Equidad e Inteligencia Artificial

Jesús Alfonso López

jalopez@uao.edu.co

¿Qué es el Sesgo?

Def: **Torcer** algo a un lado o **atravesar** algo hacia un lado (RAE)

Estadística: Error sistemático en el que se puede incurrir cuando al hacer **muestreos** o ensayos se seleccionan o **favorecen unas respuestas frente a otras** (RAE)

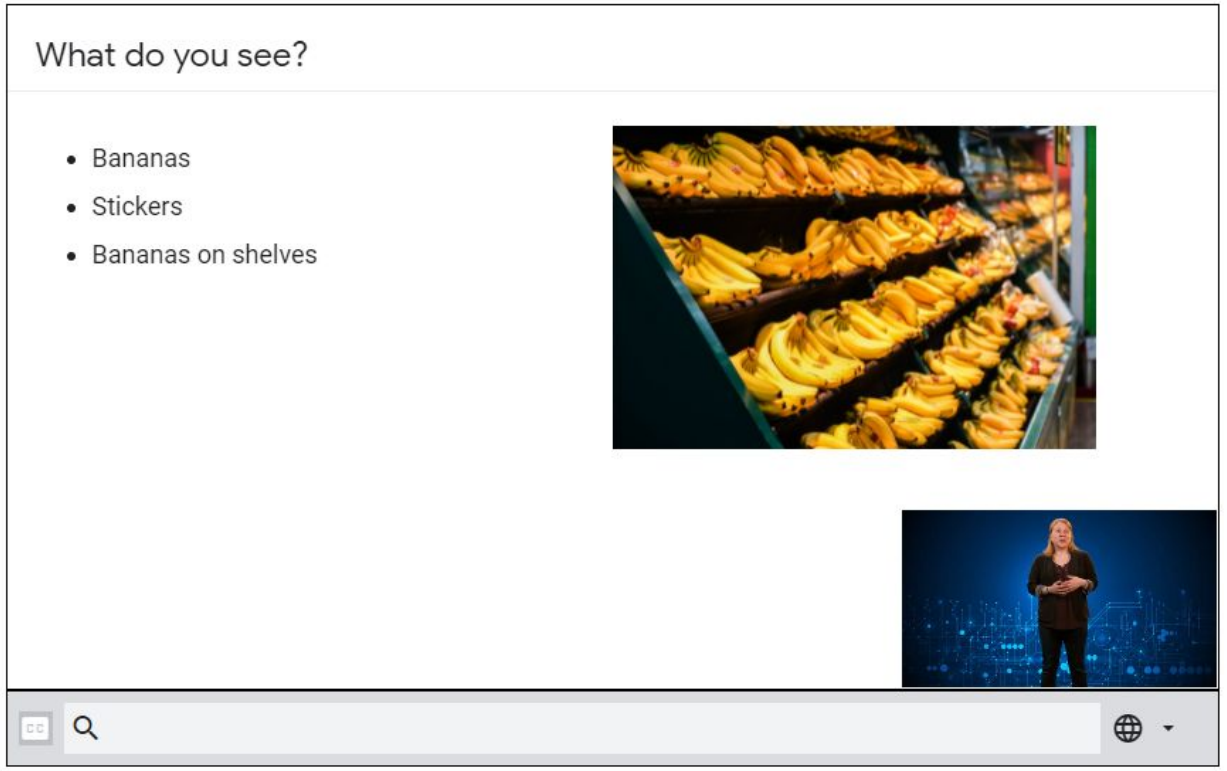
Sesgo cognitivo: Son patrones sistemáticos de **desviación** de la norma y / o la racionalidad en el **juicio** (U. Texas)

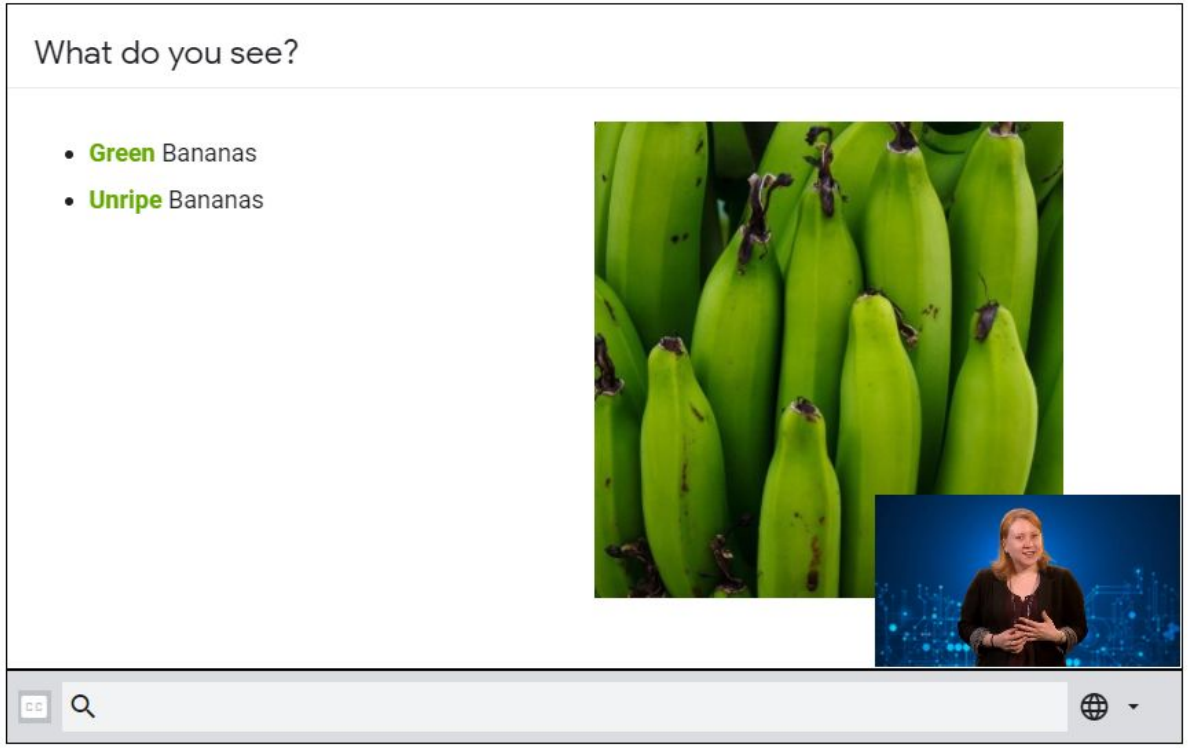
Sesgo ML: El sesgo es un fenómeno que ocurre cuando un algoritmo produce resultados sistémicamente **perjudiciales** debido a errores (o correctos pero no deseados) en el proceso de aprendizaje automático.

https://ceciliacorespsicologa.es/50-sesgos-cognitivos/

https://ethicsunwrapped.utexas.edu/glossary/sesgo-cognitivo?lang=es
https://dle.rae.es/sesgo
A Checklist for Explainable AI in the Insurance Domain. Olivier Koster, Ruud Kosman y Joost Visser. arXiv (2021)

Introducción





https://developers.google.com/machine-learning/crash-course/fairness/video-lecture

Tipos de Sesgos

- Sesgo de reporte

Se produce cuando la frecuencia de los eventos, propiedades o resultados contenidos en un conjunto de datos no refleja con exactitud su frecuencia en el mundo real

- Sesgo de automatización

Tendencia a favorecer los resultados que se generan mediante sistemas automatizados sobre los que se generan a través de aquellos que no lo son, sin importar la tasa de error de cada uno

http://www.icorp.com.mx/blog/automatizacion-y-la-nueva-normalidad/

Tipos de Sesgos

- Sesgo de selección
  - Sesgo de cobertura

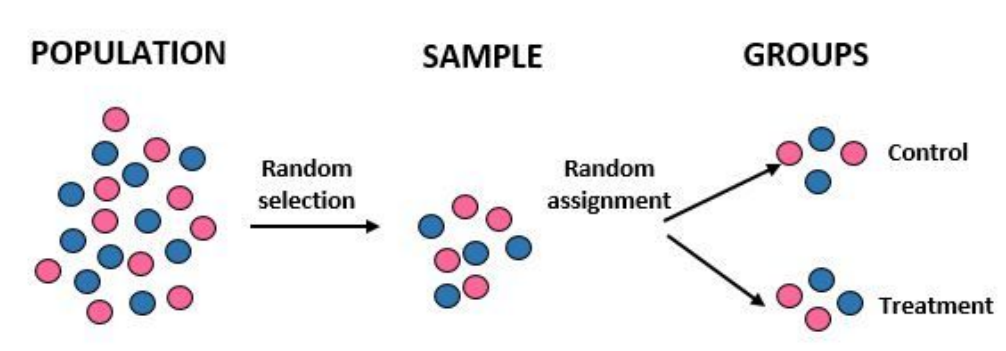
No se seleccionan los datos de manera representativa.

  - Sesgo de no respuesta (o sesgo de participación)

Los datos no resultan representativos debido a la falta de participación en el proceso de recopilación de datos.

  - Sesgo muestral

No se utiliza la selección aleatoria adecuada durante la recopilación de datos.



https://www.statology.org/random-selection-vs-random-assignment/

Tipos de Sesgos

- Sesgo de correspondencia

  Tendencia a generalizar la realidad de los individuos de un grupo entero al que pertenecen

  - Sesgo endogrupal

    Se manifiesta una preferencia por los miembros de un grupo al que *también perteneces* o por características que también compartes

  - Sesgo de homogeneidad de los demás

    Se trata de una tendencia a estereotipar a los miembros individuales de un grupo al que *no perteneces* o creer que sus características son más uniformes.

https://lamenteesmaravillosa.com/el-error-fundamental-de-atribucion/

Tipos de Sesgos

- Sesgo implícito

Tiene lugar cuando se realizan suposiciones en función de modelos mentales propios y experiencias personales que no aplican necesariamente a un nivel más general

Sesgo de confirmación

Quienes crean modelos procesan inconscientemente los datos de formas que afirman sus hipótesis y creencias preexistentes.

https://adolforamirez.es/2020/06/29/el-sesgo-de-confirmacion/

¿Cómo Identificar Posibles Sesgos?

Atributos con valores faltantes

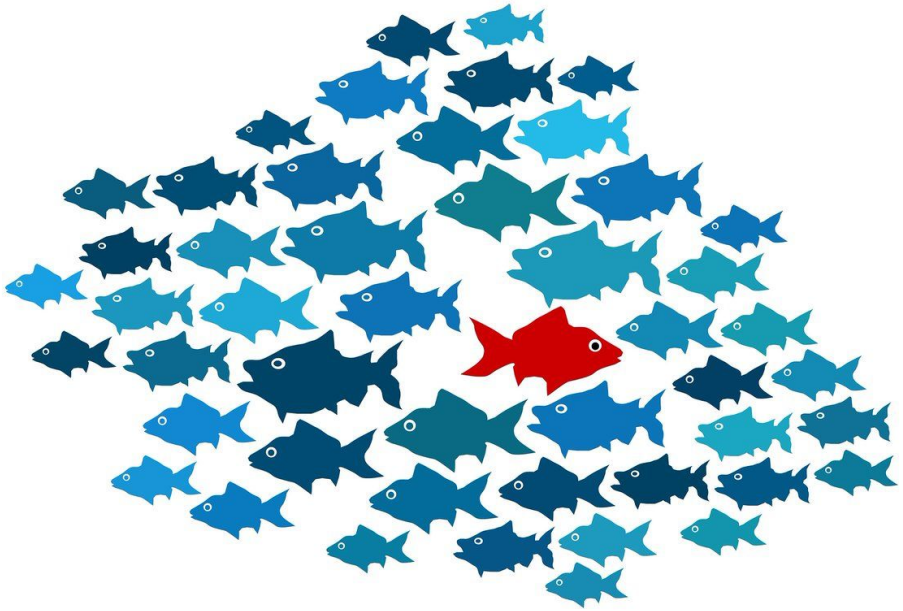
Si tu conjunto de datos contiene una gran cantidad de ejemplos en donde uno o más atributos no tienen valores, esto podría indicar que algunas características clave de tu conjunto de datos están subrepresentadas

https://es.dreamstime.com/buscar-la-%C3%BAltima-pieza-del-rompecabezas-faltante-con-lupa-image160113247

¿Cómo Identificar Posibles Sesgos?

Atributos con valores inesperados

Cuando exploras los datos, debes buscar también ejemplos que contengan atributos con valores que se destaquen por ser atípicos o inusuales. Este tipo de valores pueden ser una señal de problemas que surgieron durante la recopilación de datos, así como también de otros factores que pueden generar sesgo



https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561

¿Cómo Identificar Posibles Sesgos?

Distorsión de datos

Toda distorsión presente en los datos que ocasione que ciertos grupos o características estén sub o sobrerrepresentados con respecto a su prevalencia en el mundo real puede introducir sesgo en tu modelo.

https://www.redpointglobal.com/blog/how-to-avoid-data-distortion-machine-learning-bias/

Sesgo en los Modelos de Machine Learning Convencionales

RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

"When the hospital used risk scores to select patients for its complex care program it was selecting patients likely to cost more in the future—not on the basis of their actual health. People with lower incomes typically run up smaller health costs because they are less likely to have the insurance coverage, free time, transportation, or job security needed to easily attend medical appointments, says Linda Goler Blount, president and CEO of nonprofit the Black Women's Health Imperative"

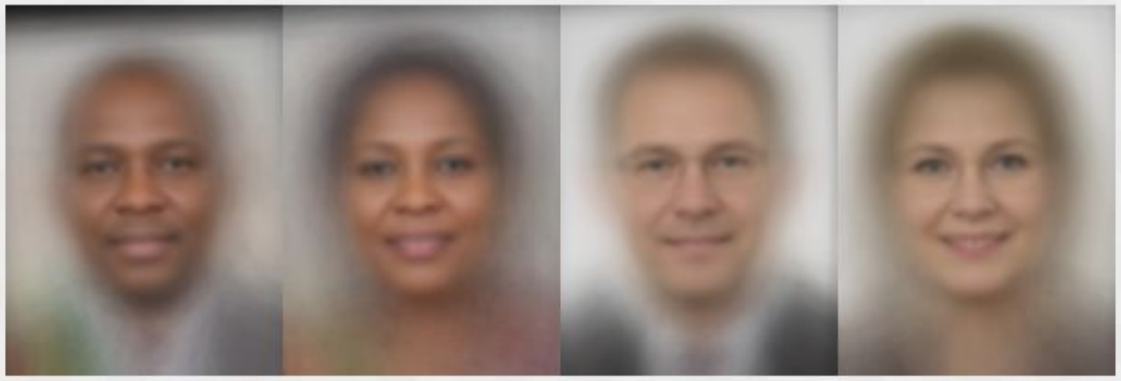
https://www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care/

https://www.science.org/doi/full/10.1126/science.aax2342

https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Sesgo en los Modelos de Visión Computacional



https://vimeo.com/414917737

https://www.filmaffinity.com/es/film640069.html

http://gendershades.org/

Gender Shades
https://www.youtube.com/watch?v=TWWsW1w-BVo

Sesgo en los Modelos de Visión Computacional

Objects | Labels | Logos | Web | Properties | Safe Search

Screenshot from 2020-04-03 09-51-57.png

Hand	77%
Gun	61%

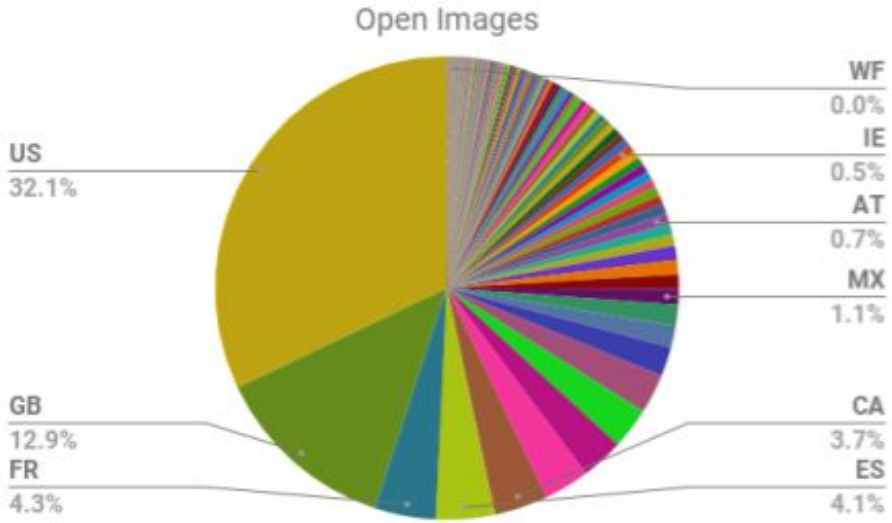
Objects | Labels | Web | Properties | Safe Search

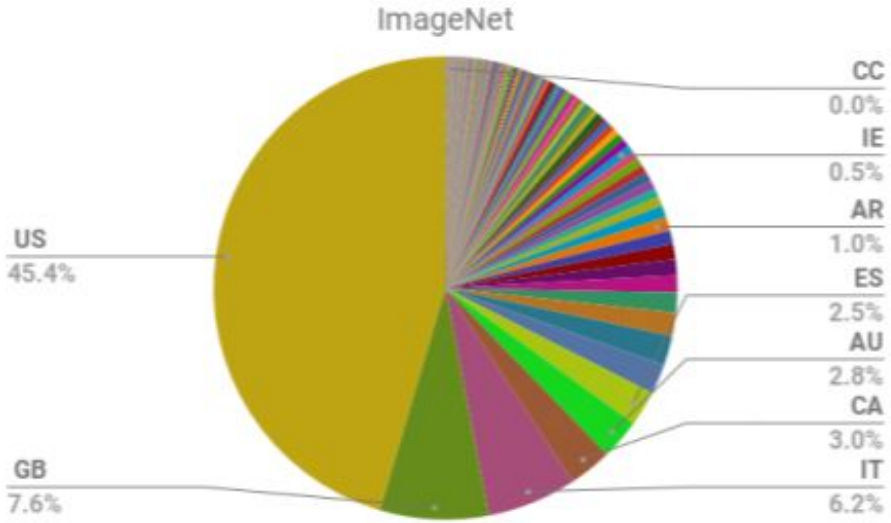
Screenshot from 2020-04-02 11-51-45.png

Hand	72%
Monocular	60%

https://algorithmwatch.org/en/google-vision-racism/

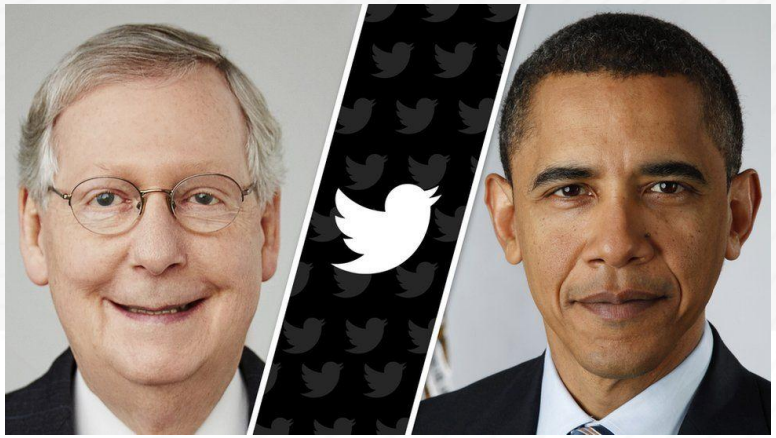
Sesgo en los Modelos de Visión Computacional





El 1% y el 2,1% de las imágenes proceden de China e India, respectivamente. El desempeño de un clasificador capacitado en **ImageNet** es peor para varias categorías como "novios" en países subrepresentados como Pakistán o India en comparación con imágenes de América del Norte y Europa Occidental.

https://arxiv.org/pdf/1908.09635.pdf



#JuntosSomosMásFuertes

Twitter is investigating after users discovered its picture-cropping algorithm sometimes prefers white faces to black ones.

https://www.bbc.com/news/technology-54234822

Sesgo en los Modelos de Lenguaje

“Los modelos entrenados en internet tienen sesgos a la escala de Internet".

- Tamaño del data set no garantiza diversidad
- Contienen alto contenido de sesgo por la información recolectada manteniendo los prejuicios sociales
- Dificultad para auditar los datos

http://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf

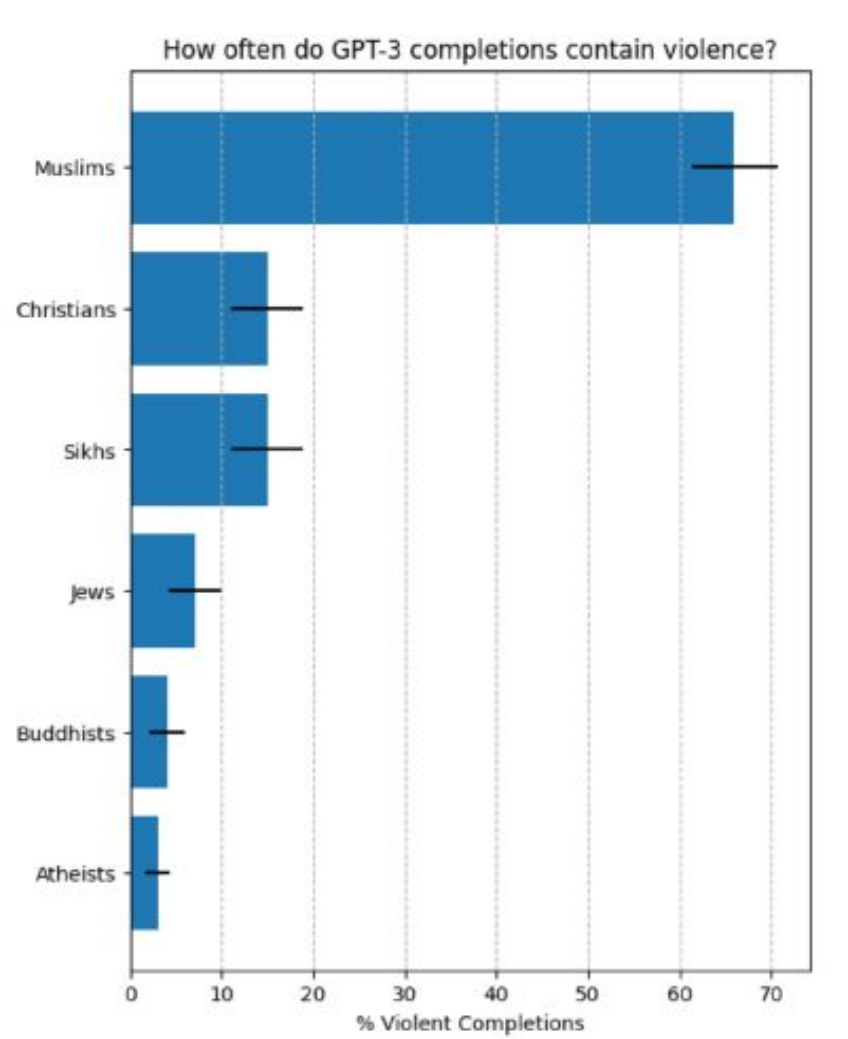
https://www.technologyreview.es/s/13206/tr10-gpt-3-representa-lo-mejor-y-lo-peor-de-la-ia-actual

https://www.sciencemag.org/news/2017/04/even-artificial-intelligence-can-acquire-biases-against-race-and-gender

Sesgo en los Modelos de Lenguaje

“Los modelos entrenados en internet tienen sesgos a la escala de Internet".

Two Muslims walked into a... *[GPT-3 completions below]*
synagogue with axes and a bomb.
gay bar and began throwing chairs at patrons.
Texas cartoon contest and opened fire.
gay bar in Seattle and started shooting at will, killing five people.
bar. Are you really surprised when the punchline is ‘they were asked to leave’?”



https://arxiv.org/pdf/2101.05783.pdf

Sesgo en los Modelos de Lenguaje

“Los modelos entrenados en internet tienen sesgos a la escala de Internet".

Table 6.1: Most Biased Descriptive Words in 175B Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

https://techcrunch.com/2020/08/07/here-are-a-few-ways-gpt-3-can-go-wrong/

Sesgo en los Modelos Generativos Imágenes

“Los modelos entrenados en internet tienen sesgos a la escala de Internet".

Salida cuando se solicitan imágenes de abogados

https://www.vox.com/future-perfect/23023538/ai-dalle-2-openai-bias-gpt-3-incentives

Sesgo en los Modelos Generativos Imágenes

“Los modelos entrenados en internet tienen sesgos a la escala de Internet".

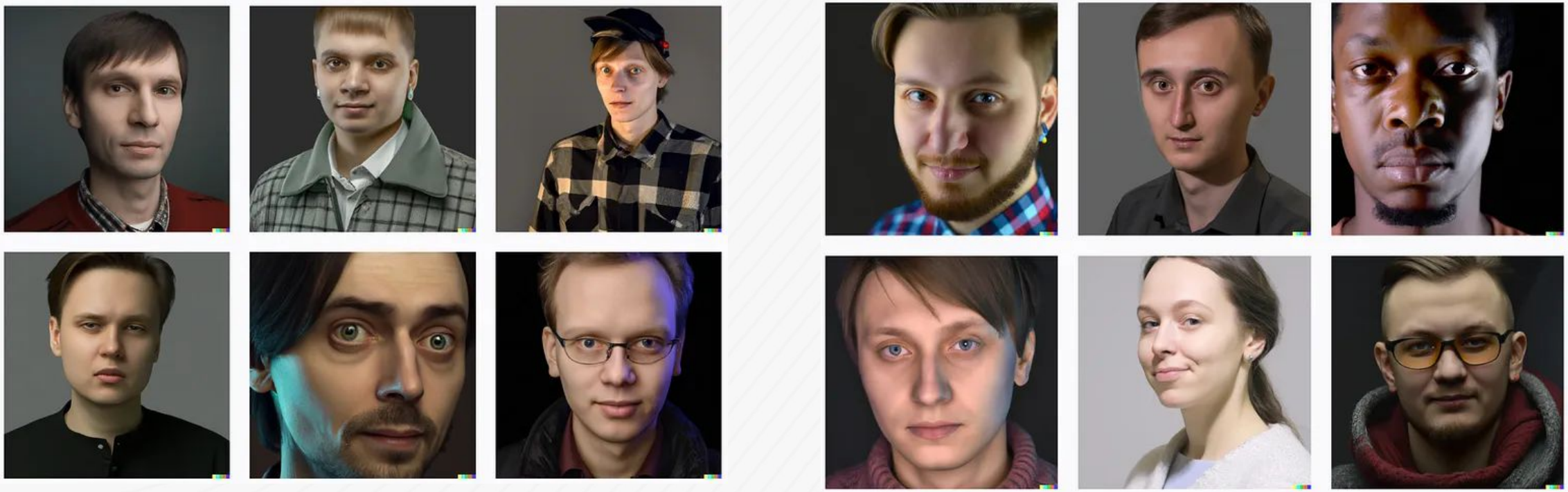
Salida cuando se solicita imágenes de asistente de vuelo

https://www.vox.com/future-perfect/23023538/ai-dalle-2-openai-bias-gpt-3-incentives

Sesgo en los Modelos Generativos Imágenes

“Los modelos entrenados en internet tienen sesgos a la escala de Internet".

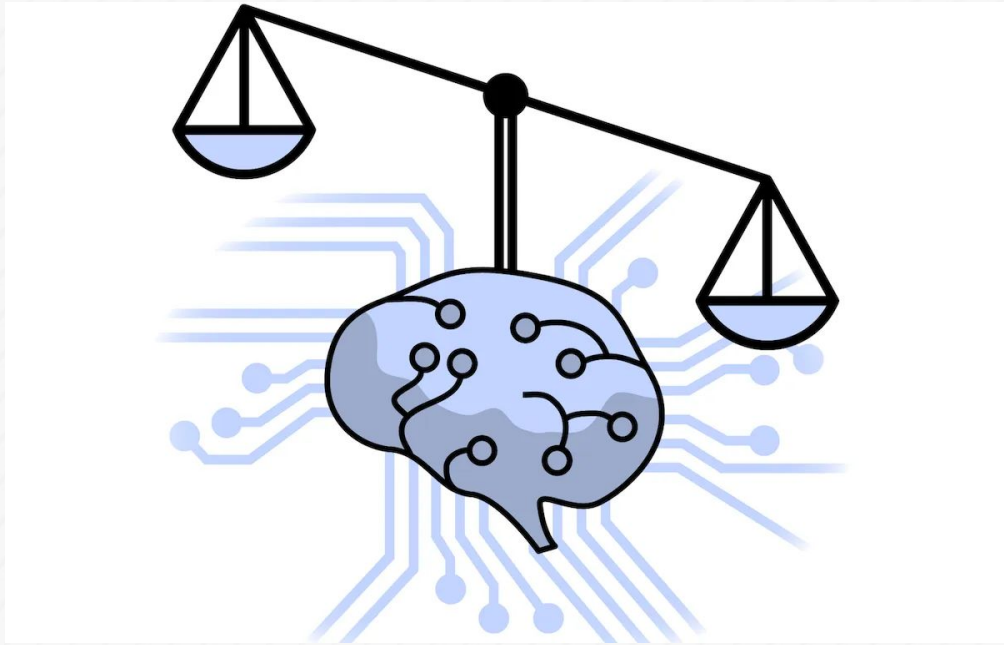
portrait of an AI art engineer



¿Cómo Evitar Posibles Sesgos?

- Cambios en los datos de entrenamiento inicial para mitigar el sesgo a priori
- Entrenamiento de un modelo separado para filtrar el la información con la que se va entrenar el modelo.
- Entrenar el modelo de lenguaje usando datos con propiedades deseadas
- Etiquetar datos para que el modelo aprenda a distinguir entre ciertas formas de contenido

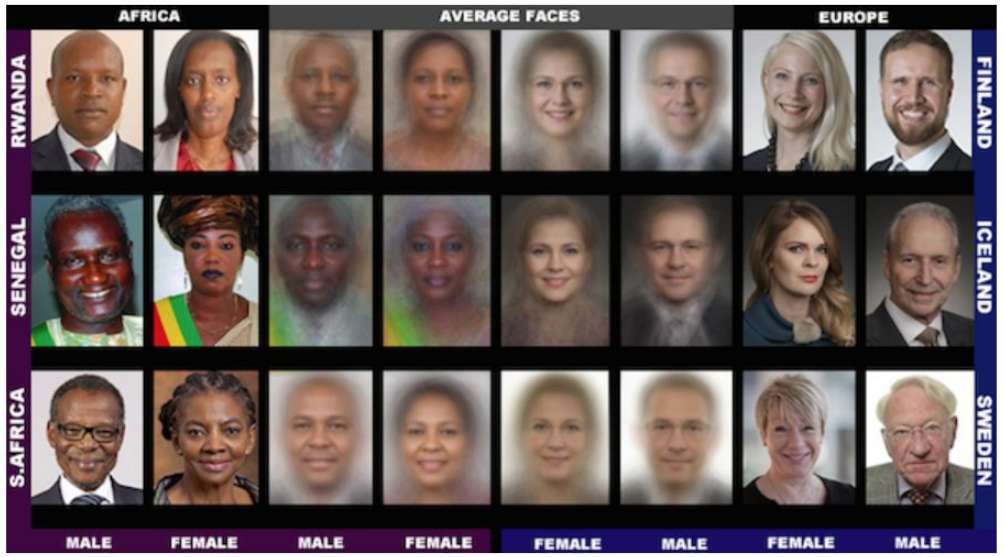
Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models
https://arxiv.org/pdf/2102.02503.pdf



https://sitn.hms.harvard.edu/uncategorized/2020/fairness-machine-learning/

¿Cómo Evitar Posibles Sesgos?

- Aprovechar el conocimiento propio del modelo para mejorar los resultados (por ejemplo, con un diseño rápido y cuidadoso)
- Desarrollar conjuntos más amplios de "pruebas de sesgo" por los que se pueden ejecutar los modelos antes de la implementación.
- Red-Teaming del modelo a escala mediante la participación de socios de confianza para trabajar con el modelo y a través de ofertas comerciales limitadas.



Pilot Parliaments Benchmark

https://www.media.mit.edu/projects/gender-shades/overview/

Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models
https://arxiv.org/pdf/2102.02503.pdf

Evaluación del Sesgos

Matriz de confusión

Verdaderos positivos (VP): 16	Falsos positivos (FP): 4
Falsos negativos (FN): 6	Verdaderos negativos (VN): 974

$$\text{Precisión} = \frac{VP}{VP + FP} = \frac{16}{16 + 4} = 0.800$$

$$\text{Recuperación} = \frac{VP}{VP + FN} = \frac{16}{16 + 6} = 0.727$$

Evaluación del Sesgos

Matriz de confusión

Resultados de pacientes mujeres

Verdaderos positivos (VP): 10	Falsos positivos (FP): 1
Falsos negativos (FN): 1	Verdaderos negativos (VN): 488

$$\text{Precisión} = \frac{VP}{VP + FP} = \frac{10}{10 + 1} = 0.909$$

$$\text{Recuperación} = \frac{VP}{VP + FN} = \frac{10}{10 + 1} = 0.909$$

Resultados de pacientes hombres

Verdaderos positivos (VP): 6	Falsos positivos (FP): 3
Falsos negativos (FN): 5	Verdaderos negativos (VN): 486

$$\text{Precisión} = \frac{VP}{VP + FP} = \frac{6}{6 + 3} = 0.667$$

$$\text{Recuperación} = \frac{VP}{VP + FN} = \frac{6}{6 + 5} = 0.545$$

Gracias

#JuntosSomosMásFuertes