Este ejemplar fue editado por el programa Leé Ciencia. Leé Futuro, una iniciativa del Ministerio de Ciencia, Tecnología e Innovación de la Nación que se propone acercar lecturas de ciencia a niños, niñas, adolescentes y jóvenes como un modo de garantizar el acceso a la cultura científica.

Introducción

Enzo Ferrante[1]

Desde hace ya más de una década, y probablemente sin que nos diéramos cuenta, la inteligencia artificial ha comenzado a permear diversas actividades humanas. Las noticias que leemos; la comida que pedimos por el celular; el camino que tomamos con el auto; o la serie que miramos a la noche luego de cenar: todas estas acciones están mediadas por sistemas de inteligencia artificial que nos recomiendan qué leer, comer, mirar o incluso qué camino tomar para evitar demoras.

Para entender de qué hablamos cuando hablamos de inteligencia artificial, uno de los conceptos fundamentales que necesitamos entender es el de algoritmo. Un algoritmo es básicamente una secuencia de pasos ordenados que, al ser ejecutados, resuelven una tarea concreta. Las recetas de cocina, los instructivos para armar un mueble, o los pasos que seguimos para multiplicar dos números en una hoja de papel son ejemplos con los que interactuamos a diario, sin saber que estamos siguiendo un algoritmo. Pero los algoritmos que nos interesan en este texto no son esos sino aquellos que pueden ser ejecutados por una computadora. Los programas de computadora, también conocidos como software o sistemas informáticos, son en realidad algoritmos escritos en un lenguaje particular (como Python, Java o C++), que puede ser entendido tanto por seres humanos como por computadoras. De esta forma, las programadoras y los programadores son personas que pueden darle instrucciones a una computadora para que ejecute acciones tales como mostrar un mensaje por pantalla, sumar dos números o pedirnos que ingresemos un texto con el teclado. Pero entonces ¿qué tienen que ver los algoritmos con la inteligencia artificial?

1. Enzo Ferrante es Doctor en Informática por la Université Paris-Saclay (París, Francia) e Ingeniero de Sistemas por la Unicen (Tandil). Realizó su postdoctorado en el Imperial College London (Reino Unido) y, a fines de 2017, volvió a la Argentina como investigador repatriado al Instituto de Señales, Sistemas e Inteligencia Computacional, sinc(i) (Conicet–UNL)

Para intentar responder a esa pregunta, investigadores e investigadoras en Inteligencia Artificial de Argentina escribimos el libro *¿Aprendizaje automágico? Un viaje al corazón de la inteligencia artificial contemporánea*. Allí explicamos que la inteligencia artificial puede ser entendida, en un sentido amplio, como la disciplina que se encarga de comprender y construir entidades artificiales inteligentes que simulan en algún sentido el comportamiento humano. Y nos interesamos principalmente en un subcampo de la inteligencia artificial: el aprendizaje automático, que ha sido el motor de una de las revoluciones más importantes de los últimos años en el campo de la computación. Los algoritmos de aprendizaje automático nos permiten entrenar a una computadora para realizar una tarea específica (como detectar la presencia de una persona en una imagen, o predecir si mañana lloverá) a partir del análisis de grandes bases de datos. Al utilizar estos algoritmos, la que aprende es la computadora. Si te interesa entender más acerca de cómo funciona el aprendizaje automático, no dejes de leer el libro entero[2].

El texto que hoy tenés en tus manos es solamente un capítulo de ese libro, el de Laura Alonso Alemany. En él, la autora reflexiona sobre un tema sumamente relevante: las implicancias éticas de la inteligencia artificial. ¿Son objetivos los algoritmos? ¿Podría suceder que las decisiones tomadas por un sistema automático beneficien más a un sector de la población que a otro? Y si es así, ¿cómo podemos hacer para prevenirlo? Laura nos introduce al complejo mundo de la equidad algorítmica, los sesgos y la importancia de las personas en el desarrollo de estos sistemas. ¡Que lo disfrutes!

2. El libro completo lo podés descargar gratuitamente del Catálogo Digital de Vera Cartonera: https://www.fhuc.unl.edu.ar/veracartonera/portfolio/aprendizaje-automagico/

Inteligencia artificial y valores

Laura Alonso Alemany

Los sistemas automáticos ¿son objetivos?

La inteligencia artificial como una tecnología que automatiza sistematicidades, ofrece predicciones a partir de ejemplos o descubre patrones en los datos. Los resultados pueden no ser perfectos, pero podemos medir el error de estos programas con métricas estándares, bien establecidas. Todo esto nos lleva a pensar que se trata de una tecnología más objetiva que los seres humanos, seres subjetivos y prejuiciosos. Nuestra creencia es que los sistemas basados en datos son especialmente objetivos porque ni siquiera incorporan la subjetividad del programador que escribe unas reglas según sus intuiciones, sino que se construyen enteramente a partir de *datos objetivos*.

Sin embargo, los sistemas de inteligencia artificial, incluso los basados en datos, incorporan las subjetividades de los equipos que los crean y de los grupos sociales que los financian. Esas subjetividades pueden llegar a resultar perjudiciales para parte de la población, incluso de formas muy sutiles, como veremos a continuación.

Los efectos van más allá de los resultados

Sabemos que el error forma parte de los sistemas de inteligencia artificial, y lo hemos aceptado e integrado en nuestra convivencia con estas tecnologías. Cuando pensamos en estos sistemas, entendemos que pueden tener algunas limitaciones. Muchas veces tratamos de adaptar nuestro comportamiento para obtener buenos resultados, como por ejemplo cuando pronunciamos bien para que un reconocedor de voz identifique correctamente las palabras que queremos comunicar.

Detengámonos un momento en esta escena. ¿Qué implica que *pronunciemos bien*? En muchos casos, no solamente implica que vamos a tratar de ser claros, sino que también vamos a adaptar nuestra forma de hablar a lo que sabemos que la máquina reconoce. Y ¿qué reconoce la máquina? El sistema reconoce lo que aprendió a reconocer a partir de ejemplos. Pero esos ejemplos, ¿de dónde salen?

El castellano tiene muchas variantes, algunas tan distintas que la comprensión entre hablantes de diferentes variantes resulta prácticamente imposible. Los hablantes del castellano, como también los de otras lenguas con variantes muy distintas, como el alemán, italiano o inglés, muchas veces consiguen entenderse entre sí porque aprendieron, además de la variante que es su lengua materna, una variante llamada estándar que facilita la intercomprensión entre hablantes. ¿Cómo se establece cuál es la variante estándar? Por lo general, se trata de la variante de un grupo dominante, como por ejemplo el castellano de Castilla (la lengua de los conquistadores) o el latinoamericano neutro (la variante del castellano que eligen los grandes medios de comunicación internacionales). Cuando interactuamos con otras personas o instituciones, desplegamos nuestro conocimiento social y cultural sobre las variantes del castellano, y lo ponemos en juego de forma bastante consciente.

Cuando tratamos de adaptar nuestra forma de hablar para que una máquina reconozca lo que queremos decir, nuestra postura puede ser diferente de cuando interactuamos con personas. Muchas veces descartamos cuestionamientos que quizás sí plantearíamos a una persona o institución al encontrarnos en un contexto mediado por una tecnología como la inteligencia artificial, compleja y prestigiosa, pero también con limitaciones. Nos damos cuenta de que el sistema solo funciona bien si hablamos de cierta forma, pero no le atribuimos una intencionalidad, sino que asumimos que se trata de un mecanismo objetivo y simplemente tratamos de adaptarnos a sus limitaciones como algo no intencional. Efectivamente, hasta donde sabemos, las máquinas no tienen voluntad propia, pero el contexto de mediación tecnológica, tan nuevo, tan complejo, tan rodeado de grandes prestigios y grandes expectativas, dificulta

que entendamos qué voluntades pueden estar involucradas en esa tecnología, más allá de la máquina que la implementa.

Pero incluso si no llegamos a identificar las voluntades involucradas en el desarrollo de las tecnologías que encontramos en nuestras vidas, sí podemos observar y entender el efecto de estas tecnologías. Por ejemplo, ¿qué efectos puede tener que adaptemos nuestra forma de hablar para facilitar que una máquina reconozca nuestras palabras? Puede suceder, que empecemos a considerar que nuestra variante del castellano no es moderna, no está alineada con el progreso tecnológico, no nos sirve para tener éxito en el mundo actual. Puede suceder que eso nos lleve a relegar nuestra variante materna, con la consiguiente pérdida de capacidad expresiva e incluso de identidad. Puede ser, también, que si no conseguimos adaptar nuestro dialecto, la máquina no reconozca lo que queremos decir, y eso puede tener efectos todavía más profundos: podemos sentirnos inútiles, incapaces de funcionar con éxito en el mundo actual. Puede contribuir a una imagen de nosotros mismos como ineptos que termine convirtiéndose en un obstáculo para proyectarnos y funcionar de forma satisfactoria en una sociedad cada vez más mediada por tecnología.

Entonces, el comportamiento de un sistema de inteligencia artificial puede tener efectos mucho más allá de la simple interacción puntual entre la persona y la máquina. Si bien es cierto que resulta difícil predecir todos los efectos que puede tener una determinada tecnología en algo tan complejo como su uso en una sociedad, también es cierto que esa responsabilidad recae especialmente sobre los equipos que conciben, desarrollan e implementan esas tecnologías, ya que son los que las conocen mejor. Profundicemos un poco en cómo podemos empezar a abordar estas complejidades.

Si es sistemático no es error, es sesgo

Hemos dicho que no le atribuimos intencionalidad a la máquina, y todo parece indicar que, efectivamente, las máquinas no tienen intenciones.

Pero la concepción, desarrollo y despliegue de la máquina están determinados por intenciones de grupos humanos.

En varias ocasiones hemos visto cómo los responsables de algunos sistemas de inteligencia artificial piden disculpas por efectos perjudiciales imprevistos de los sistemas que desarrollan. Por ejemplo, en el documental "El dilema de las redes sociales[3]", algunos de los entrevistados, personas involucradas en el desarrollo de estas tecnologías, explican que nunca imaginaron los efectos perniciosos que resultaron teniendo las redes sociales en cuanto a adicciones, su impacto en salud mental (por ejemplo, aumentando el índice de suicidios entre adolescentes), entre muchos otros aspectos.

En las disculpas, estos efectos perjudiciales se presentan como errores no intencionales. Sin embargo, en otro documental, "Prejuicio cifrado[4]", se describe cómo los efectos de muchos sistemas que involucran inteligencia artificial son el producto de los prejuicios de sus creadores. En este documental se muestra cómo un sistema de reconocimiento de imágenes identifica con gran precisión rostros de personas con piel clara pero comete muchos más errores si se encuentra ante el rostro de una persona de piel más oscura.

Vemos aquí una gran diferencia entre un error accidental y un error sistemático. En el caso de un error sistemático, incluso si no es intencional o ni siquiera consciente, los efectos son también sistemáticos y, por lo tanto, pueden ser identificados, solucionados y, en el mejor de los casos, también prevenidos.

En los últimos tiempos hemos observado sistematicidades preocupantes en los errores de algunos sistemas de inteligencia artificial.

3. "El dilema de las redes sociales" es un documental combinado con elementos de ficción.
4. "Prejuicio cifrado" es un documental dirigido por Shalini Kantayya y estrenado en 2020.

Hemos entendido que los errores afectan de forma más perniciosa a personas de grupos minorizados, mientras que tienen un mejor funcionamiento con respecto a las personas de grupos dominantes.

Por ejemplo, un sistema de filtrado automático de candidatos para lugares de trabajo para Amazon descartaba sistemáticamente a mujeres. Twitter creaba automáticamente recortes de imágenes grandes en las que sistemáticamente se mostraban las caras de las personas en la imagen, priorizando personas blancas por encima de personas de pieles más oscuras. El servicio de traducción automática de Google traduce los nombres de profesiones que tienen género neutro en inglés al género estereotipado para esas profesiones en castellano, contribuyendo de esta forma a reforzar estereotipos de género y a dificultar el acceso a determinadas profesiones para grandes sectores de la población. Por ejemplo, se traducen doctor y nurse, palabras que pueden aplicarse a personas de cualquier género en inglés, sistemáticamente como doctor y enfermera en castellano.

Este tipo de comportamientos es especialmente grave si tenemos en cuenta que estos sistemas tienen injerencia en derechos humanos fundamentales como educación, salud o justicia. Por ejemplo, un sistema de estimación de calificaciones escolares de Reino Unido asignó notas más bajas de lo que realmente habrían obtenido a personas de barrios de renta baja, pero hizo estimaciones más acordes con el resultado final para personas de barrios con mayor renta per cápita. Varios cuerpos de policía alrededor del mundo usan sistemas de reconocimiento facial para vigilar los espacios públicos y encontrar personas con orden de búsqueda y captura, pero como hemos mencionado más arriba, estos sistemas poseen menos exactitud al clasificar personas de pieles más oscuras que en personas de pieles más claras. En la justicia del estado de Florida, en Estados Unidos, un sistema que determinaba el riesgo de reincidencia en personas que solicitan libertad condicional estimaba un riesgo mayor al real para personas tipificadas como de etnia negra, y un riesgo menor al real para personas tipificadas como de etnia blanca.

A este tipo de comportamiento sistemático se lo conoce como sesgo, porque proviene de la intervención humana en la creación del sistema.

Podemos tratar el sesgo

Por su sistematicidad, estos sesgos se pueden detectar con métricas bien establecidas, las llamadas *métricas de equidad* (*fairness* en inglés), siempre que se haya identificado el grupo social al que se está discriminando. Este grupo social se representa mediante uno o más atributos protegidos. Mediante estos atributos, las métricas de equidad describen con precisión si las predicciones de un modelo se distribuyen de forma indistinta entre la población que tiene el atributo protegido y la que no lo tiene. De esta forma se puede detectar si un sistema está actuando de forma discriminatoria con respecto a un grupo social que ya hemos identificado como vulnerable. Sin embargo, resulta mucho más complejo identificar comportamientos dañinos si no hemos identificado previamente a quiénes pueden afectar de forma sistemática. En el ejemplo con el que iniciábamos este capítulo no resulta fácil caracterizar las personas que pueden verse afectadas porque el sistema no reconoce sus palabras: puede tratarse de personas de ciertas regiones, pero también de ciertos grupos sociales, con voces más agudas o más graves, con ciertas particularidades neurológicas o motoras.

Afortunadamente, las métricas de equidad no son la única forma de inspeccionar el comportamiento de un sistema de inteligencia artificial. En los sistemas programados explícitamente se puede revisar el código para obtener una descripción de las acciones que podría llevar a cabo el sistema. Pero los sistemas basados en aprendizaje automático suelen producir modelos que resultan muy difíciles de comprender para los seres humanos. Sin embargo, se pueden aplicar mecanismos para que esos modelos ofrezcan, además de una predicción, también una explicación de las razones en las que se basa esa predicción. En esas explicaciones se pueden detectar razones inaceptables para nuestra sociedad, como por ejemplo la discriminación por etnia o género.

Dada la gravedad de estos efectos dañinos, sería muy importante poder prevenirlos en el momento de diseñar un sistema, en lugar de detectarlos recién cuando el sistema ya está funcionando en la sociedad y afectando la vida de las personas. La principal dificultad para prevenirlos está en nuestras propias limitaciones cognitivas.

El sesgo es un mecanismo cognitivo básico de los seres humanos, y resulta invisible para las personas que lo tienen. Por lo tanto, es prácticamente inevitable que un sistema diseñado por personas incorpore el sesgo de esas mismas personas. ¿Cómo hacer, entonces, para evitar los comportamientos dañinos sistemáticos? La mejor propuesta que tenemos hasta el momento no consiste en evitar los sesgos, sino en multiplicarlos. Es decir, incorporar la mayor cantidad de perspectivas distintas en la creación de un sistema, para integrar, desde el diseño, posibles contextos de uso en los que el sistema tendría efectos diferentes.

Atención: ¡inteligencia artificial en construcción!

Hemos visto cómo los sistemas de inteligencia artificial incorporan los sesgos propios de sus creadores, y por esta razón pueden llegar a tener errores sistemáticos con efectos discriminatorios.

Ante la sistematicidad de los errores, los responsables de estos sistemas muchas veces alegan que los modelos predictivos sencillamente están reproduciendo las tendencias estadísticas que encontraron en los datos con los que fueron entrenados. Es decir, que los sesgos de los sistemas no se originan en los equipos que los crearon, sino que son tendencias propias de la sociedad. Sin embargo, al inspeccionar estos comportamientos en detalle, observamos que las sistematicidades encontradas se alinean con los valores de los grupos sociales dominantes que idean y financian estas tecnologías a más alto nivel, y no necesariamente con los fenómenos que efectivamente ocurren en la sociedad.

En cualquier caso, si el comportamiento de los sistemas es pernicioso, independientemente de cuáles sean las razones por las que llegó a serlo, es necesario remediarlo. Contamos con leyes que garantizan muchos derechos fundamentales, como el derecho a la no discriminación, pero resulta difícil aplicar estos principios generales a casos concretos, y a menudo sutiles, que involucran tecnologías sofisticadas en interacciones complejas con la sociedad. Afortunadamente, en muchos países y también a nivel internacional, se están diseñando normativas específicas que determinan las responsabilidades, proveen mecanismos de control o disponibilizan canales para recibir las quejas y comentarios de los usuarios, de la misma forma que se desarrolló para otras áreas como alimentos, productos farmacéuticos, o derecho del consumidor en general. Resultan especialmente esperanzadoras las regulaciones que exigen la auditabilidad de los sistemas que impactan en derechos fundamentales de las personas, como la ley *rider* en España[5]. Estas exigencias regulatorias resultan totalmente factibles a nivel técnico. Afortunadamente, el área Inteligencia Artificial Responsable se ha desarrollado mucho en los últimos años y hoy contamos con herramientas como métricas para supervisar el comportamiento de los sistemas, metodologías para obtener explicaciones sobre las predicciones de los modelos y entornos de trabajo que facilitan estas herramientas.

Queremos cerrar este capítulo con un llamado a la acción. A pesar de las complejidades técnicas, los conceptos fundamentales en los que se basan los modelos de inteligencia artificial son intuitivos.

También podemos comprender sin mucha dificultad cómo se comportan estos sistemas, aunque desconozcamos el detalle de cómo funcionan internamente, y tenemos instrumentos para detectar los sesgos. De esta forma, podemos convertirnos en agentes de cambio, ser una parte activa para mejorar estos sistemas y ayudar a construir una inteligencia artificial mejor para todos.

5. La ley *rider* (del inglés, ciclista) establece una serie de medidas de protección a los derechos laborales de las personas que se dedican al reparto domiciliario a través de plataformas digitales en España (2021)

Sobre la autora

Laura Alonso Alemany es doctora en Lingüística Computacional por la Universitat de Barcelona, profesora e investigadora en la Universidad Nacional de Córdoba y forma parte del grupo de investigación en Procesamiento de Lenguaje Natural de la Facultad de Matemática, Astronomía, Física y Computación en dicha universidad. Además, es miembro del equipo de Ética en Inteligencia Artificial de la Fundación Vía Libre. Es especialista en inteligencia artificial aplicada al procesamiento del lenguaje.