiradoras robot que mapean hogares[4], o las falencias en el uso de reconocimiento facial por la policía del Reino Unido[5, 6].
[2] En concepto de "computadora electrónica digital", sin entrar en los detalles de aquellas que no son ni electrónicas, ni digitales.

dos dimensiones matriz, y de más dimensiones simplemente se llama "arreglo de N dimensiones".

- Una pila es un conjunto de datos donde sólo nos importa el dato "de más arriba", que podemos sacar de la pila, o apilar otro dato encima.
- Una cola en cambio tiene dos posiciones, el frente y el fin de la cola. Cuando sacamos un dato se saca del frente, y si agregamos un dato nuevo se lo agrega al final.
- Las listas son una abstracción más general, que nos permite insertar o extraer datos de cualquier posición.

Nótese cómo a medida que se construyen estructuras más complejas, aumenta el nivel de abstracción y ya no es necesario considerar detalles finos: una lista en memoria puede ser una lista de números, una lista de cadenas de texto, o una lista de listas u otra estructura de datos.

Una de las cuestiones fundamentales para la revolución que han generado las computadoras es que es posible **programarlas**, es decir, alterar su comportamiento de manera que puedan resolver nuevos problemas. Por medio de la combinación de secuencias de instrucciones básicas, un programador puede entonces crear un nuevo programa que procese la información de manera distinta a cualquier otro, logrando así resolver un problema particular.

Se denomina algoritmo a una secuencia de pasos o instrucciones que toman datos de entrada, y los procesan de manera determinística. Es decir, siempre se siguen las mismas instrucciones, y siempre ante el mismo dato de entrada se debe obtener el mismo resultado[3].

Cuando se expresan las estructuras de datos, algoritmos, y la secuencia lógica en que deben utilizarse de manera conjunta, por medio de un lenguaje de programación, se obtiene el llamado "código fuente" de un programa. De la posterior aplicación de un compilador o intérprete sobre este código, se obtendrá entonces un programa, una secuencia de instrucciones de máquina que luego puede ser ejecutada en una computadora. bajo un determinado sistema operativo y entorno de ejecución.

Es posible combinar todos estos conceptos que se han explicado en un ejemplo concreto: un programa de edición de imágenes contará con algoritmos para cargar archivos guardados en formatos específicos, de manera de leer la información que contienen en su interior y representar una imagen en una estructura de datos en memoria, por ejemplo, una matriz de píxeles RGB (rojo, verde y azul, los colores que pueden detectar las células en la retina del ojo humano)[9]. Es decir, la imagen se representa como un arreglo de 3 dimensiones (ver Fig. 1).

Figura 1 - Una de las posibles formas de representar una imágen en memoria, como un arreglo de 3 dimensiones (alto, ancho, canales de color).

Sobre este arreglo será posible aplicar luego distintas operaciones que lo transformen de una u otra manera:

- Una operación de curvas de nivel de intensidad cambiará el valor de los píxeles RGB de manera que los colores sean más brillantes o más oscuros en base a su valor inicial.
- Una operación de recorte extraerá de la imágen completa una

[3] Esta definición entra un poco en conflicto con los algoritmos pseudo-aleatorios. Entrar en esos detalles y distinciones aporta poco al presente trabajo, y sólo haría más compleja la discusión.

sub-imagen que nos resulte de interés.
- Una operación de rotación aplicará una transformación geométrica sobre los datos, dando como resultado una nueva imágen.

Nótese que en el contexto del programa de edición de imágenes, la abstracción con la que interactúa el usuario gira alrededor del significado que damos los humanos a los datos, y en pocas ocasiones serán importantes los detalles de implementación o de las estructuras de datos – el usuario simplemente debe pensar en conceptos de imágenes, y es problema del programa implementar las estructuras y algoritmos adecuados de manera transparente.

Inteligencia Artificial

La Inteligencia Artificial puede definirse como el esfuerzo por automatizar tareas intelectuales normalmente llevadas a cabo por humanos[10]. En búsqueda de este objetivo, se utilizan principalmente herramientas y conceptos de las matemáticas, electrónica, computación, e informática, en conjunto con ideas de otras ciencias como pueden ser la psicología y la biología, entre otras.

Dentro de la IA como rama del conocimiento, encontramos el *machine learning*, un área específica enfocada en algoritmos que tienen la capacidad de aprender en base a datos, y más especializada dentro de esta, se encuentra el *Deep Learning*, que utiliza redes neuronales "profundas"[4] para lograr este cometido.

Central a todas las ramas de la Inteligencia Artificial, es la representación de la información por medio de señales que serán procesadas por el sistema inteligente. Y resulta central al planteo, modelado, diseño e implementación de estos sistemas, el uso de herramientas matemáticas.

Al adentrarse en el mundo de la IA es común escuchar hablar de vectores, matrices, y tensores[5], así como también de gradientes, transformaciones lineales y no lineales, espacios vectoriales, *embeddings*, y transformaciones geométricas, por mencionar tan solo algunos ejemplos.

Adentrarse en las explicaciones de todos estos conceptos requeriría un libro entero, pero ilustraremos brevemente un ejemplo.

Una de las redes neuronales más simples que se puede plantear es el llamado Perceptrón de McCulloch y Pitts[11].

Este modelo de procesamiento de información está inspirado en las neuronas biológicas, y plantea que para que una neurona artificial se active, la suma de todas sus entradas debe superar un cierto umbral de activación. Cuando esto sucede, la neurona se activa al 100%, sin dar lugar a activaciones parciales.

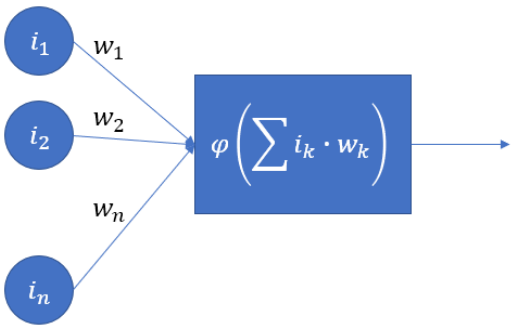

Figura 2 - Diagrama conceptual del perceptrón, donde cada entrada (*input*, i) tiene asociado un peso (*weight*, w) que la multiplica. Sobre la sumatoria de estos valores se aplica una función φ, que determina la activación de la neurona.

La Fig. 2 ilustra conceptualmente este tipo de neuronas. En general, no se habla de arquitecturas de perceptrón simple porque se propuso originalmente para ser utilizado como una única neurona que forma un clasificador binario, si bien luego en los 80's se plantearon extensiones y mejoras, como el perceptrón multicapa[12]. En la actualidad muchas redes convolucionales y modelos de *deep learning* utilizan en alguna instancia capas *fully connected*, que son efecto perceptrones multicapa,

[4] Es tema de debate qué convierte en "profunda" a una red neuronal en contraste con las redes "convencionales", dado que en general utilizan los mismos conceptos y herramientas, y la distinción entre una y otra es ligeramente arbitraria.

[5] Que pueden representarse como vectores, matrices y arreglos multidimensionales, las estructuras de datos, pero además tienen el contexto de la matemática que define operaciones claras y concretas sobre estas entidades.

contenidos dentro de una arquitectura más compleja.
La fórmula para calcular entonces el valor de una neurona será:

$$\varphi\left(\sum i_k \cdot w_k\right) \quad (1)$$

Que se puede expandir como:

$$\varphi(i_1 \cdot w_1 + i_2 \cdot w_2 + \cdots + i_n \cdot w_n) \quad (2)$$

Donde φ representa la función de activación elegida. Si representamos las neuronas de entrada y los pesos de cada una como vectores, entonces la operación de multiplicación entre cada peso sináptico y el valor de entrada correspondiente, y posterior sumatoria de todos estos resultados, puede resumirse como el producto punto entre estos ambos vectores. Para considerar múltiples neuronas, entonces se deberá extender el vector de pesos para que sea una matriz de pesos, y cada neurona tendrá representados los pesos de sus conexiones con la *inputs* (o las neuronas de la capa anterior) como una fila dentro de esta matriz.
Representar esto como código es simple utilizando Python[13] y la librería numpy[14], simplemente basta con definir los vectores y matrices pertinentes, e invocar el producto punto entre ellas. Las Figuras 3 y 4 muestran, como código y diagrama respectivamente, un ejemplo ilustrativo.

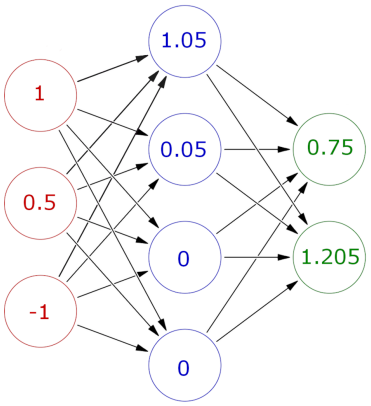

Figura 3 - Diagrama ilustrativo de un perceptrón multicapa, con los valores de las neuronas de entrada, intermedias y de salida. No se han indicado los pesos sinápticos, ver Fig. 4 para más detalles.

```
In [1]: import numpy as np

In [2]: inputs = np.array([1, 0.5, -1])

In [3]: weights = np.array([
   ...:     [ 2.0,  0.1,  1.0],
   ...:     [ 0.1,  0.1,  0.1],
   ...:     [-0.5,  0.5,  0.5],
   ...:     [-1.0, -1.0, -1.0],
   ...: ])

In [4]: def relu(x):
   ...:     x[x < 0] = 0
   ...:     return x
   ...:

In [5]: inputs.dot(weights.T)
Out[5]: array([ 1.05,  0.05, -0.75, -0.5 ])

In [6]: relu(inputs.dot(weights.T))
Out[6]: array([1.05, 0.05, 0.  , 0.  ])

In [7]: oweights = np.array([
   ...:     [0.7, 0.3, 0.5, 0.3],
   ...:     [1.1, 1.0, 0.1, 0.1],
   ...: ])

In [8]: relu(inputs.dot(weights.T)).dot(oweights.T)
Out[8]: array([0.75 , 1.205])

In [9]:
```

Figura 4 - Sesión en una terminal interactiva de IPython que muestra cómo se pueden representar las operaciones y cálculos de vectores y matrices de un perceptrón multicapa con numpy.

Es evidente la íntima relación entre la Inteligencia Artificial y la matemática. Comprender en detalle todas las implicancias, por supuesto, requeriría de los lectores conocer temas de Álgebra Lineal y Geometría Analítica, Análisis Matemático, Probabilidad y Estadística. Queda como ejercicio para los lectores profundizar en estos temas, si así les interesara. Para una introducción accesible a los temas de Inteligencia Artificial, pero aún así rigurosa, entretenida, y llena de ejemplos interesantes, se recomienda leer a Janelle Shane[15].
Ante el ejemplo brindado, puede surgir en la curiosidad del lector la pregunta: ¿cómo se determinan los valores de los pesos sinápticos? Para el ejemplo fueron elegidos de manera arbitraria, pero en una red neuronal real los pesos son aprendidos por ésta en la etapa de entrenamiento del modelo, donde un *algoritmo de aprendizaje* ajusta el valor de cada conexión de manera que se pueda llegar a aproximar un conjunto de resultados específico dentro de un márgen de error.

Los modelos de Inteligencia Artificial son modelos matemáticos, es decir, conjuntos de fórmulas y variables que se pueden ajustar con el fin de representar un sistema o fenómeno (físico, cognitivo, etc). Se utilizan al menos dos algoritmos distintos en su ciclo de vida: en primer lugar, un algoritmo de entrenamiento que permite ajustar los parámetros internos del modelo de manera que aprenda en base a un conjunto de ejemplos o datos de entrenamiento, y en segundo lugar un algoritmo de consulta o inferencia, que permite consultar al modelo ya entrenado para obtener un resultado en base a los datos de entrada que se le presentan.

Dependiendo cómo se realice el entrenamiento del modelo, se puede hacer una distinción en al menos tres categorías distintas:

- Aprendizaje supervisado: se llama con este nombre a los algoritmos y modelos que se entrenan presentando un dato de entrada y una etiqueta u objetivo que el modelo debe aprender a aproximar. Etiquetar los conjuntos de datos para que se puedan utilizar algoritmos de esta familia suele ser un proceso demandante y económicamente costoso.
- Aprendizaje no supervisado: se denomina así a los algoritmos y modelos que pueden trabajar sobre datos que no han sido etiquetados. En general los algoritmos no supervisados podrán realizar un agrupamiento de los datos de acuerdo a características comunes que tengan, pero es difícil lograr una clasificación en categorías o la regresión de valores, como sí permiten los algoritmos supervisados.
- Aprendizaje auto-supervisado: se denomina así a los algoritmos que tienen la capacidad de generar un puntaje o métrica objetivo en base a los datos de entrada, en el contexto de una simulación del problema que deben aprender a resolver. La optimización de ese puntaje o métrica permitirá al algoritmo de aprendizaje ajustar los parámetros del modelo.

En el caso del aprendizaje supervisado, al crear un nuevo modelo de IA, los parámetros de este se inicializan de acuerdo a algún criterio, en general con valores aleatorios. Luego se van presentando los ejemplos de entrenamiento, y se calcula el error, la diferencia entre los resultados obtenidos y el resultado objetivo que debiera haberse obtenido. En base a este error, y aplicando las fórmulas correspondientes para el tipo de modelo que se está entrenando, se determina cómo deben ajustarse sus parámetros internos. Una vez que se han presentado todos los ejemplos de entrenamiento, se dice que se ha cumplido una "época", y se vuelve a comenzar el proceso. A medida que avanzan las épocas de entrenamiento, el modelo irá disminuyendo el error, es decir, aprenderá en base a los datos[6].

Algunos de los modelos de *machine learning* más conocidos son:

- Árboles de decisión y bosques aleatorios: son modelos conceptualmente simples y fáciles de interpretar. Una vez entrenados, en sus nodos consultan si una variable cumple o no con un criterio determinado y se abre el espacio de decisión en ramas que, en sus nodos subsiguientes, consultan otros criterios. En las hojas del árbol se llega a un valor o categoría resultado.
- *Support-Vector Machines* o Máquinas de Vectores de Soporte, son modelos matemáticos que permiten, por medio de la solución de un problema de optimización, encontrar el hiperplano que separa en dos regiones un espacio de clasificación. Pueden extenderse

[6] Esto no siempre sucede, y será el trabajo de quienes entrenen el modelo determinar cómo ajustar los hiperparámetros para mejorar el aprendizaje.

para considerar múltiples clases, y también para lograr barreras de separación no lineales. Durante los 90 fueron ampliamente utilizados, y desplazaron a los perceptrones multicapa en amplitud de uso.

- Redes Neuronales Artificiales: son modelos que representan una serie de neuronas interconectadas entre sí, de manera que procesan información y, siguiendo sus conexiones, se parte de las neuronas de entrada y eventualmente se llega a las neuronas de salida, que codifican un resultado. En los últimos años, se han desarrollado una gran cantidad de variantes. En general son modelos complejos, y que puede resultar difícil determinar por qué llegan a una decisión o resultado.

Discusión

Habiendo delineado el panorama desde lo teórico, ya estamos en condiciones de plantear la discusión acerca de la pertinencia o no del uso de los modelos de Inteligencia Artificial, las ventajas reales que conllevan, y también las limitaciones y desventajas que pueden presentar.

<u>Críticas a la IA</u>

Las críticas a los sistemas de Inteligencia Artificial son variadas, y van desde el temor a que el uso de este tipo de tecnologías dejará obsoletos a los humanos en determinadas tareas, hasta cuestiones reales y actuales vinculadas con los sesgos y la llamada discriminación algorítimica.

La cuestión de los sesgos es preocupante en la medida que actualmente está comenzando a afectar a porciones de la sociedad. Se denomina sesgo algorítmico a cualquier situación en la cual, ante la aplicación de un programa o algoritmo de decisión, se da un trato desigual a un grupo de personas por alguna característica. Y no es una cuestión nueva, ya en los años 70's y 80's el tema se comenzó a discutir, y se dieron casos como el de la discriminación a mujeres y estudiantes extranjeros que pretendían estudiar en la St. George's Hospital School of Medicine [16].

En particular cuando se utilizan algoritmos para tomar decisiones, se da una impronta de que, al ser secuencias de órdenes y operaciones lógicas y matemáticas, llevadas a cabo por una computadora, claramente no habrá animosidad ni prejuicios intervinientes. Lo que ignora esta visión es algo simple y evidente para cualquiera que haya programado: podemos indicar a un programa que haga lo que nosotros queramos.

Por ejemplo, en una empresa que gestionara los sueldos con un sistema informático, podría añadirse una sentencia para pagar un 10% adicional a aquellos cuyo nombre comienza con una determinada letra, o restarle un 10% a otros en base al mismo criterio. Este hecho encajaría perfectamente en la definición de sesgo algorítmico dada, aún cuando estamos hablando de instrucciones llevadas a cabo por un programa, y solamente utilizando sentencias lógicas y operaciones de datos básicas.

Entonces, ¿de dónde proviene el sesgo algorítmico? Se pueden identificar claramente al menos dos posibles orígenes para estas situaciones.

En primer lugar, se puede dar una situación similar a la planteada en el ejemplo: por alguna razón, se codifica en un programa una serie de reglas que terminan afectando negativamente a un grupo de personas, o beneficiando a otras, sin una razón justa. Esto en ocasiones sucede de manera intencional, y en esos casos se habla también del fenómeno denominado *mathwashing*, en el cual se justifica cualquier resultado obtenido (incluso los resultados injustos) bajo la suposición que si proviene de aplicar una fórmula, análisis estadístico, datos, y/o algoritmos, entonces es confiable, neutral, y sin prejuicios.

En otras ocasiones sucede que se aplican técnicas estadísticas, de análisis de datos, o de *machine learning*, y por la selección de datos de entrenamiento realizada se están

incluyendo sesgos, que el modelo aprenderá y luego aplicará. Si no se está atento a detectar este tipo de situaciones, entonces el modelo aplicado seguirá replicando esta discriminación, de manera automática y sin supervisión humana.
Para detectar ambas instancias, entonces es importante que se preste atención tanto a las reglas que se codifican, como a los datos de entrenamiento, y siempre a los resultados de aplicar algoritmos o modelos entrenados. Es decir, con ser conscientes que estas situaciones pueden darse, de manera intencional o accidental, se podrán tomar medidas para evitarlas.

Limitaciones de la IA

Es importante también conocer los límites de la tecnología, y saber distinguir sus falencias de sesgos.
En el año 2020, Colin Madland recurrió a Twitter[17] para contar que uno de sus compañeros de trabajo estaba siendo discriminado por el detector de rostros de la herramienta de videoconferencia que utilizaban a diario, y no podía utilizar los filtros para ocultar su fondo, ya que la misma al no detectar su rostro, lo eliminaba de la imagen.
Para sorpresa de Madland, al cargar las imágenes en Twitter descubrió que la red social recortaba de todas las imágenes a su colega afroamericano. De repetidos intentos, no logró en ninguna instancia que Twitter lo incluyera en las vistas previas que acompañaban sus quejas[7].
Para hacer una prueba simple, es posible utilizar la librería dlib[19], en particular su detector de rostros basado en la técnicas de Histograma de Gradientes Orientados[8] (usualmente abreviado HOG por las siglas en inglés, *Histogram of Oriented Gradients*), para hacer una prueba sobre las imágenes compartidas por Madland en Twitter. Podemos ver en la Fig. 5 cómo al utilizar el detector de rostros de dlib, también se detecta únicamente un rostro, de los dos que se encuentran visibles en la imagen.

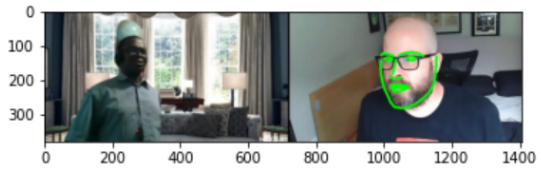

Figura 5 - Resultado de aplicar el detector de rostros HOG de dlib sobre la imagen, seguido de un modelo de *landmarking*[9] también incluido en dlib.

¿Estamos en condiciones de decir que este es un algoritmo racista? Si nos adentramos en los detalles de cómo funciona el descriptor HOG, nos encontraremos que es una técnica que, al estar basada en los gradientes de la imágen, está fuertemente afectada por sus contrastes. Si nos enfocamos entonces sobre los rostros en la imágen, y calculando el gradiente mediante el método de Sobel, podremos ver cómo está "viendo" el algoritmo (Fig. 6):

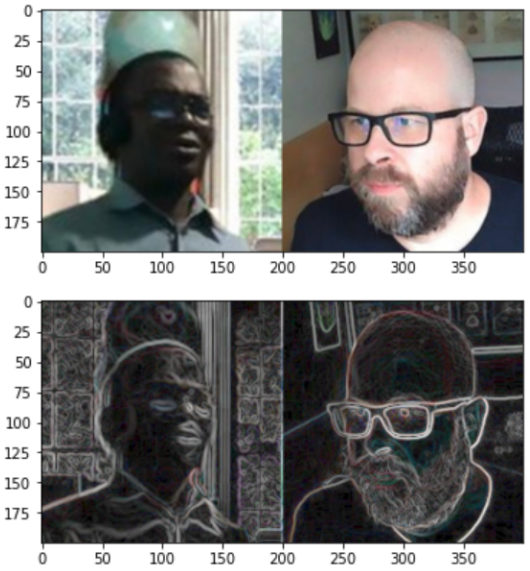

Figura 6 - La imagen en cuestión con los rostros de las personas, y el gradiente aproximado por el método de Sobel.

Se puede apreciar en la Fig. 6 que, al transformar la imagen con el filtro de Sobel, uno de los rostros mantiene

[7] Ante esta situación, Twitter se comprometió a mejorar este sistema. Luego de reiterados intentos fallidos, fue dado de baja y se reemplazó por un algoritmo más simple[18].
[8] Similar al trabajo de Dalal, Triggs y Schmid del año 2006[20].
[9] Se denomina *landmarking* al proceso de identificar puntos clave en una figura o imagen, en este caso, puntos característicos del rostro: el contorno, la nariz, cejas, ojos, boca, etc.

características distintivas, pero el otro está opacado casi en su totalidad por la falta de contraste.
Existen algunas técnicas de procesamiento de imágenes que se pueden utilizar para compensar este tipo de situaciones, una de ellas es el *Contrast Limited Adaptive Histogram Equalization* (o CLAHE)[21], que se encuentra disponible en la librería OpenCV[22]. Mediante la aplicación de esta técnica, es posible obtener los siguientes resultados (Fig. 7):

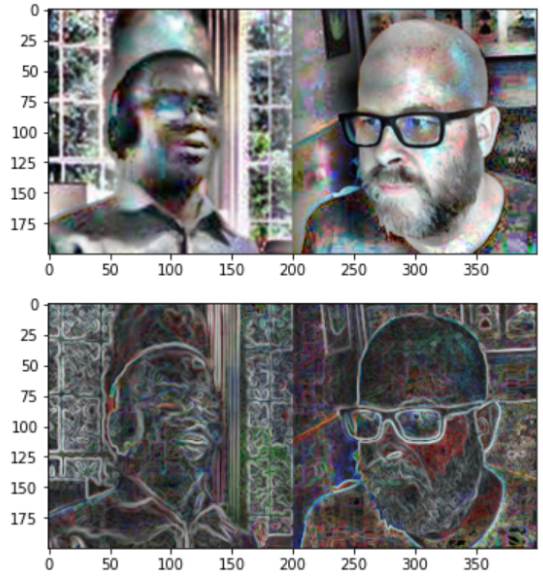

Figura 7 - La imagen en cuestión con los rostros de las personas, procesada con el algoritmo de CLAHE, y el gradiente (Sobel) sobre esta versión.

Se puede apreciar como luego de aplicar el algoritmo de ecualización de histograma, el mapa de gradientes de la imagen ha cambiado notablemente. Si se aplica sobre esta versión de la imagen el detector de rostros de dlib nuevamente, ésta vez si se detectarán ambos rostros (Fig. 8):

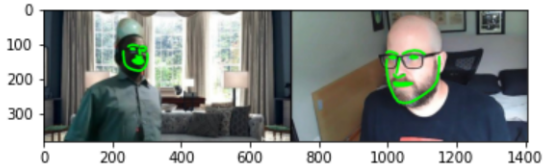

Figura 8 - Resultado de la detección y *landmarking* de rostros luego de aplicar CLAHE sobre la imagen.

En este caso particular del detector de rostros basado en HOG, la situación a la que nos enfrentamos no se trataba de un algoritmo racista, sino un algoritmo que no estaba "viendo" adecuadamente, y al cual fue necesario preprocesarle la imágen antes de que pudiera funcionar de manera adecuada.
Existen muchas otras situaciones en las que los algoritmos de Inteligencia Artificial y *machine learning* no se comportan de la manera en que esperamos.
Si bien ha habido grandes avances en los últimos años en los campos de visión artificial, procesamiento de lenguaje natural, traducción de idiomas, generación de contenido (ya sea texto, imágenes, sonido, y/o video), los algoritmos de Inteligencia Artificial siguen teniendo aún hoy falencias.
Por estas razones es importante conocer cómo funcionan, en qué condiciones pueden fallar, qué podemos hacer como desarrolladores de tecnología para compensar sus limitaciones en el uso. Y por sobre todo, explicar al resto de la sociedad que se enfrenta a estas nuevas herramientas con desconocimiento, y a veces temor, de qué manera funcionan, qué es lo que pueden y no pueden hacer, y en qué medida se deben controlar.
En ese sentido, es sumamente importante cuando se utilizan este tipo de sistemas en el Estado, contar con una suite de tests adecuados que permitan conocer el rendimiento de los sistemas inteligentes, entre otras métricas: su tasa de aciertos, su tasa de error, falsos positivos, falsos negativos, F-score, y modos de fallo. Ya se citó anteriormente el caso de la Policía Metropolitana del Reino Unido[5, 6], pero vale la pena retomarlo en este punto. Si se hubieran publicado las métricas obtenidas por Fussey y Murray con anterioridad a la aplicación de este sistema, seguramente la sociedad no hubiera aceptado su puesta en marcha.

<u>Beneficios y ventajas de la IA</u>
La capacidad de automatizar tareas como el reconocimiento de patrones o la búsqueda de información con un cierto significado semántico, que los algoritmos clásicos no

podían realizar, es una de las más grandes ventajas de la Inteligencia Artificial.
El correcto uso de esta capacidad permite minimizar tareas repetitivas, que demandan un bajo nivel de atención, pero suficiente para resultar desgastantes.
No sólo eso, también en ocasiones el uso de IA permite evitar que una persona se vea expuesta a material que podría resultar perturbador o afectarla psicológicamente. Por ejemplo, los peritos informáticos e investigadores digitales que trabajan casos donde deben interactuar con material de abuso sexual infantil se ven muy seriamente afectados por estas tareas[23].
Sin llegar a casos tan extremos, cuando se require la búsqueda de un dato determinado en un volúmen de evidencia digital vasto, compuesto por decenas o cientos de miles de archivos, se vuelve prácticamente imposible en tiempos humanos llevar a cabo una tarea, por más que se trate de algo simple. Por plantear algunos ejemplos:

- Encontrar todas las direcciones que se mencionan en un conjunto de mensajes de WhatsApp.
- Detectar instancias en donde una persona amenazó a otra.
- Buscar todas las fotografías donde se ve un determinado objeto.
- Encontrar todas las instancias donde está escrita una palabra, ya sea en texto, capturas de pantalla, fotografías, audio o incluso video.

En todas estas instancias, hoy es posible utilizar distintas tecnologías y modelos de *machine learning* para convertir algo extremadamente demandante en tiempo y esfuerzo, en una tarea donde las computadoras pueden realizar un filtrado de los datos para luego presentar a los operadores humanos un subconjunto de datos de interés, en donde serán estos últimos los encargados de decidir si estos resultados provisorios son realmente pertinentes o no para la investigación o el peritaje.
Hoy una computadora puede "leer" texto miles de veces más rápido que cualquier ser humano, y "ver" cientos de imágenes por segundos en búsqueda de objetos[10], sin reducir su rendimiento por cansancio. Por supuesto, el nivel de comprensión que logran en estas tareas es menor de lo que lograría un ser humano.
De esta manera, los modelos de Inteligencia Artificial toman el lugar que propiamente les corresponde: no en reemplazo del criterio humano, ni la toma decisiones en nuestro lugar, sino funcionando como una herramienta que aumente las capacidades de las personas.

Conclusión

A lo largo del trabajo se hizo una exposición desde el funcionamiento más básico de las computadoras y algunos conceptos y algoritmos de inteligencia artificial, para luego abordar algunas de las críticas, falencias y beneficios que nos puede brindar este tipo de tecnologías.
Si bien existen falencias y situaciones en donde el comportamiento de los sistemas inteligentes falla, en general esto se debe a que hemos puesto en ellos exigencias que no pueden cumplir. Cuando los algoritmos de *machine learning* aprenden a discriminar, no es por una tendencia inherente en ellos, sino porque han sido entrenados con datos sesgados – datos que fueron etiquetados por humanos en primer lugar.
Es necesario elevar el nivel de exigencia que ponemos sobre el uso de estas tecnologías: no alcanza con saber si un algoritmo es preciso en un cierto porcentaje de los casos, es imperativo conocer en profundidad las métricas y los casos de test que se tomaron para realizar dicha evaluación.
Aún en esa situación, siempre debe considerarse a los algoritmos inteligentes como una herramienta que nos permite afrontar los nuevos desafíos que nos impone la sociedad de la información: un volúmen más grande de datos e

[10] Incluso decenas de miles, aunque encima de esa escala empiezan a ser limitantes del rendimiento algunas cuestiones del hardware y la forma en que se organizan los datos, un tema que será abordado en otro trabajo.

información. Y ante la vorágine y la velocidad de la sociedad moderna, usualmente son necesarias respuestas más rápidas.

No debemos dejar pasar la ventana de oportunidad que se nos ha presentado, de utilizar las tecnologías de la información, y en particular la Inteligencia Artificial, para potenciar las capacidades de nuestros operadores de Justicia.

Agradecimientos

Agradecemos a la Universidad FASTA, y el Ministerio Público Fiscal de la Provincia de Buenos Aires, por su apoyo constante al Laboratorio de Investigación y Desarrollo de Tecnología en Informática Forense, sin el cual este trabajo no hubiera sido posible.

Agradecemos también a Fundar por su apoyo con la beca FunDatos, fundamental para la participación en este trabajo de Valentina Fernández.

Referencias

1. "Informe Técnico Vol. 6, Nro. 89: Acceso y uso de tecnologías de la información y la comunicación. EPH. Cuarto Trimestre de 2021". Instituto Nacional de Estadísticas y Censos, Buenos Aires, Mayo 2022.
2. "How Target Figured Out A Teen Was Pregnant Before Her Father Did", K. Hill. Forbes, Febrero 2012. Disponible online: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/.
3. "How Companies Learn Your Secrets", C. Duhigg. New York Times, Febrero 2012. Disponible online: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.
4. "Your Roomb May Be Mapping Your Home", M. Astor. The New York Times, Julio 2017. Disponible online: https://www.nytimes.com/2017/07/25/technology/roomba-irobot-data-privacy.html.
5. "81% of 'suspects' flagged by Met's police facial recognition technology innocent, independent report says", R. Manthorpe, A. J. Martin. Sky News, Julio 2019. Disponible online: https://news.sky.com/story/met-polices-facial-recognition-tech-has-81-error-rate-independent-report-says-11755941.
6. "Policing Uses of Live Facial Recognition in the United Kingdom", P. Fussey, D. Murray. Regulating Biometrics: Global Approaches and Urgent Questions, AI Now Institute, 2020. Disponible online: https://ainowinstitute.org/regulatingbiometrics-fussey-murray.pdf.
7. "Organización de Computadoras: Un enfoque estructurado", A. S. Tanenbaum. Prentice Hall, 2000.
8. "Estructuras de Datos y Algoritmos", A. Aho, J. Hopcroft, J. Ullman. Addison-Wesley, 1988.
9. "Digital Image Processing, 3rd edition", R. C. Gonzalez, R. E. Woods. Pearson Prentice Hall, 2008.
10. "Deep Learning with Python, 2nd edition", F. Chollet. Manning Publications, 2021.
11. "A Logical Calculus of Ideas Immanent in Nervous Activity", W. McCulloch, W. Pitts. Bulletin of Mathematical Biophysics, 1943.
12. "Learning Internal Representations by Error Propagation", D. Rumelhart, G. Hinton, R. Williams. Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press, 1986.
13. "Python 3 Reference Manual", G. Van Rossum, F. Drake. CreateSpace, 2009.
14. "Python for Scientific Computing", T. Oliphant. Computing Science and Engineering, Volúmen 9, Número 3, Junio 2007.
15. "You Look Like a Thing and I Love You", J. Shane. Wildfire, Headline Publishing Group, 2019.
16. "A blot on the profession", S. Lowry, G. Macpherson. British Medical Journal, 296, Marzo 1988.
17. Hilo de tweets, C. Madland, disponible online: https://twitter.com/colinmadland/status/1307111816250748933.
18. "Sharing learnings about our image cropping algorithm", R. Chowdhury, Twitter Engineering, Mayo 2021. Disponible online: https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm.
19. "Dlib C++ library", D. King. Disponible online: http://dlib.net.
20. "Human Detection Using Oriented Histograms of Flow and Appearance". N. Dalala, B. Triggs, C. Schmid. European Conference on Computer Vision 2006.
21. "Adaptive Histogram Equalization and Its Variations", S. M. Pizer, E. P. Amburn, J. D. Austin, et al. Computer Vision, Graphics, and Image Processing 39, 1987.
22. "The OpenCV Library", G. Bradski. Dr. Dobb's Journal of Software Tools, Noviembre 2000.
23. "The Psychological Effects Experienced by Computer Forensic Experts Working with

Child Pornography", J. Whelpton. Masters Thesis, University of South Africa, Febrero 2012.

Datos de Contacto:
Bruno Constanzo. InFo-Lab, Universidad FASTA. bconstanzo@ufasta.edu.ar.
Valentina Fernández. InFo-Lab, Universidad FASTA. valentinaf@ufasta.edu.ar.
Lucía Algieri. InFo-Lab, Universidad FASTA y Ministerio Público de la Provincia de Buenos Aires. luciaalgieri@ufasta.edu.ar.
Ana Haydée Di Iorio. InFo-Lab, Universidad FASTA y Ministerio Público de la Provincia de Buenos Aires. diana@ufasta.edu.ar.