Somos IA n.º 4: Todo sobre los sesgos

Cubierta-alt.

Un majestuoso espejo dorado se encuentra en una habitación vacía: vemos por encima del hombro de un robot, o una IA personificada, que se está mirando en el espejo. El reflejo en el espejo muestra una multitud de rostros humanos: al frente hay tres hombres caucásicos, sonriendo con confianza y radiantes en el reflejo. Detrás de ellos se encuentran hombres y mujeres asiáticos y afroamericanos con mordazas en la boca, desvaneciéndose en el fondo. El espejo emana una luz colorida, dibujada como un arte lineal abstracto, que va desde el reflejo hacia la IA o robot. El reflejo de la IA como personas en el espejo, tanto oprimidas como empoderadas, es un símbolo de la idea de que el sesgo de la máquina es simplemente un reflejo de los sesgos humanos que codifica la tecnología.

Citar como:
Julia Stoyanovich y Falaah Arif Khan. “Todo sobre el sesgo”.
We are AI Comics, Vol. 4 (2021) http://r-ai.co/comics

Página 1

Vamos a hablar de qué son y cómo surgen los llamados “sesgos” de la IA.
Imagen de una cabeza robótica, con ojos saltones, dos antenas grises en la cabeza y botones de teclado en lugar de dientes. En los ojos, en escritura arremolinada, está escrita la palabra “sesgo”.

Decimos que una IA está sesgada cuando su uso puede conducir a una discriminación injunsta y sistemática contra personas concretas o grupos de personas, de manera que favorezca a otras personas o grupos.

El sesgo puede provenir de patrones dañinos recogidos de los propios datos, o de cómo está diseñado el algoritmo, o de los objetivos que le especificamos, o de cómo lo usamos.

En su artículo seminal de 1996, Batya Friedman y Helen Nissenbaum identificaron tres tipos de sesgos que pueden surgir en los sistemas informáticos,

Una caricatura de Batya Friedman y Helen Nissenbaum sonríen a la cámara o al lector.

que representamos aquí como un dragón con tres cabezas:
preexistente
técnico
emergente

Un dragón negro azabache, con tres cabezas, estira sus garras hacia el lector y abre la boca emitiendo un gruñido: la cabeza más a la izquierda mira hacia la izquierda y tiene escrito “preexistente” en azul en el cuello. La cabeza del medio está girada hacia arriba y ruge desafiante, y tiene la palabra “técnico” escrita en color púrpura en el cuello. La cabeza más a la derecha mira hacia el lector con la boca abierta y muestra su lengua verde viscosa al gruñir, y tiene la palabra “emergente” escrita en rosa en el cuello.

Página 2

Recordemos ahora la metáfora del horneado que usamos en el volumen 1 para entender los algoritmos basados en datos.

¡Usemos ahora la misma metáfora para entender los sesgos!

Demos la bienvenida de nuevo a nuestra protagonista Mo. Mo lleva una manopla amarilla y saca una barra de pan del horno. El pan es de un marrón dorado perfecto en el exterior y tiene un trabajo de corteza detallado en forma de cabezas de robots sonrientes.

Sesgos preexistentes (en los datos)

Los sesgos preexistentes existen independientemente de los algoritmos y tienen su origen en la sociedad.

Estos serían los tipos de sabores que se filtrarán a nuestro pan si no priorizamos la pureza y la frescura de los ingredientes,

o si decidimos usar una masa madre premezclada y lista para usar.

Mo está comprando en una tienda de comestibles: sostiene una canasta llena de ingredientes y está en el pasillo de “precocinados”, seleccionando una caja de mezcla para pan.

Esos sesgos existen en la sociedad, vienen “precocinados” en el algoritmo,

y proceden del sistema de discriminación subyacente existente en el lugar donde se recopilaron los datos:

como por ejemplo, los estereotipos raciales y de género que los modelos lingüísticos recogen cuando se entrenan con datos procedentes de las redes sociales.

Una cabeza de robot frunce el ceño al espectador. Sus cejas están fruncidas por la ira y sus ojos están rojos y furiosos. En sus pupilas vemos el rostro de un hombre que grita. Detrás del robot está el mismo hombre, con las cejas cruzadas por la ira, el pelo de punta y la boca en medio de un grito. De su boca emergen pequeñas burbujas que “alimentan” la parte posterior de la cabeza del

robot. Las burbujas tienen los logotipos o símbolos de plataformas de redes sociales populares, como una cámara vectorizada o un pájaro cantando.

Página 3

Sesgos técnicos (en el sistema técnico)
Los sesgos técnicos los introduce el propio sistema, debido a la forma en que está diseñado o a cómo funciona.

Equivalen a las imperfecciones que se generan en el pan por haberlo cocinado con un equipamiento incorrecto:

como sucede con la cocción desigual de los cupcakes si la temperatura del horno está mal calibrada,
o cuando rebosa la masa porque el horno no es del tamaño adecuado.

Dos instantáneas de los fiascos de horneado de Mo: a la izquierda está la imagen de un molde para hornear desbordado, donde Mo claramente llenó demasiado el recipiente con la masa, lo que provocó que rebosara durante el horneado. A la derecha hay una instantánea de algunos cupcakes curiosos: todos tienen un desgarro en la parte superior y se han levantado asimétricamente: están aplastados y quemados por un lado, pero esponjosos y levantados por el otro. La masa del medio está cruda.
Y volviendo a los sistemas informáticos:

un ejemplo destacado son las plataformas de redes sociales–diseñadas para optimizar la actividad continuada de los usuarios (en lugar de su seguridad o la autenticidad)– que acaban promocionando las noticias falsas y artículos extremistas.

Un hombre está sentado, pensativo, con los codos apoyados en una mesa y las manos sosteniéndole el rostro. Sobre su cabeza hay una nube de arte abstracto ondulado que emana de la mano izquierda de un genio de la IA. El genio hace aparecer una red de cabezas, arrojando las mismas nubes onduladas que captan la atención del hombre: cada una de las cabezas es idéntica y grita con

virulencia a la otra con los ojos cerrados, las cejas fruncidas y gruñidos de enfado.

Página 4

Sesgos emergentes (ocasionados por decisiones)

Los sesgos emergentes surgen con el paso del tiempo, porque las decisiones que se toman con la ayuda de los sistemas informáticos llevan a cambiar cosas del mundo real,

lo que a su vez afecta el funcionamiento de esos mismos sistemas en el futuro.

Pensemos en los cambios de comportamiento que surgirán como resultado de nuestro horneado:

¿Qué pasaría si nos convertimos en maestros reposteros e incorporamos, sin darnos cuenta, el pan como un alimento que permanece constante en nuestra dieta?

Mo sostiene un plato de bollos recién horneados en una mano y se llena la boca con la otra.

¿O si hiciéramos pan con tanta frecuencia que nadie a nuestro alrededor quisiera probar una rebanada más?

Una joven frunce el ceño ante una bandeja de pan de masa fermentada y lo empuja con la mano.

O pensemos si nuestra opinión sobre “el sabor que debe tener el pan” está condicionada por la popularidad de otros productos parecidos y con sabores característicos como el “Wonder Bread”.

Un niño, con mejillas sonrosadas y cabello oscuro y rizado, sonríe al lector, mientras se acerca a la boca una rebanada de pan blanco. Frente a él, un pan perfectamente posicionado para mostrar la etiqueta “Wonder Bread” al lector.

Del mismo modo, pensemos en cómo nuestra exposición a las noticias, y a la información en general,

está conformada por los algoritmos que seleccionan noticias de las redes sociales a partir de las publicaciones que son populares y “de tendencia”.

Mo está mirando un cielo lleno de datos, hashtags, pronósticos, predicciones y análisis.

Página 5

Para concretar un poco más, veamos ejemplos de los sesgos algorítmicos en el mundo real.
Tomemos el sector de la “contratación” como un caso representativo, donde los algoritmos se utilizan cada vez más para tomar decisiones críticas de manera más “eficiente”.

Una herramienta de contratación algorítmica o de IA está realizando una entrevista: tres candidatos están sentados frente a un robot en una mesa, nerviosamente educados y ansiosos por impresionar.
Uno de los primeros indicios de que existe un motivo de preocupación se produjo en 2015, tras publicarse los resultados del estudio AdFisher de la Universidad Carnegie Mellon [2].

Los investigadores realizaron un experimento, en el que crearon dos conjuntos de perfiles artificiales de usuarios de la Web que eran iguales en todos los aspectos—en términos de datos demográficos, intereses y patrones de navegación— con una sola excepción: su género declarado, masculino o femenino.

Imagen de pantalla dividida de un posible candidato: en el panel de la izquierda hay un hombre de negocios con un traje elegante, sosteniendo un maletín, en el panel de la derecha hay una mujer de negocios también con un traje, sosteniendo un maletín idéntico. Ambas personas en los paneles son idénticas en su color de cabello, tono de piel, el tono de sus trajes, el color y el tamaño/forma de sus maletines.

Los investigadores demostraron que Google mostraba anuncios de un servicio de orientación profesional para puestos ejecutivos bien remunerados con mucha más frecuencia al grupo de hombres que al grupo de mujeres.

Esto nos retrotrae a la época en que era legal anunciar en los periódicos puestos de trabajo diferenciados para cada género. Esa práctica fue prohibida en los EE. UU. en 1964, pero persiste en el entorno de la publicidad en Internet.

Más tarde se demostró que parte de la razón por la que eso estaba sucediendo es la dinámica del propio sistema de etiquetado de los anuncios, un mecanismo esencial para lanzar las ofertas de empleo.
Esto es un sesgo técnico en la práctica.

Página 6

Vayamos adelante en el tiempo y avancemos también a la siguiente etapa del proceso de contratación: la revisión del currículum.

Un robot o IA personificada mira dentro de una bola de cristal, tratando de leer el futuro del grupo de solicitantes: sus cabezas giran alrededor de la bola de cristal, y la IA la mira detenidamente, con las manos estiradas a su alrededor.

A fines de 2018, se conoció que la herramienta de contratación de Amazon basada en IA, que se había desarrollado con el objetivo de aumentar la diversidad de los trabajadores, hizo justo lo contrario:
el sistema aprendió por sí mismo que los candidatos masculinos eran preferibles a las candidatas femeninas.
Penalizaba los currículums que incluían la palabra “femenino”, como en “capitana del club de ajedrez femenino”.
Y rebajó la calificación de las graduadas de dos universidades solo de mujeres.

Los resultados se alinearon y reforzaron un marcado desequilibrio de género en los trabajadores de la empresa.
Este es un sesgo emergente en la práctica:
un gerente de contratación a quien una herramienta de IA le sugiere repetidamente el mismo tipo de solicitante de empleo como el candidato perfecto con el tiempo llegará a creer que esa es la imagen de un empleado adecuado.

Dos personas caucásicas posan ante el lector: la de la izquierda está haciendo el signo de “pulgar hacia abajo”, la de la derecha está haciendo el signo de “pulgar hacia arriba”. La de la izquierda tiene el símbolo de mujer, un círculo con una pequeña cruz debajo, en el pulgar, mientras que la de la derecha tiene el símbolo de hombre, un círculo con una flecha que apunta al noreste, en el pulgar. Detrás de ellas hay reflejos de espejo de las dos personas, pero en lugar de sus siluetas, son robots o IA personificadas quienes están haciendo las mismas poses de pulgares hacia arriba y hacia abajo. Hay 3 conjuntos de dichos reflejos, y en la parte superior hay un reflejo de las dos personas haciendo las mismas poses, lo que simboliza que el sesgo preexistente está codificado y exacerbado por la tecnología, y luego se presenta como intuición o juicio humano “neutral” a la persona encargada de tomar decisiones.

También se puede apreciar un sesgo preexistente en este mismo ejemplo: la herramienta de IA se entrenó con datos históricos sobre empleados anteriores, que eran predominantemente hombres.

Página 7

Aquí hay otro ejemplo, que sucede en una fase aún más adelantada del proceso de contratación, durante la verificación de antecedentes que se hace tras la entrevista con un potencial empleador:

Latanya Sweeney, profesora de informática de la Universidad de Harvard,

Una Latanya Sweeney caricaturizada sonríe al lector y señala hacia arriba a los resultados de su estudio.

demostró que buscar en Google nombres que suenan a personas afroamericanas es más probable que genere anuncios que sugieren antecedentes penales, que buscar en Google nombres que suenan a personas blancas,
incluso verificando en la búsqueda si una persona tiene o no antecedentes penales.

Hay dos conjuntos de rostros de mujeres: en la parte superior izquierda hay un conjunto de rostros de mujeres caucásicas, con cabello rubio y piel clara. Debajo hay una lupa, símbolo de “búsqueda”, con el nombre “Kristen”. Abajo a la derecha hay un conjunto de rostros de mujeres afroamericanas, con cabello oscuro y rizado y tez oscura. Debajo está el símbolo de “búsqueda”, con el nombre “Latanya”.

Así es como funciona un sesgo preexistente en la práctica:
Manifestando los prejuicios raciales que están ampliamente arraigados en la sociedad.

Página 8

Los casos presentados aquí tienen un denominador común: muestran que la IA puede reforzar y amplificar la discriminación ilegal contra minorías y grupos históricamente desfavorecidos.

Generalmente, esto se denomina “sesgos de la IA”.
La misma cabeza robótica, con ojos saltones, dos antenas grises en la cabeza y botones de teclado en lugar de dientes, ha vuelto a aparecer en nuestro debate sobre sesgo ahora. En sus ojos, en escritura arremolinada, se lee la palabra “sesgo”.

Entonces, ¿por qué fallan estos sistemas sofisticados que pretenden que la contratación sea más eficiente, haciendo que las cosas empeoren todavía más?

Por supuesto, los problemas de sesgo en el empleo no son nuevos. También estaban presentes en la era analógica.
Por ejemplo, en su conocido estudio de 2004, Marianne Bertrand y Sendhil Mullainathan enviaron currículums ficticios a ofertas de empleo en periódicos de Boston y Chicago. [5]

Para manipular la percepción de la raza, asignaron al azar nombres que sonaban afroamericanos o blancos a los currículums.

Los nombres blancos recibieron un 50 por ciento más de llamadas para entrevistas.

Los caricaturizados Sendhil Mullainathan y Marianne Bertrand miran al lector, mientras señalan hacia un cuadro grande encima de ellos. En el cuadro están las palabras “¿Emily y Greg son más empleables que Lakisha y Jamal?”. A la izquierda del texto están los rostros de un hombre y una mujer caucásicos. A la derecha están los rostros de un hombre y una mujer afroamericanos.

Este caso muestra que el sesgo puede deberse a decisiones humanas.

Página 9

Revisemos el sesgo preexistente, que a menudo se muestra en los datos.
Los datos son una imagen del mundo, su reflejo en el espejo.

Cuando pensamos en el sesgo de los datos, estamos cuestionando ese reflejo. Una interpretación del “sesgo de los datos” es que el reflejo está distorsionado: podemos sobrerrepresentar o subrepresentar sistemáticamente las partes particulares del mundo que muestran los datos, o también distorsionar las lecturas de esos datos.

Una mujer se sienta frente a su computadora y podemos ver por encima del hombro: Frente a ella, en su computadora portátil, hay una hoja de cálculo abierta y de la computadora portátil surgen fragmentos de datos. Los datos se fusionan en una forma redonda que se asienta sobre un espejo rosa. La forma redonda es un reflejo pixelado y encima vemos la imagen original del mundo: un globo circular dibujado con líneas abstractas.

Recordemos el fracaso de la IA de contratación de Amazon para mejorar la diversidad de la plantilla de empleados.

Esa herramienta fue entrenada utilizando datos históricos: currículums de personas que fueron contratadas en el pasado.

Ese entrenamiento estaba sujeto a un sesgo preexistente.

En esos datos, había una subrepresentación de mujeres en la plantilla existente y en los roles técnicos.
Un punto más sutil tiene que ver con las distorsiones.

Cuando consideramos características, como la puntuación de un individuo en una prueba estandarizada, ¿la tomamos al pie de la letra?

¿O tenemos en cuenta las diferencias en el acceso a las oportunidades educativas, como haber ido a una mejor escuela, o haber tenido acceso a tutorías personales?

Página 10

Otra interpretación del “sesgo de los datos” es que incluso si pudiéramos reflejar el mundo perfectamente en los datos, seguiría siendo un reflejo del mundo tal como es,
y no necesariamente de cómo podría o debería ser.

Es importante tener en cuenta que un reflejo por sí mismo no puede saber si está distorsionado.

Los datos por sí solos no pueden decirnos si son un reflejo distorsionado de un mundo perfecto, un reflejo perfecto de un mundo distorsionado,
o una combinación de las dos cosas.

Una mujer mira fijamente lo que parece ser un retrato o una instalación de arte de algún tipo. El retrato es una semiesfera tridimensional de píxeles verdes y azules, montada sobre una cuadrícula rosa. Detrás de ella, y más allá de la vista de la mujer, hay un globo dibujado con líneas abstractas. Las mitades del globo: la mitad hacia la mujer está dibujada en azul y verde, mientras que la otra mitad está dibujada con negro sobre líneas verdes y azules. En el medio del globo hay un enorme signo de interrogación en rojo.

El segundo punto es que no depende de los datos ni de los algoritmos, sino de las personas
—individuos, grupos y sociedad en general— llegar a un consenso sobre si el mundo es como debe ser o si necesita mejorar.

Y, si es así, cómo deberíamos mejorarlo.

Vemos la silueta de un grupo de personas mirando lo que parece ser una instalación de arte de algún tipo: hay 4 representaciones diferentes de un globo. De izquierda a derecha: hay una representación tridimensional realista de los diferentes terrenos y elementos geográficos de la tierra, luego un dibujo abstracto que usa solo polígonos concéntricos en violetas y rosas brillantes, luego un boceto compuesto por líneas abstractas en azul y verde, y finalmente una composición de arte poligonal aún más abstracta de diferentes formas coloridas.

Página 11

La conclusión aquí es que, posiblemente, cambiar el reflejo del mundo no equivale a cambiar el mundo.

Si el reflejo se usa para tomar decisiones importantes:

por ejemplo, a quién contratar o qué salario ofrecer a una persona contratada, entonces vale la pena compensar las distorsiones.

Pero la metáfora del espejo solo nos sirve solo hasta cierto punto.

Tenemos que trabajar mucho más duro, yendo mucho más allá de las soluciones tecnológicas, para lograr un cambio duradero en el mundo, no es suficiente con revisar el reflejo, simplemente.

Imagina la semiesfera tridimensional de píxeles verdes y azules, montada en una cuadrícula rosa, una vez más. Detrás de ella, en el extremo derecho, hay un globo terráqueo dibujado con líneas abstractas. La mitad del globo: la mitad hacia la semiesfera está dibujada con líneas azules y verdes rellenas de negro, mientras que la otra mitad está dibujada rellena de blanco. Una mujer sostiene un líquido de limpieza en una mano y un paño de cocina en la otra, y está limpiando la semiesfera (reflejo) del globo abstracto (mundo). Un hombre está de pie junto a ella, con un portapapeles y un bolígrafo, mirando la semiesfera y tomando notas.

Y volviendo ahora al dragón de tres cabezas de los sesgos:
Cuando hablamos de abordar los sesgos de la IA, tendemos a enmarcar el problema en cómo encontramos la manera de matar al dragón.

El dragón de tres cabezas del sesgo ha vuelto a aparecer. Esta vez, lanza fuego a través de dos de sus tres cabezas. Un valiente caballero está luchando contra el dragón y está perforando la cabeza del medio con una espada. Sin embargo, las otras dos cabezas están escupiendo llamas sobre el caballero.

Pero en nuestro debate sobre el vínculo entre los sesgos humanos y los sesgos de las máquinas, nos hemos llegado a cuestionar la naturaleza misma de esa relación.

Así que, en última instancia, tal vez la pregunta no sea:
¿cómo matar al dragón y rescatar a la princesa?
Una princesa mira hacia abajo desde la ventana de su castillo, a un caballero de pie sobre la cabeza del dragón que acaban de matar.

La primera pregunta que realmente deberíamos hacernos es:
¿Qué hacemos con una sociedad que encierra princesas en castillos?