Este artículo es copia fiel del publicado en la revista Nueva Sociedad Nº 294,
julio-agosto de 2021, ISSN: 0251-3552, <www.nuso.org>.

Inteligencia artificial y sesgos algorítmicos

¿Por qué deberían importarnos?

Enzo Ferrante

Cada vez más, la inteligencia artificial es parte de nuestras vidas, a menudo de manera imperceptible. Ya no se trata de utopías tecnológicas sobre el futuro, sino de un presente muy concreto. Pero detrás de avances que incluyen desde diagnósticos médicos hasta vigilancia masiva están los algoritmos, cuyos «sesgos» amenazan con perpetuar e incluso profundizar las desigualdades del presente. Poner el foco en los datos, los modelos y las personas puede servir para construir una inteligencia artificial más «justa».

Una cámara enfoca las escalinatas de entrada del Instituto de Tecnología de Massachusetts (MIT). La investigadora Joy Buolamwini sube algunos escalones y se escucha su voz en *off*:

> Una de las cosas que me atrajeron de las ciencias de la computación fue que podía programar y alejarme de los problemas del mundo real. Quería aprender a hacer tecnología que fuera interesante. Así que vine al MIT y trabajé en proyectos de arte que usaban visión artificial[1].

Enzo Ferrante: estudió Ingeniería de Sistemas en la Universidad Nacional del Centro de la Provincia de Buenos Aires (Unicen) y se doctoró en Informática en la Université Paris-Saclay y el Institut National de Recherche en Informatique et en Automatique (INRIA) de Francia. Realizó su posdoctorado en el Imperial College de Londres y regresó a Argentina como científico repatriado. Actualmente es investigador adjunto del Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina (Conicet) y docente en la Universidad Nacional del Litoral (UNL), donde trabaja en el desarrollo de métodos de aprendizaje automático para el análisis de imágenes biomédicas.

Palabras claves: aprendizaje automático, desigualdades, inteligencia artificial, sesgo.

1. La visión artificial (*computer vision*) es una rama de las ciencias de la computación que se encarga de construir algoritmos y programas de computación capaces de identificar contenido en las imágenes e interpretarlas.

> Durante mi primer semestre en el Media Lab hice un curso sobre «invención científica». Lees ciencia ficción y eso te inspira a crear algo que seguramente sería poco práctico si no tuvieras el curso como excusa para hacerlo. Yo quise construir un espejo que me inspirara por las mañanas. Lo llamé Espejo Aspire. El espejo me colocaba leones sobre el rostro, o gente que me inspirara, como Serena Williams. Le coloqué una cámara y con un *software* de visión artificial, se suponía que debía detectar los movimientos de mi cara. Pero el problema era que no funcionaba bien, hasta que me puse una máscara blanca. Cuando me ponía la máscara, me detectaba. Cuando me la quitaba, ya no me detectaba.

Así comienza *Prejuicio cifrado* (*Coded Bias*), el documental dirigido por la cineasta Shalini Kantayya y estrenado en 2020 que narra cómo Buolamwini tomó conciencia del sesgo racial existente en los algoritmos de reconocimiento facial y analiza sus consecuencias. Buolamwini es una mujer negra, especialista en informática, activista y fundadora de la Liga por la Justicia Algorítimica (Algorithmic Justice League), y hace algunos años descubrió que varios sistemas comerciales de reconocimiento facial diseñados por Amazon, IBM y Microsoft funcionaban mejor con el rostro de sus amigos blancos que con el suyo[2]. Más allá de los dilemas éticos sobre el desarrollo de sistemas de reconocimiento facial[3], el caso de Buolamwini muestra claramente cómo un sistema basado en inteligencia artificial puede adquirir un sesgo y cumplir mejor la tarea para la que fue diseñado en un grupo de individuos que en otro.

Esta no es una cuestión menor. La expresión «inteligencia artificial» dejó de ser propiedad exclusiva de las novelas de ciencia ficción y de los libros de computación. Noticias sobre avances fascinantes –como computadoras capaces de asistir al personal médico en tareas de diagnóstico o de manejar automáticamente vehículos no tripulados– aparecen cada vez con más frecuencia y se vinculan cada vez más con nuestras vidas. Sin embargo, no todas las noticias son tan alentadoras. Lo que experimentó Buolamwini no es un caso aislado: durante los últimos años, hemos visto desde sistemas para reconocimiento facial que alcanzan un peor rendimiento[4] en mujeres de piel negra que en hombres blancos, hasta traductores del inglés al español que perpetúan estereotipos de género. Estos ejemplos ilustran un fenómeno

2. Si bien estos sistemas son más conocidos por su uso en vigilancia masiva o publicidad, también es posible encontrarlos en otros contextos, como en cámaras fotográficas (para hacer foco en el rostro de manera automática) o en redes sociales (para etiquetar a personas de manera automática).
3. «Facial-Recognition Research Needs an Ethical Reckoning», editorial en *Nature*, 18/11/2020.
4. En el contexto de este artículo, utilizaremos el término «rendimiento» de un sistema de inteligencia artificial para hacer referencia al nivel de acierto de las predicciones que realiza.

conocido como «sesgo algorítmico»: sistemas cuyas predicciones benefician sistemáticamente a un grupo de individuos frente a otro, resultando así injustas o desiguales. Pero ¿cuáles son las razones que llevan a estos sistemas a generar predicciones sesgadas? Para entenderlo, comencemos por definir algunos conceptos que nos serán útiles a lo largo de este ensayo: «inteligencia artificial» y «aprendizaje automático».

Cuando la inteligencia deviene artificial y el aprendizaje, automático

Existen muchas definiciones de «inteligencia artificial». Aquí usaremos una definición general ofrecida en uno de los libros fundamentales del campo, que describe la inteligencia artificial como la disciplina que se encarga de comprender y construir entidades inteligentes (pero artificiales)[5]. Esta definición es muy amplia y abarca conceptos que van desde los sistemas de razonamiento deductivo basados en reglas lógicas hasta algoritmos de aprendizaje automático que buscan detectar automáticamente patrones en conjuntos de datos y luego usarlos para realizar predicciones[6]. Un elemento central para este último subcampo de la inteligencia artificial son entonces los datos, que constituyen la materia prima utilizada para automatizar el proceso de aprendizaje en el que los sistemas son entrenados para realizar predicciones.

Un elemento central para este último subcampo de la inteligencia artificial son los datos

Los datos pueden ser imágenes, sonidos, texto escrito, redes, posiciones de un GPS, tablas o cualquier representación que se nos ocurra. En todo caso, la idea central es que los modelos de aprendizaje automático *aprenden* a partir de los datos. Esta noción resulta central en la actualidad, dado que la gran mayoría de las tecnologías disruptivas adoptadas masivamente en el siglo XXI y que son presentadas como inteligencia artificial utilizan en realidad métodos de aprendizaje automático. Pero ¿cómo aprenden estos sistemas?

Existen distintos paradigmas de aprendizaje. Uno de los más utilizados es el del aprendizaje supervisado, en el que los sistemas son sometidos a un proceso de entrenamiento que es guiado por anotaciones o etiquetas. La idea es simple: se intenta asociar características o patrones propios de los datos con las correspondientes etiquetas. Es decir, se analizan los datos en busca de

5. Peter Norvig y Stuart Russell: *Artificial Intelligence: A Modern Approach*, Pearson, Londres, 2002.
6. Kevin P. Murphy: *Machine Learning: A Probabilistic Perspective,* MIT Press, Cambridge, 2012.

patrones distintivos que permitan separar una categoría de la otra. Tomemos un ejemplo: imaginemos que queremos entrenar un sistema para que pueda decirnos si el contenido de una imagen corresponde a un perro o a un gato. Bajo el paradigma del aprendizaje supervisado, lo que necesitaremos es una base de datos compuesta por imágenes de perros y gatos, con la correspondiente etiqueta asociada a cada una. Durante el proceso de entrenamiento, el algoritmo tomará esas imágenes y comenzará a hacer predicciones a partir de ellas, asociando características (información de la imagen) con etiquetas. De forma simplificada, podemos pensar que estas características están dadas por diferentes patrones presentes en la imagen, como el color, el brillo, la cantidad de patas, el tamaño del cuerpo o la forma de las orejas. Si nos detenemos a pensar en estas características, algunas serán más útiles que otras para distinguir entre perros y gatos. Por ejemplo, la cantidad de patas no parece ser una característica útil para diferenciarlos; sin embargo, el tamaño del cuerpo sí podría serlo. La idea es que, por medio del entrenamiento, los sistemas aprendan a asociar patrones en estas características con las correspondientes categorías. Al principio estas asociaciones serán seguramente incorrectas; pero a medida que avance el proceso de entrenamiento, el modelo se irá ajustando y mejorando su desempeño en la tarea asignada.

Esta idea que ilustramos con imágenes es extrapolable a otros tipos de datos sobre los que hablábamos: si quisiéramos entrenar un sistema para aprender a traducir texto de inglés a español, necesitaríamos muchos textos escritos en ambos idiomas. Para inferir el estado de ánimo de una persona a partir de su voz, necesitaríamos grabaciones de audio de personas hablando, y la correspondiente etiqueta que indique si se encuentran alegres o tristes. Si pensáramos en un sistema que detecte patologías automáticamente a partir de imágenes radiográficas, necesitaríamos pares de imágenes con su correspondiente diagnóstico médico. O si quisiéramos entrenar un modelo para detectar rostros en imágenes, necesitaríamos una base de datos de fotografías de personas, con etiquetas que indiquen en qué lugar se encuentra el rostro de cada una.

Como vemos, los datos juegan un rol esencial en el entrenamiento de sistemas por medio de aprendizaje automático, dado que son la fuente de información que le indicará al sistema cuándo ha llegado a conclusiones correctas y cuándo no. Algo que resulta fundamental en este proceso, y que no siempre es tenido en cuenta, es que un sistema raramente se construye para realizar predicciones con los datos con que fue entrenado. Por el contrario, se espera que los modelos puedan sacar conclusiones acertadas sobre datos nunca vistos durante el «aprendizaje» –los datos de prueba– y cuyas etiquetas no se conocen. Esta capacidad de generalización es un rasgo primordial, dado que de nada serviría un modelo predictivo que solo acertara

en situaciones conocidas. Imaginemos un detector de patologías en imágenes radiográficas que puede predecir si una persona tiene o no neumonía utilizando solamente imágenes de esa misma persona. O un traductor de inglés a español que solo puede traducir textos que ya estaban traducidos. En general, la hipótesis de trabajo de estos sistemas es que los datos de prueba serán de alguna manera similares a los datos de entrenamiento, pero no los mismos. Por ejemplo, si entrenamos un modelo para detectar neumonía en humanos, el modelo será utilizado en otros humanos, pero no en animales. O si entrenamos un sistema para traducir del español al inglés, los textos de prueba serán distintos de los de entrenamiento, pero estarán siempre escritos en español, y no en francés. En este caso, resulta evidente que un sistema que aprendió utilizando textos en español no podrá generalizar al francés. ¿O acaso le pediríamos a un intérprete de francés que traduzca mandarín? Sin embargo, existen variaciones entre los datos de entrenamiento y prueba que pueden ser más sutiles que el cambio de español a francés o de humanos a animales, pero que igualmente producen un efecto devastador en la calidad de las predicciones.

Volvamos a imaginar el caso del sistema para distinguir entre imágenes de perros y gatos, pero con una pequeña variación: nuestra base de datos solo está compuesta por perros negros y gatos blancos. En este caso, el color del animal será una característica sumamente útil para distinguir entre ambas clases. De hecho, nos dará una predicción perfecta: si el color predominante en el cuerpo del animal es negro, será un perro; y si es blanco, será un gato. Ahora imaginemos que en nuestro conjunto de prueba hay una sutil diferencia: aparecen perros de color blanco. ¿Qué creen que sucederá con las predicciones sobre los perros blancos? El sistema seguramente les asignará la etiqueta «gato» de forma incorrecta, resultando en un rendimiento más bajo para este subconjunto de la población objetivo. Tomar en cuenta estos factores al entrenar sistemas de inteligencia artificial basados en aprendizaje automático es clave si queremos evitar el sesgo algorítmico en varios sentidos. Veamos algunos ejemplos.

Sobre datos, modelos y personas

Hace algunos años, llegó a mis manos, por recomendación de colegas, un artículo que se titulaba «AI is Sexist and Racist. It's Time to Make it Fair»[7] [La inteligencia artificial es sexista y racista. Es hora de volverla justa], de James Zou y Londa Schiebinger. El artículo discutía un aspecto sobre el

7. J. Zou y L. Schiebinger: «AI is Sexist and Racist –It's Time to Make it Fair» en *Nature*, 18/7/2018.

que hasta ese momento no me había detenido a pensar respecto de los modelos de inteligencia artificial que yo mismo estaba implementando: estos modelos pueden ser sexistas y racistas. En otras palabras, pueden adquirir un sesgo que los lleve a presentar un rendimiento dispar en grupos caracterizados por distintos atributos demográficos, lo que redunda en un comportamiento desigual o discriminatorio. Y una de las razones detrás de este comportamiento eran justamente los datos que usaba para entrenarlos.

Estos modelos pueden ser sexistas y racistas. En otras palabras, pueden adquirir un sesgo

Los ejemplos de sesgo algorítmico adquirido a través de los datos son variados y muchas veces tienen que ver con bases de datos que no representan en realidad al conjunto de la población. En el caso reportado por Joy Bowlamwini y Timnit Gebru[8], en el que diversos sistemas comerciales de reconocimiento facial muestran un rendimiento dispar respecto a variables demográficas como el género y el color de la piel, son las mujeres de piel negra el grupo para el cual los modelos presentan peor rendimiento. Este hecho está posiblemente relacionado con la falta de representatividad de mujeres negras en las bases de datos utilizadas para el entrenamiento. Ejemplos similares se encuentran al analizar ImageNet, una de las bases de datos de imágenes etiquetadas más grandes del mundo, que ha sido motor del desarrollo de los modelos más populares de clasificación de imágenes[9]. ImageNet posee millones de imágenes clasificadas en miles de categorías. Sin embargo, pese a que es utilizada mundialmente, más de 45% de las imágenes provienen de Estados Unidos y reflejan una realidad localizada en el hemisferio norte y que encarna representaciones propias de la cultura occidental. No resulta sorpresivo entonces el ejemplo citado por Zou y Schiebinger: sistemas de inteligencia artificial entrenados con ImageNet asignan las categorías «novia», «vestido», «mujer» o «boda» a la imagen de una novia occidental vestida de blanco, pero identifican como «arte de performance» o «disfraz» la imagen de una novia vestida con el típico atuendo usado en la India, que ciertamente difiere del occidental.

Otro ejemplo está dado por los traductores automáticos como Google Translate, donde se encontró que el sistema asignaba un género específico al traducir palabras que son neutras en un idioma y no en otro[10], perpetuando

8. J. Buolamwini y T. Gebru: «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification», trabajo presentado en la 1ª Conferencia sobre Equidad, Responsabilidad y Transparencia, disponible en *Proceedings of Machine Learning Research (PMLR)* vol. 81, 2018.
9. Olga Russakovsky et al.: «ImageNet Large Scale Visual Recognition Challenge» en *International Journal of Computer Vision* vol. 115 Nº 3, 2015.
10. Gabriel Stanovsky, Noah A. Smith y Luke Zettlemoyer: «Evaluating Gender Bias in Machine Translation» en *Actas de la 57ª Reunión Annual de la Asociación de Lingüística Computacional (ACL)*, 7/2019.

así estereotipos de género como la asignación del género femenino a la palabra «*nurse*» y masculino a «*doctor*», palabras que en inglés valen para ambos géneros. Es posible que en los textos utilizados para entrenar el modelo la probabilidad de encontrar la palabra «*nurse*» traducida como «enfermera» ciertamente fuera más alta, y por tanto el modelo minimiza las chances de error al asignar ese género en situaciones de incerteza, y lo mismo vale con «*doctor*». Un caso relacionado es el de los sistemas de puntuación para la asignación de préstamos bancarios o límites de gasto en tarjetas de crédito: frente a una pareja con ingresos, gastos y deudas similares, la empresa de tarjetas de crédito estableció un límite para la mujer de casi la mitad del límite del esposo[11]. La brecha salarial entre hombres y mujeres es una realidad del mundo desigual en que vivimos, y probablemente los datos con los que fue entrenado el modelo la reflejaran, por lo que su recomendación era asignarle mayor límite de gasto al hombre que a la mujer. Es decir, los datos son un reflejo (acotado) de la realidad actual. Sin embargo, en estas situaciones cabe preguntarse: ¿realmente queremos que el modelo perpetúe (y hasta en ocasiones amplifique) las desigualdades, por el solo hecho de que vivimos en una sociedad desigual? ¿O queremos modificar esta realidad? El recorte que se hace de estos datos, la población utilizada para construir las muestras, las variables que se miden: todas son decisiones humanas que están lejos de ser neutrales. El aura de neutralidad que muchas veces se atribuye a los sistemas automáticos se desvanece en el instante mismo en que comprendemos la relación entre los datos, los modelos y las personas. Y la necesidad de auditar la equidad de nuestros modelos tomando en cuenta una perspectiva interseccional se vuelve sumamente relevante.

En ocasiones, cuando detectamos posibles sesgos o rendimientos dispares en estos modelos, es posible pensar en soluciones para mitigarlos. Una de ellas sería balancear de alguna forma los datos, para evitar que los modelos resulten discriminatorios o injustos, dependiendo de la situación que estamos modelando. Otra opción podría ser inducir al sistema a que utilice representaciones «justas» de los datos, en el sentido de que no estén asociadas a las características que son fuente de discriminación. O, directamente, obligarlo a ignorar estos atributos protegidos, como el género u otras características demográficas, al momento de tomar una decisión. Sin embargo, debemos ser cuidadosos al diseñar estas soluciones: aunque ocultemos ciertos atributos a un sistema, como el género o el grupo étnico al que pertenece una persona, la correlación entre esos atributos y otras variables seguirá existiendo. Recordemos que si hay algo que los modelos de aprendizaje automático hacen bien

11. Genevieve Smith e Ishita Rustagi: «When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity» en *Stanford Social Innovation Review*, 2021.

es encontrar patrones y también correlaciones. Por eso, si bien la comunidad académica de investigación en equidad algorítmica (*fairness*) ha trabajado arduamente durante los últimos años en pos de construir modelos justos y que no discriminen, el factor humano en el diseño de estos sistemas resulta primordial. Aunque existen en la actualidad diversas formalizaciones del concepto de *fairness*, muchas de ellas resultan mutuamente incompatibles, en el sentido de que no es posible maximizarlas al mismo tiempo[12], y por tanto se debe optar por aquellas que se desee maximizar.

No alcanza entonces con generar bases de datos representativas o modelos justos en algún sentido específico. Los sistemas de inteligencia artificial están diseñados por personas con sus propias visiones del mundo, prejuicios, valoraciones de los hechos y sesgos adquiridos a lo largo de su experiencia de vida, que pueden filtrarse en el diseño y la definición de criterios de evaluación para estos modelos. Si esos grupos de trabajo no son lo suficientemente diversos como para reflejar una amplia variedad de visiones, muy probablemente no lleguen siquiera a darse cuenta de la existencia de los sesgos, y por tanto a corregirlos. No hay ejemplo más claro que el caso de Joy Buolamwini, quien descubrió el sesgo racial de los sistemas de detección facial al usarlos en su propio rostro.

Los sistemas de inteligencia artificial están diseñados por personas con sus propias visiones del mundo, prejuicios, valoraciones

Ahora bien, si la diversidad en los equipos que conciben estos sistemas resulta tan relevante, esperaríamos que en la práctica esos grupos fueran realmente diversos, no solo en términos de género, sino también de clases sociales, etnias, creencias, edad u orientación sexual, solo por dar algunos ejemplos. Pero la respuesta no siempre es la que deseamos, y en palabras del AI Now Institute de la Universidad de Nueva York, la industria de la inteligencia artificial está viviendo una crisis de diversidad «desastrosa»[13]. Según su informe, elaborado en 2019, estudios recientes encontraron que solo 18% de los trabajos publicados en las principales conferencias de inteligencia artificial son realizados por mujeres, y que más de 80% de quienes son docentes de inteligencia artificial son hombres. Esta disparidad también se refleja en la industria, donde, por ejemplo, las mujeres representan solo 15% del personal de investigación de inteligencia artificial en Facebook y

12. Sorelle Friedler, Carlos Scheidegger y Suresh Venkatasubramanian: «The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making» en *Communications of the ACM* vol. 64 Nº 4, 4/2021.

13. Sarah Myers West, Meredith Whittaker y Kate Crawford: «Discriminating Systems: Gender, Race and Power in AI», AI Now Institute, 4/2019.

10% en Google, dos de las empresas líderes en el área a escala global. Por otro lado, no se cuenta con datos públicos sobre personas trans o con otras identidades de género. Y a escala regional la situación tampoco mejora. Por ejemplo, según un informe elaborado por la Asociación Chicas en Tecnología y el Instituto para la Integración de América Latina y el Caribe del Banco Interamericano de Desarrollo (INTAL-BID)[14] sobre mujeres en el sistema universitario argentino entre 2010 y 2016, existen grandes brechas de género en el ingreso y egreso de las estudiantes de las disciplinas CTIM (ciencia, tecnología, ingeniería y matemática). Así, se observa un registro de 33% de mujeres y 67% de varones.

Ahora bien, aunque este escenario suena desolador y muchas de las situaciones que hemos discutido a lo largo de este artículo resaltan aspectos negativos potencialmente asociados al uso de estas tecnologías, muchos de los esfuerzos realizados en los últimos años para crear conciencia sobre estos riesgos y aumentar la diversidad de la comunidad de inteligencia artificial, tanto en el ámbito académico como en la industria, comienzan a sentar las bases para un futuro más promisorio. Iniciativas como la de Chicas en Tecnología o el Observatorio de Datos con Perspectiva de Género en Argentina, o WomenInML, QueerInAI, BlackInAI y LatinXInAI a escala global, solo por nombrar algunas, comienzan a poner en debate y a cuestionar esta realidad. Los gobiernos empiezan a preocuparse por la necesidad de regular el uso y desarrollo de estas tecnologías. La emergencia de foros de discusión especializados en estas temáticas y el interés de todas las ramas de la ciencia por conocer las implicancias y potenciales aplicaciones de la inteligencia artificial en sus propios campos de estudio abren nuevos horizontes para el desarrollo científico guiado por los datos. Porque no se trata de obstaculizar el avance de la inteligencia artificial como disciplina, sino de que tanto quienes la utilizan como quienes la desarrollan sean conscientes de sus limitaciones, y de que las tomemos en cuenta a la hora de concebir y hacer uso de estas tecnologías.

14. Ana Inés Basco, Cecilia Lavena y Chicas en Tecnología: «Un potencial con barreras. La participación de las mujeres en el área de Ciencia y Tecnología en Argentina», Nota Técnica Nº IDB-TN-01644, BID, 2019.