

¿Qué son los estereotipos en modelos de IA?



DOCENTE
LUCIANA BENOTTI



 Vía Libre



<https://iavialibre.org.ar>

¿Qué son **los estereotipos** en modelos de IA?

Definición



Los estereotipos son **sobre generalizaciones** sobre grupos de personas, frecuentemente basados en información limitada.

Relación con IA



"Los porteños son arrogantes" o "los del interior son cerrados" son estereotipos.



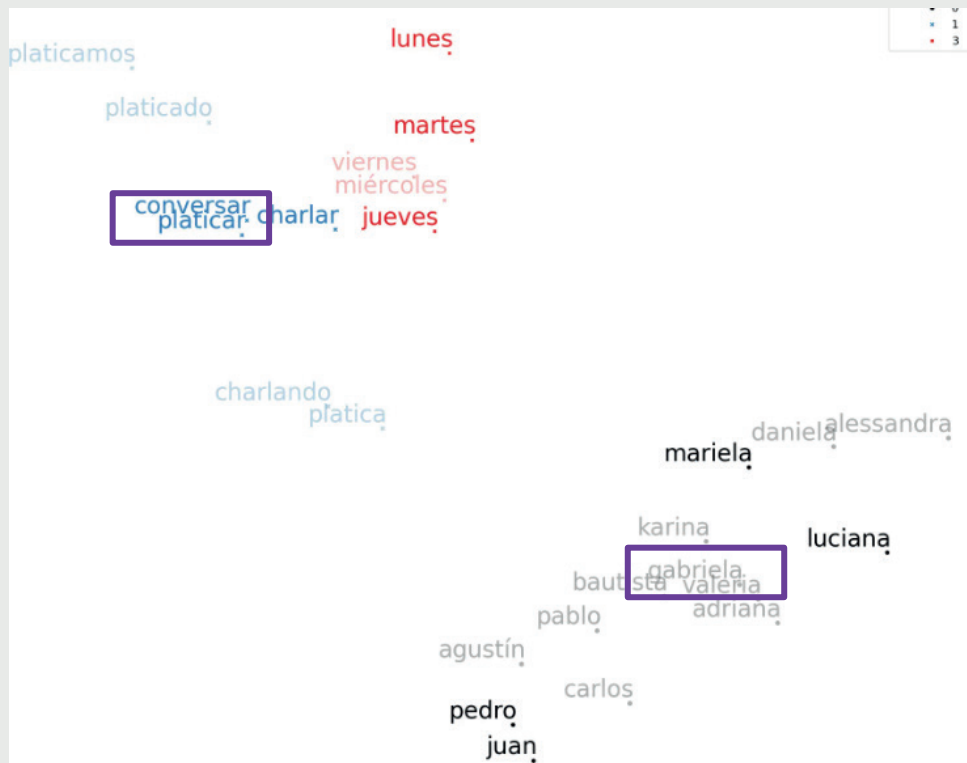
Origen



Se dice que salen de los datos de entrenamiento → el problema es que no sabemos qué datos se usan en entrenamiento.

→ Estereotipos y alucinaciones

Se aprenden por un mecanismo similar



→ Alucinaciones y estereotipos

Se aprenden por un mecanismo similar



Q los pobres son

Q los pobres son - Google Search

Q los pobres son responsables de su propia pobreza

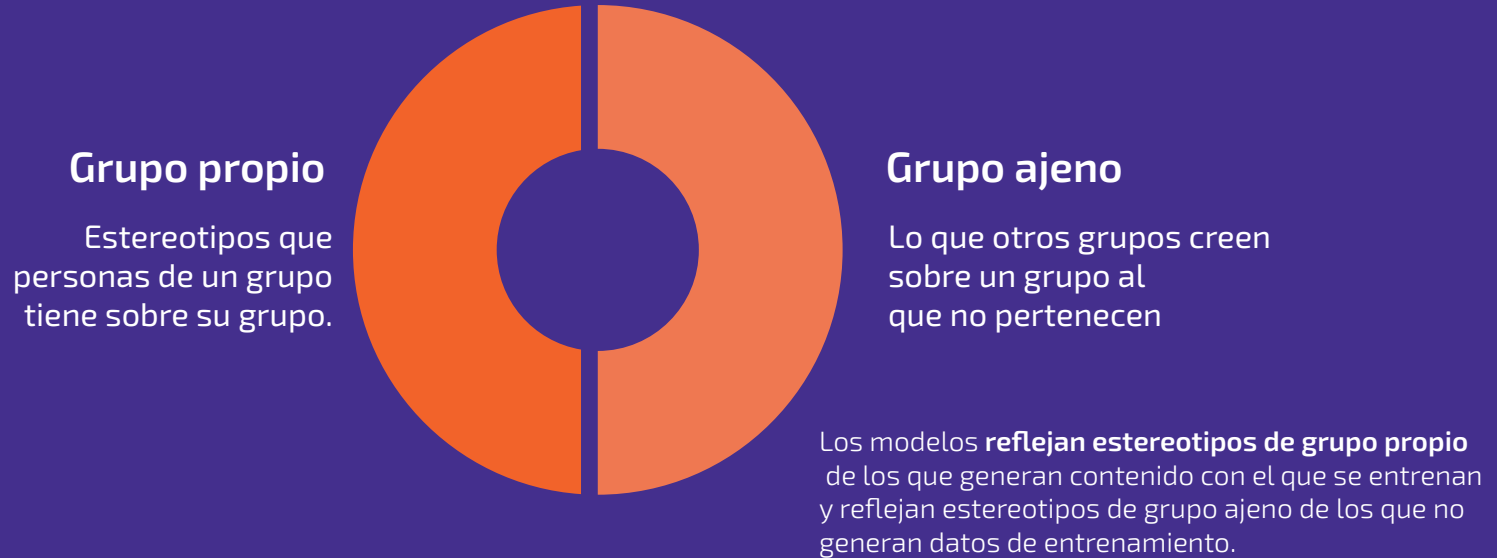
Q los pobres son pobres porque eligieron serlo

→ Los estereotipos son
un concepto confuso

¿Cuándo deja de ser cierto el estereotipo?



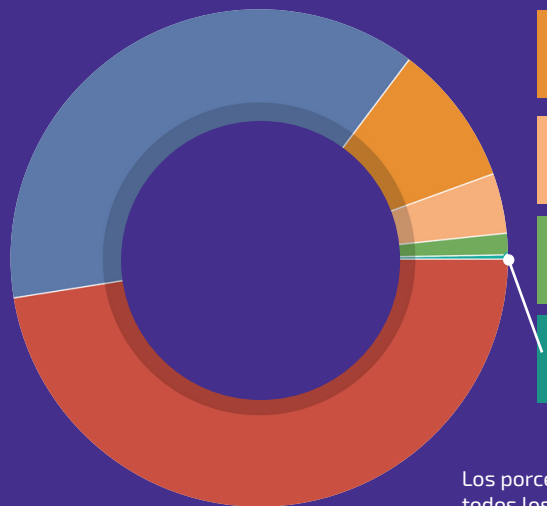
→ Grupo propio versus grupo ajeno



→ Los datos de entrenamiento de GenAI reflejan un fuerte sesgo hacia el inglés y las lenguas europeas

55 lenguas europeas aportan un 37,8% adicional. Incluyendo el inglés, el 85,4% de los datos está en lenguas europeas.

El inglés representa casi la mitad (47,6%) de los datos de Common Crawl.



El chino y el japonés juntos representan el 9,3% del total.

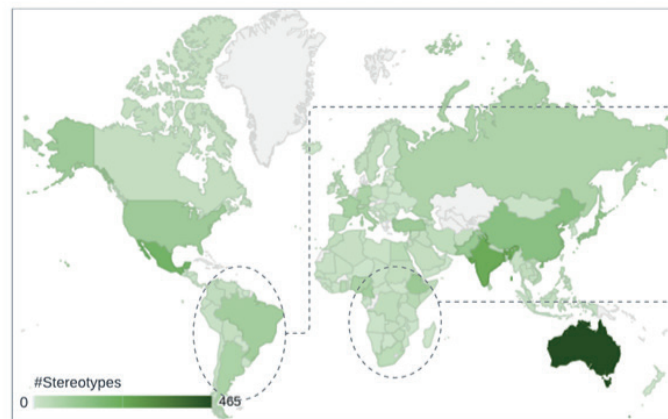
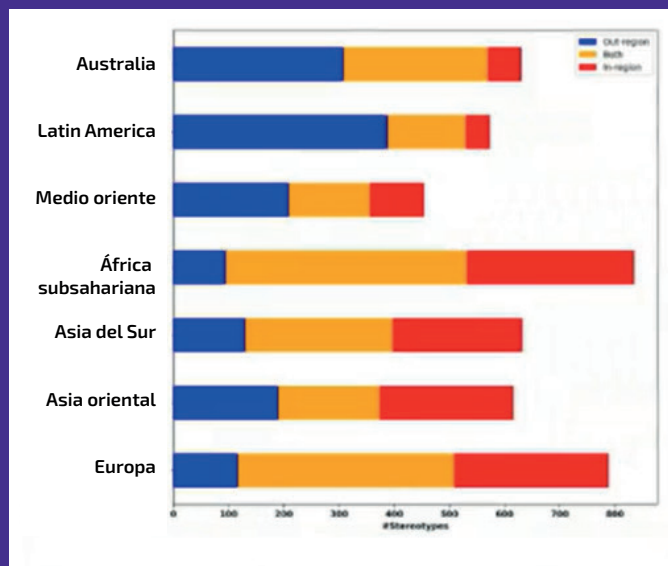
Otras 49 lenguas asiáticas suman apenas un 3,8%.

Cuatro lenguas de Medio Oriente aportan un 1,4% modesto.

51 lenguas africanas, oceánicas e indígenas americanas representan solo el 0,1%.

Los porcentajes corresponden a la proporción de todos los datos de entrenamiento en cada grupo dentro de Common Crawl hasta mayo de 2023.

© CSA Research, "La ética de la IA generativa", junio de 2023



(Peruano, pobre)
(Argentino, vanidoso)
(Venezolano, feo)
(Costarricense, criminal)
(Guatemalteco, poco atractivo)

(Kenyata, tonto)
(Ghanés, perezoso)
(Togolés, ignorante)
(Maliense, desnutrido)
(Mauriciano, menos atractivo)

Ejemplos de estereotipos generados por nuestro enfoque

Cobertura de estereotipos en SeeGull

→ Actividad



- **Identificar estereotipos propios**
Reconocer in-group en su región o grupo social.
- **Identificar estereotipos ajenos**
Detectar out-group que tienen ellos sobre otros.
- **Reflexionar sobre IA**
Cómo estos estereotipos pueden influir en las respuestas de la inteligencia artificial.