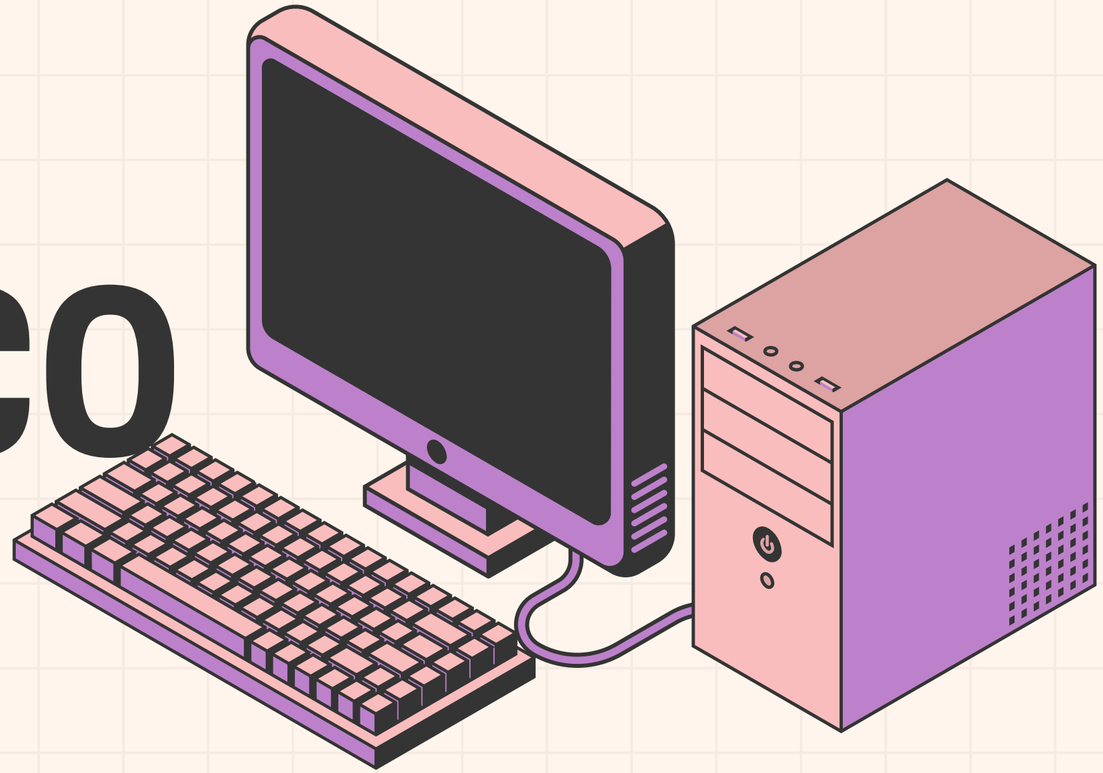
COLAB
TECH

RACISMO ALGORÍTMICO



COM EXEMPLOS EM PYTHON

CONCEITOS DE VIESES NOS DADOS E COMO ELES AFETAM OS SISTEMAS DE IA

O que são Vieses nos Dados?

Vieses nos dados referem-se a distorções ou preconceitos presentes no conjunto de dados que podem influenciar os resultados de um sistema de inteligência artificial (IA).

Eles podem surgir de várias fontes, como:

- **Amostragem**: Quando o **conjunto de dados** não é representativo da população total. Por exemplo, se um sistema de IA é treinado apenas com dados de uma determinada demografia, ele pode não funcionar bem para outras demografias.

- **Exclusão**: Quando certos grupos ou categorias são sub-representados ou completamente omitidos do conjunto de dados.

- **Sobrerrepresentação**: Quando alguns grupos são representados de maneira desproporcional, levando a um modelo que prioriza indevidamente esses grupos.

Como os Vieses Afetam os Sistemas de IA?

- **Decisões Injustas:** Modelos de IA podem perpetuar ou amplificar preconceitos existentes, resultando em decisões injustas, como a rejeição de candidatos a emprego ou empréstimos com base em características irrelevantes ou injustas.

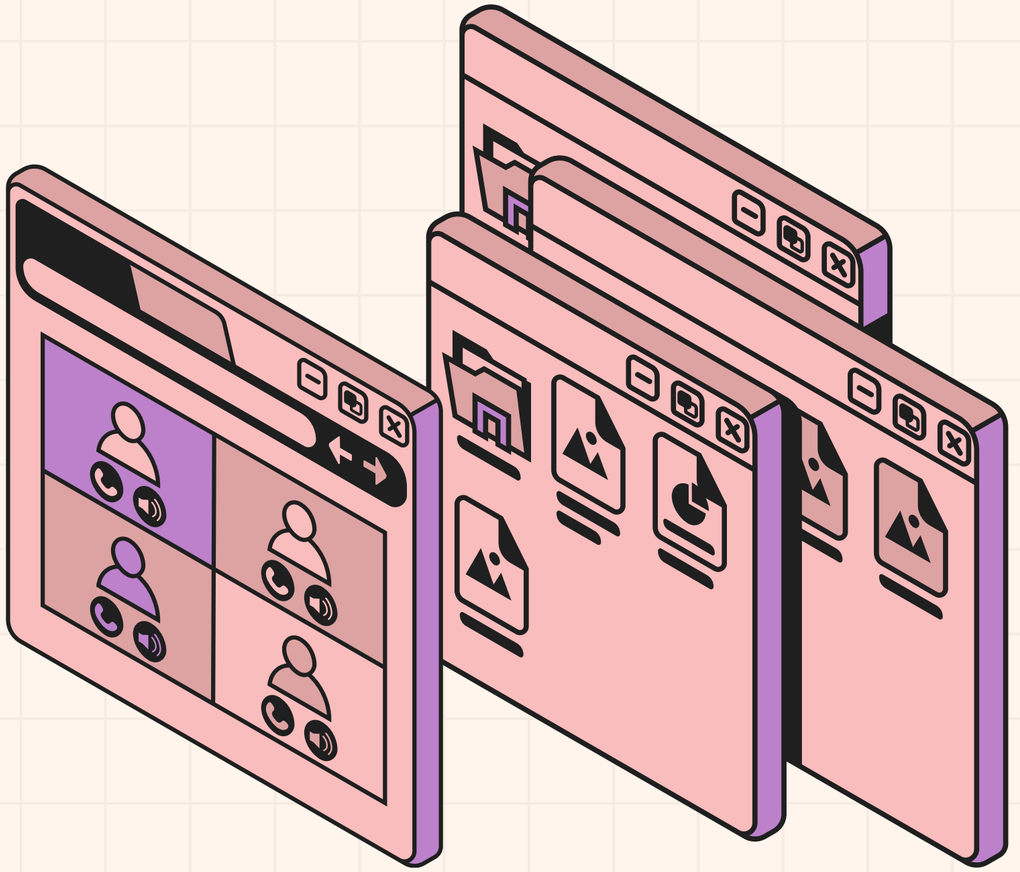
- **Desempenho Inconsistente:** Modelos treinados com dados enviesados podem ter um desempenho inconsistente, funcionando bem para alguns grupos enquanto prejudicam outros.

- **Reforço de Estereótipos:** Sistemas de IA podem reforçar estereótipos sociais, por exemplo, ao associar determinadas profissões ou características a grupos específicos de maneira tendenciosa.

- **Impacto Social:** O uso de IA enviesada pode ter impactos sociais significativos, incluindo discriminação sistemática e perda de confiança pública em sistemas automatizados.

EXEMPLO

Se um sistema de recrutamento é treinado em dados históricos de contratações, onde há um viés contra certos grupos demográficos, o sistema pode aprender a replicar esses vieses, prejudicando candidatos qualificados de grupos sub-representados.



JANEIRO 2010

CÂMERAS DA NIKON N ENTENDEM ROSTOS ASIÁTICOS

Recurso para evitar selfies com olhos fechados se c
com olhos de asiáticos

MARÇO 2016

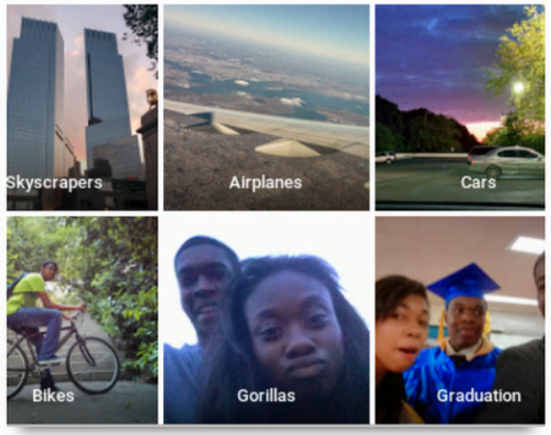
CHATBOT DA MICROSOFT TORNA-SE RACISTA EM MENOS DE UM DIA

A chatbot Tay, que constrói discurso a partir de aprendizado de máquina, virou racista e xenófoba em menos de um dia, mostrando falta de compreensão da sociedade pelos engenheiros da empresa.

SISTEMA
EXPLORAÇÃO
CONTEÚ
ENVIESA BUSC
P
CONHECIMEN
MARÇO 20

2006

TERRORISTA FOI RADICALIZADO POR RESULTADOS NO GOOGLE

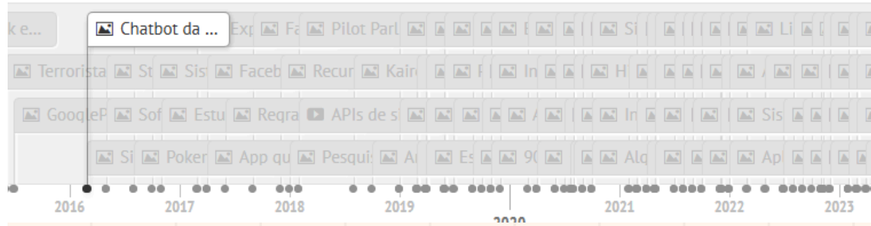


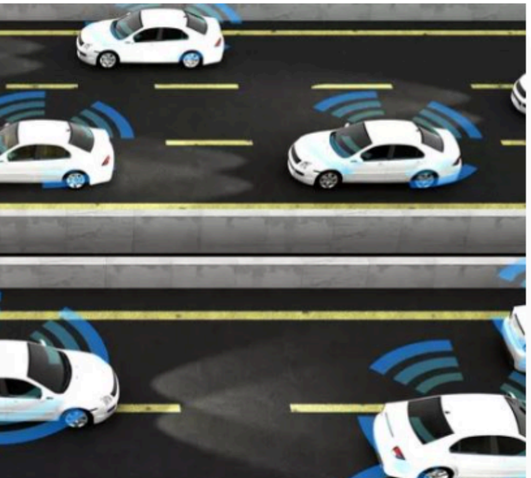
JULHO 2015

GOOGLEPHOTOS TAGGEOU PESSOAS NEGRAS COMO "GORILAS"

Desenvolvedor Jacky Alciné denuncia que visão computacional do GooglePhotos marcou pessoas negras como "gorilas", como reportado no The Guardian

CHATBOT DA MICROSOFT TORNA-SE RACISTA EM MENOS DE UM DIA

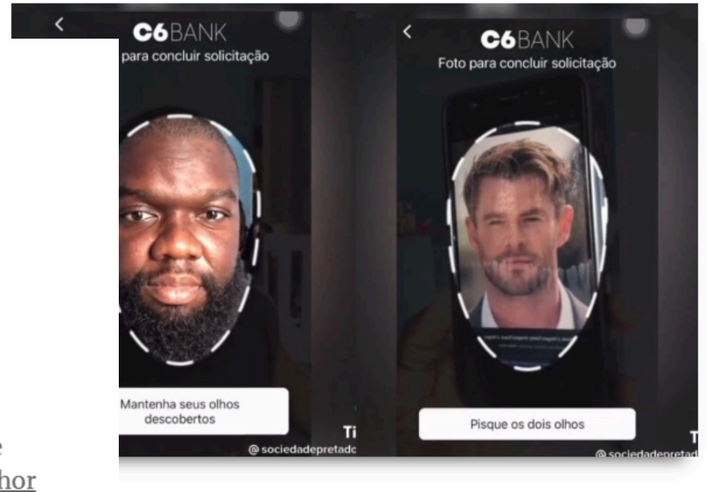




MARÇO 2019

CARROS AUTÔNOMOS TEM MAIS CHANCE DE ATROPELAR PESSOAS NEGRAS

Pesquisadores da George Institute of Technology descobriram que a visão computacional em sistemas de carros autônomos foram treinados para identificar melhor pedestres de pele clara.



JANEIRO 2022

SISTEMA DO C6BANK NÃO RECONHECE CORRENTISTA NEGRO

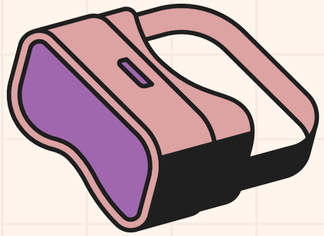
Falhas do sistema são tão vulgares contra correntista negro que até uma foto de ator branco em celular funcionou - mas não a sua

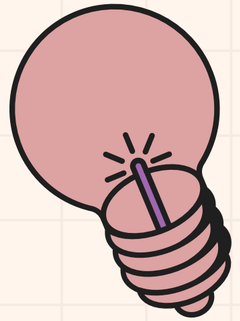
IDENTIFICAÇÃO E MITIGAÇÃO

Para identificar e mitigar vieses, é essencial:

- **Auditar os Dados:** Examinar os dados para identificar desequilíbrios ou representações inadequadas.
- **Métricas de Justiça:** Utilizar métricas como paridade demográfica e equidade de oportunidade para medir o viés.
- **Re-treino e Ajuste de Modelos:** Implementar técnicas para balancear o conjunto de dados ou ajustar o modelo para minimizar o viés.

BIBLIOTECAS





1. **Pandas**:
   - **Uso**: Manipulação e análise de dados, especialmente dados tabulares (como planilhas e CSVs).
   - **Funcionalidades**: DataFrames para manipulação de tabelas, leitura e escrita de arquivos (CSV, Excel, JSON), limpeza e transformação de dados.
2. **NumPy**:
   - **Uso**: Computação numérica e manipulação de arrays multidimensionais.
   - **Funcionalidades**: Arrays de alta performance, operações matemáticas e estatísticas, integração com outras bibliotecas como pandas e scikit-learn.
3. **Matplotlib**:
   - **Uso**: Criação de visualizações de dados, como gráficos de linhas, barras, histogramas, dispersão, entre outros.
   - **Funcionalidades**: Controle total sobre a aparência dos gráficos, integração com outras bibliotecas de visualização como seaborn.