

Sesgo: Problemas

Calibración de cámaras con personas blancas



Sesgo contra colores oscuros



CelebA dataset : 80% blancos, 80% hombres

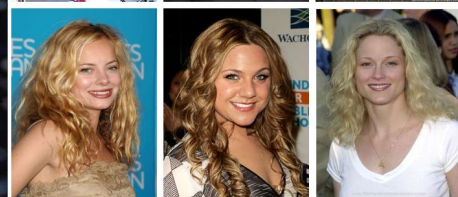
Eyeglasses



Wearing Hat



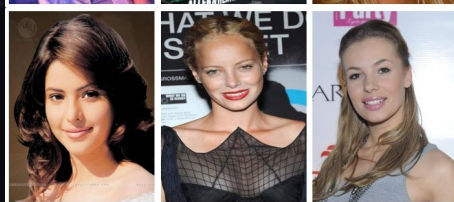
Wavy Hair



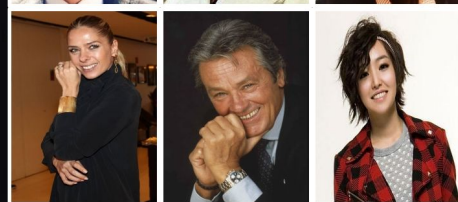
Mustache



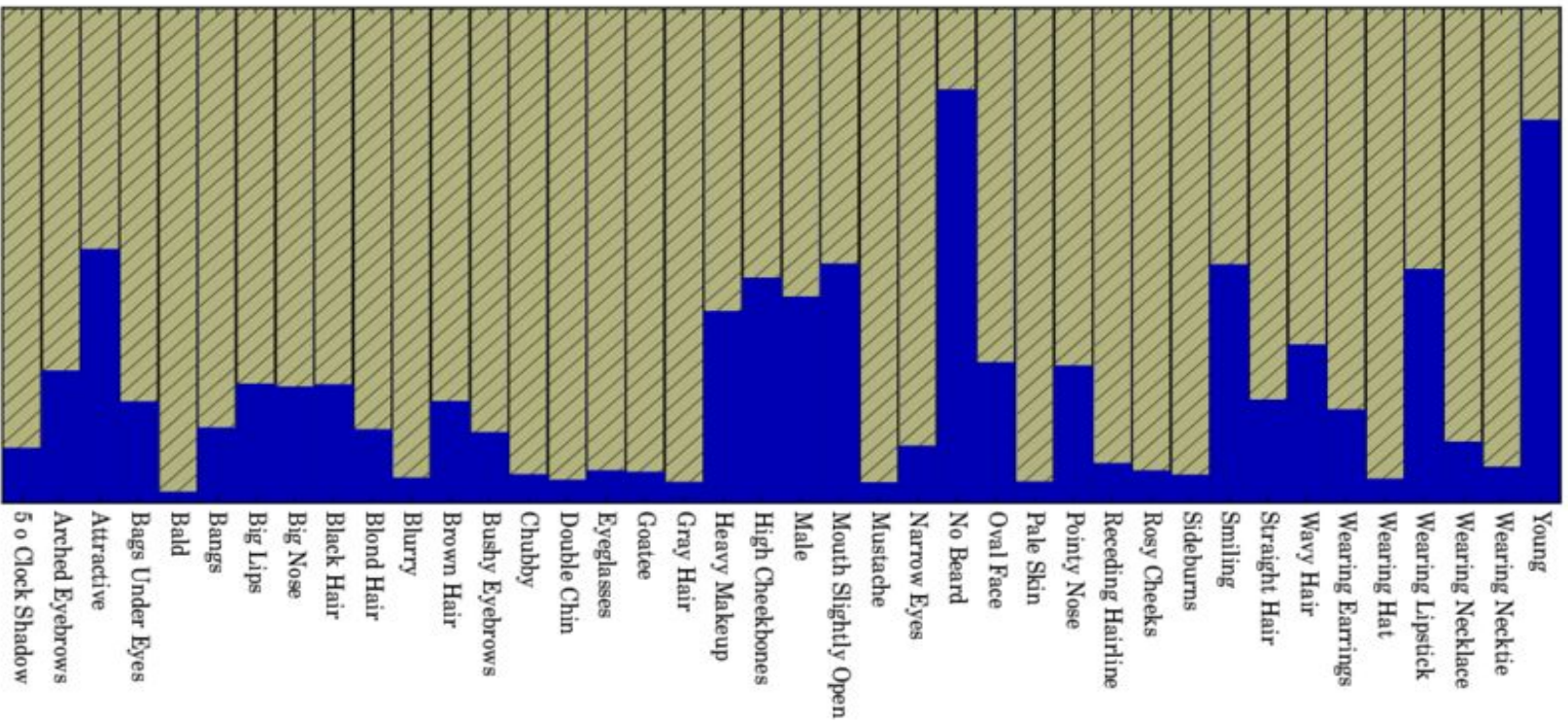
Oval Face



Smiling



CelebA: otros sesgos



Google Photos reconoce personas negras como gorilas

“Solución”: Quitar la clase Gorilla

Año 2015

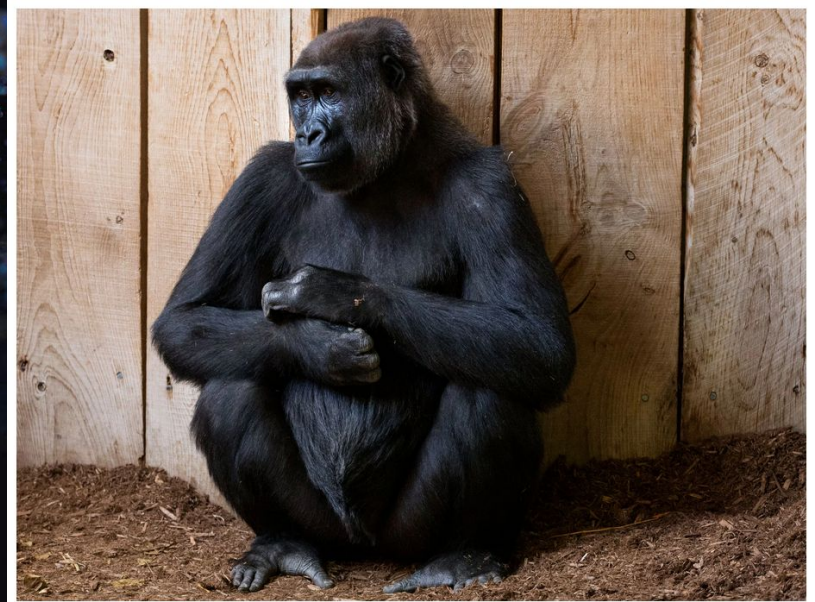
Google Photos, y'all [redacted] ed up. My friend's not a gorilla.
pic.twitter.com/SMkMCsNVX4

— diri noir avec banan (@jackyalcine) [June 29, 2015](#)

Yonatan Zunger, Google's chief social architect, responded quickly.

[@jackyalcine](#) Holy [redacted] G+ CA here. No, this is not how you determine someone's target market. This is 100% Not OK.

— Yonatan Zunger (@yonatanzunger) [June 29, 2015](#)



In WIRED's tests, Google Photos did identify some primates, but no gorillas like this one were to be found. RICK MADONIK/TORONTO STAR/GETTY IMAGES

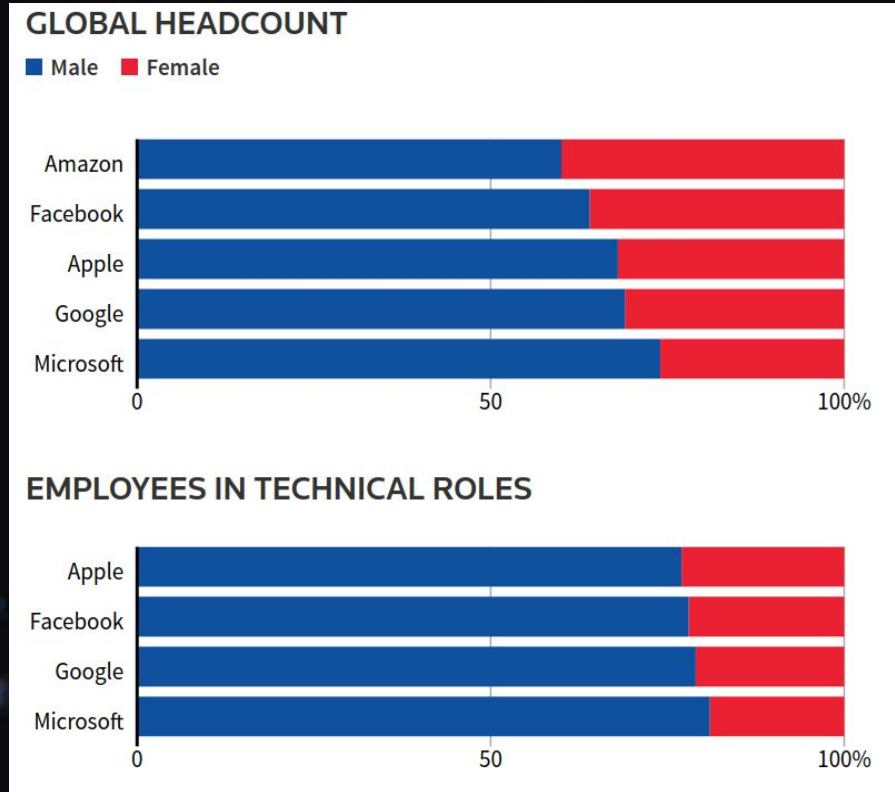
Detección de “parpadeos”



Reclutamiento en Amazon

- Reclutamiento automático
 - Entrada N cvs
 - Salida: mejores K
- Sesgo contra mujeres
 - $P(\text{contratar} \mid \text{atributos y mujer}) < P(\text{contratar} \mid \text{atributos y hombre})$
- Ya no lo usan
 - Pero existe LinkedIn

Reclutamiento: Humanos vs AI



Ofrecen más trabajos en ciencias a hombres (más barato)

An empirical study into apparent gender-based discrimination in the display of STEM career ads

We explore data from a field test of how an algorithm delivered ads promoting job opportunities in the Science, Technology, Engineering and Math (STEM) fields. This ad was explicitly intended to be gender neutral in its delivery. Empirically, however, fewer women saw the ad than men. This happened because younger women are a prized

demographic and are more likely to see ads which simply optimizes cost. The algorithm delivered ads that were intended to be gender neutral in a discriminatory way, due to a regularity that extends to other

Want A Piece Of The 18 Trillion Dollar Female Economy? Start With Gender Bias

Categoría “Odio a Judíos” en Facebook

MACHINE BIAS

Facebook Enabled Advertisers to Reach ‘Jew Haters’

After being contacted by ProPublica, Facebook removed several anti-Semitic ad categories and promised to improve monitoring.

Sistema COMPAS

- Correctional Offender Management Profiling for Alternative Sanctions
- Se utiliza en EEUU
- Calcula $P(\text{criminal reincidente} \mid \text{Antecedentes})$
 - Antecedentes: 137 atributos
 - Padre/madre fue a la cárcel
 - Raza/color de piel
 - Es ilegal preguntarle esto a un acusado
 - Pero el sistema lo usa igual
- Algoritmo privado

Sistema COMPAS

- Evidencia de que está sesgado en contra de afroamericanos
 - Sesgado: $P(\text{criminal reincidente} \mid \text{Antecedentes y afroamericano}) > P(\text{criminal reincidente} \mid \text{Antecedentes y blanco})$
- Estudios critican métodos del estudio anterior
 - Pero no las conclusiones

Sistema COMPAS: sesgo de interacción

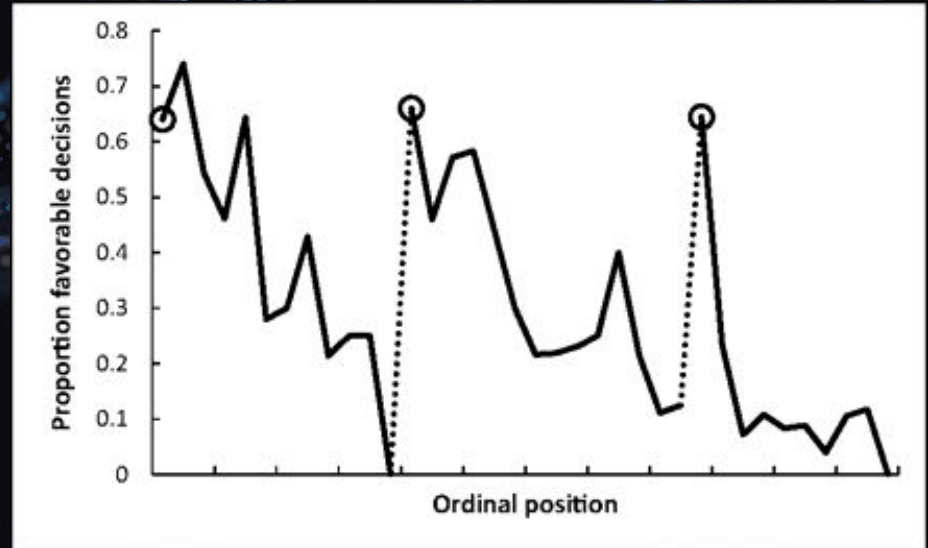
- Utiliza cantidad de condenas previas como atributo
- Se usa desde el 2001
- Asumamos sesgo en 2001 contra negros
 - Si $P(\text{condenar 2001} | \text{antecedentes y negro}) = \text{Media}$
 - Igual encarcelan
 - Persona liberada en 2005
 - Vuelven a arrestar en 2010
 - $P(\text{condenar 2010} | \text{antecedentes y negro y condenado en 2005}) = \text{Alta}$
 - **Genera su propio dataset sesgado**
 - **Aún si se quitan los sesgos del modelo**

COMPAS VS 20 humanos

- COMPAS no es mejor que 20 humanos inexpertos
 - “We show, that the widely used commercial risk assessment software **COMPAS is no more accurate or fair than** predictions made by **people** with little or no criminal justice expertise.”
 - “In addition, despite COMPAS’s collection of 137 features, the same accuracy can be achieved with a simple linear classifier with **only two features.**”

COMPAS vs Jueces

- Estudio sobre fallos de jueces
- Probabilidad de dar libertad condicional vs tiempo del día
- Subjetividad humana
- Estado de ánimo



Predicción de investigación a familias

- Estados Unidos
- Herramienta para trabajadores sociales
- Recomendación qué familias investigar por
 - Abuso
 - Negligencia
- Sesgo: probabilidad de investigación aumenta para niños de color de piel oscura
 - 66% vs 50% para la población general