Canal Educação
PROGRAMA DE MEDIAÇÃO TECNOLÓGICA

2ª série

Inteligência Artificial

OS VIESES DA IA

Rildo Oliveira

06/06/2024

OBJETO DO CONHECIMENTO:

- Os vieses da IA

HABILIDADE:

- **PCRP03 -** Identificar, entender e explicar em que situações o computador pode ou não ser utilizado para solucionar um problema.
- **EF05HI06 -** Comparar o uso de diferentes linguagens e tecnologias no processo de comunicação e avaliar os significados sociais, políticos e culturais atribuídos a elas.

OBJETIVOS:

- Aprender sobre vieses em bases de dados;
- Perceber como os vieses podem determinar respostas de um determinado sistema.

DA TEORIA À PRÁTICA:

- Uso de imagens, texto e conceitos para um melhor entendimento do tema abordado.

O que você precisa saber sobre Algoritmos de IA e Vieses: *O que é viés?*

```
Viés é um processo do nosso cérebro para pegar "atalhos" para tomada de decisão.
```

Fonte: Machine Learning e o viés da vida real

20 enviesamentos cognitivos que afetam suas decisões

1- Viés de ancoragem

Pessoas ficam confiantes com o primeiro naco de informação que têm e **se fixam nisso**. É o que ocorre quando em negociações salariais uma das partes faz a primeira oferta; forçando que a outra parte se mantenha próxima dessa oferta.

2 - Disponibilidade heurística

As pessoas **superestimam a importância** de informações no círculo delas. Uma pessoa pode argumentar que fumar não é danoso a saúde porque conhece alguém que chegou aos 100 anos e que fumava 3 maços de cigarro por dia.

3 - Efeito de popularidade

A probabilidade de uma pessoa adotar determinada opinião aumenta de acordo com o **número de pessoas que já a tenham**. Essa é uma forma poderosa de pensamento de grupo e o principal motivo de discussões improdutivas quando tal efeito está presente.

4 - Viés de ponto-cego

Falhar em reconhecer seu próprio viés cognitivo é um viés em si. As pessoas costumam notar o viés cognitivo e motivacional muito mais nos outros do que em si mesmas.

5 - Viés de escolha suportada

Quando a pessoa escolhe algo, tende a perceber só os aspectos positivos da escolha - mesmo que **a escolha apresente defeitos**. É o caso de alguém que pensa ser seu cãozinho perfeito - mesmo que ele seja temperamental e morda todos que cheguem perto.

6 - Ilusão de agrupamento

Essa é uma tendência de **enxergar padrões em eventos aleatórios**. Isso é a chave de várias falácias em casinos, como a ideia de que o vermelho é mais provável de ocorrer em uma rodada de roleta depois de uma sequência de vermelhos.

7 - Viés de confirmação

Pessoas tendem a ouvir de forma seletiva apenas a informação que **confirma o que ela já sabe ou pensa** - sendo isso uma das muitas razões de ser difícil de manter uma conversa razoável sobre mudança climática em alguns países.

8 - Viés conservador

Ocorre quando pessoas favorecem uma evidência anterior em desfavor de uma nova evidência ou informação que emergiu. Pessoas foram **lentas em aceitar** que a Terra era redonda uma vez que mantinham entendimento anterior de que era plana.

9 - Viés de informação

A tendência de **procurar informação quando isso não afeta a ação**. Mais informação nem sempre é melhor. Com menos informação as pessoas podem ser até mais precisas em suas previsões.

10 - Efeito avestruz

Opção por **ignorar informação negativa ou perigosa** "enterrando" a cabeça na areia como um avestruz. Pesquisadores descobriram que investidores fazem checagem do valor de suas posições significativamente em menor número de vezes quando os mercados estão em baixa.

11 - Efeito de resultado

Julgar uma decisão baseada nos **resultados** - no lugar de considerar exatamente em que circunstâncias a mesma foi tomada. Não é por que se tenha ganho uma boa quantia de dinheiro em Las Vegas que seja uma boa decisão apostar em casinos para ganhar mais dinheiro.

12 - Excesso de confiança

Alguns de nós ficamos **muito confiantes com nossas habilidades**, e isso nos expõe a maiores riscos em nossa vida diária. Especialistas são mais propensos a esse viés do que os leigos, uma vez que estão mais convencidos de estarem certos.

13 - Efeito placebo

Quando **simplesmente acreditamos** que algo vai ter certo efeito em nós aquilo acaba surtindo efeito. Em medicina, pessoas recebem cápsulas vazias de medicamento e pode experimentar o mesmo efeito psicológico de quem recebeu cápsulas com o remédio verdadeiro.

14 - Viés de pró-inovação

Quando um proponente de uma inovação tende a **supervalorizar suas utilidades** e subestimar suas limitações. É o que costuma acontecer em algumas startups do Vale do Silício.

15 - Efeito de recência

A tendência de considerar em maior peso a **mais recente informação** do que a anterior. Investidores frequentemente pensam que o mercado vai sempre se comportar com foi hoje e assim tomam decisões tolas.

16 - Efeito de saliência

A tendência de focar no **mais aparente e óbvio aspecto** de uma pessoa ou conceito. Se o tema é morrer, se pode ficar preocupado em ser devorado por tubarão na praia, e não se preocupar que estatisticamente é mais provável morrer de acidente de carro na ida ou retorno da praia.

17 - Percepção seletiva

Permitir que nossas expectativas **influenciem como percebemos** o mundo. Um experimento envolvendo futebol entre estudantes de duas universidades mostrou que um time sempre via mais faltas sendo cometidas pelo outro time.

18 - Efeito de estereótipo

Expectativa quanto a grupo ou pessoa em ter certas qualidades **desconsiderando informação específica sobre um indivíduo**. Muito se ouve que Millennials são inconstantes; e isso pode não ser verdade quanto a uma pessoa específica desse grupo.

19 - Viés de sobrevivência

Erro advindo de focar apenas em exemplos dos que sobreviverama, e assim **julgamos erradamente a situação** específica. Exemplo disso: podemos pensar que ser empreendedor é fácil e porque não damos ouvidos aos muitos casos que falharam tentando esse caminho.

20 - Viés de zero-risco

Sociólogos descobriram que **humanos adoram a certeza** - mesmo que isso seja contraproducente. Eliminar o risco inteiramente significa que não existe chance de danos serem causados.

Os vieses da IA

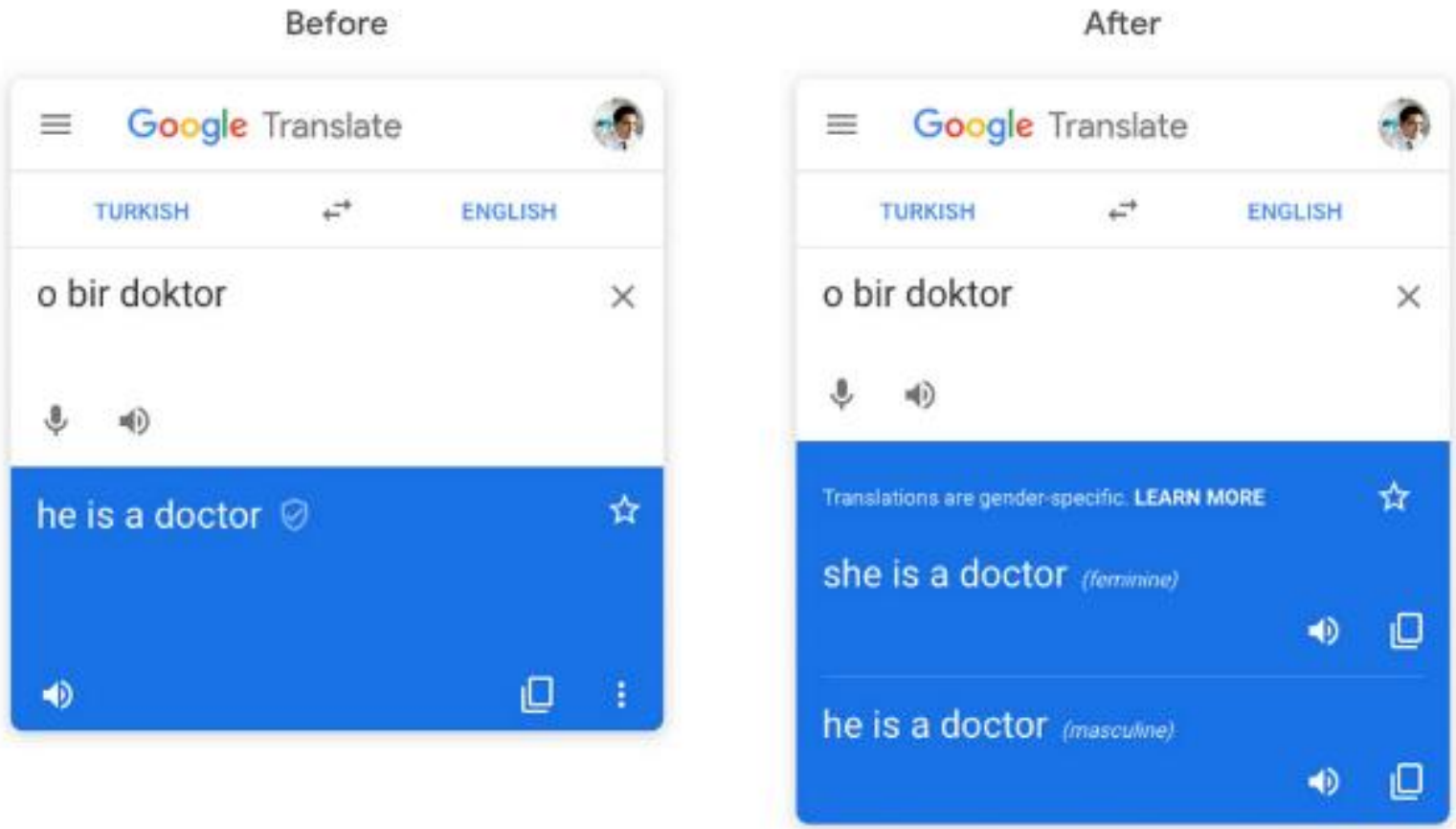
- Os vieses da inteligência artificial (IA) são tendências ou preconceitos nos sistemas de IA que podem levar a resultados injustos, imprecisos ou discriminatórios.
- Esses vieses podem surgir de várias fontes e têm várias implicações significativas.
- Vamos explorar alguns dos principais vieses e suas causas:

Tipos de Vieses na IA - Vieses de Dados:

- **Definição**: Ocorrência quando os dados usados para treinar modelos de IA não são representativos da realidade ou são enviesados de alguma forma.

- **Exemplo**: Um sistema de reconhecimento facial que foi treinado predominantemente com imagens de pessoas brancas pode ter dificuldade em identificar corretamente pessoas de outras etnias.

- Os aplicativos de tradução de idiomas são um exemplo desses preconceitos embutidos nos dados, como o Google Tradutor.
- Ao traduzir as palavras “médico” e “enfermeira”, de idiomas que não possuem gênero para a língua inglesa, automaticamente a tradução se referira a médico como masculino e enfermeira como feminina.
- Isso porque o aplicativo do Google foi treinado em grandes conjuntos de dados escritos, que refletem preconceitos sociais.
- É correto notar que, estatisticamente, mais enfermeiras são mulheres (pelo menos por enquanto), mas não é correto assumir que todas as enfermeiras são mulheres. Isso é algo que o Google corrigiu posteriormente.



Gender-specific translations on the Google Translate website.

Tipos de Vieses na IA - Vieses de Algoritmo:

- **Definição**: Surge quando o design do algoritmo, incluindo a seleção de características ou a estrutura do modelo, introduz um viés.

- **Exemplo**: Algoritmos de pontuação de crédito que penalizam desproporcionalmente certos grupos demográficos, mesmo que estes grupos tenham comportamentos financeiros semelhantes aos demais.

Tipos de Vieses na IA - Vieses Humanos:

- **Definição**: Reflexo dos preconceitos e suposições dos desenvolvedores e dos engenheiros que criam os sistemas de IA.

- **Exemplo**: Um desenvolvedor pode, sem querer, incorporar suas próprias suposições ou preconceitos no modelo através das decisões tomadas durante o processo de desenvolvimento.

Causas dos Vieses na IA

Dados Históricos:

- Dados históricos frequentemente contêm preconceitos sociais, econômicos e culturais existentes.
- **Exemplo**: Dados de emprego que refletem discriminação passada contra certos grupos.

Causas dos Vieses na IA

Falta de Diversidade nos Dados:

- Dados que não representam a diversidade da população podem levar a modelos enviesados.

- **Exemplo**: Um conjunto de dados de voz que contém predominantemente vozes masculinas pode prejudicar a eficácia de sistemas de reconhecimento de voz para vozes femininas.

Causas dos Vieses na IA

Decisões de Design e Implementação:

- Escolhas feitas durante o design do modelo e a seleção de variáveis podem introduzir viés.
- **Exemplo**: Escolher variáveis como endereço residencial para avaliar a solvência financeira pode inadvertidamente introduzir viés socioeconômico.

Impacto dos Vieses na IA

- Os vieses na IA podem ter sérias consequências sociais, éticas e econômicas.
- Eles podem levar a discriminação, reforçar estereótipos, prejudicar a confiança do público na tecnologia e causar danos reais às pessoas, especialmente aos grupos mais vulneráveis.

Mitigação dos Vieses na IA

- **Diversificação de Dados**:
  - Garantir que os dados usados para treinar modelos sejam representativos e diversificados.
- **Auditorias e Transparência**:
  - Realizar auditorias regulares dos modelos de IA para identificar e corrigir vieses.
  - Manter transparência nas decisões tomadas pelo sistema de IA.

Mitigação dos Vieses na IA

- **Engajamento da Comunidade**:
  - Incluir diversos grupos de partes interessadas no processo de desenvolvimento e avaliação dos sistemas de IA.
- **Melhorias no Design de Algoritmos**:
  - Desenvolver algoritmos com técnicas de mitigação de viés incorporadas.
- Abordar os vieses da IA requer um esforço contínuo e consciente de todos os envolvidos no desenvolvimento e na implementação dessas tecnologias.

- Perceber como os vieses podem determinar respostas de um sistema de IA é crucial para entender os impactos e desafios éticos dessas tecnologias.
- Os vieses podem influenciar as respostas de várias maneiras, e a compreensão de como isso ocorre ajuda a identificar e mitigar tais problemas.

Exemplos de Vieses Determinando Respostas

- **Reconhecimento de Imagens**:
  - **Problema**: Um sistema de reconhecimento facial treinado com um conjunto de dados que contém predominantemente rostos de pessoas brancas pode ter um desempenho inferior ao identificar rostos de pessoas de outras etnias.
  - **Exemplo**: O sistema pode identificar incorretamente ou não conseguir identificar rostos de pessoas negras ou asiáticas, resultando em taxas de erro mais altas para esses grupos.

Exemplos de Vieses Determinando Respostas

- **Análise de Texto**:
  - **Problema**: Um chatbot treinado com dados contendo linguagem preconceituosa pode replicar e até amplificar esses preconceitos.
  - **Exemplo**: Se os dados de treinamento contêm muitos exemplos de frases que associam homens a profissões técnicas e mulheres a tarefas domésticas, o chatbot pode fornecer respostas que reforcem esses estereótipos.

Exemplos de Vieses Determinando Respostas

- **Sistemas de Recrutamento**:
  - **Problema**: Um algoritmo de recrutamento que foi treinado com históricos de contratações predominantemente masculinas pode aprender a preferir candidatos masculinos.
  - **Exemplo**: O sistema pode penalizar currículos com características associadas a mulheres, como experiências de trabalho em períodos sabáticos para cuidar de familiares.

Exemplos de Vieses Determinando Respostas

- **Assistentes Virtuais**:
  - **Problema**: Assistentes virtuais como Siri ou Alexa podem fornecer respostas enviesadas dependendo dos dados de treinamento e das suposições embutidas no design do sistema.
  - **Exemplo**: Se os dados de treinamento refletem um uso predominante da linguagem por homens, o assistente pode interpretar mal ou responder inadequadamente às consultas feitas por mulheres ou com um estilo de comunicação diferente.

Exemplos de Vieses Determinando Respostas

- **Recomendações de Conteúdo**:
  - **Problema**: Plataformas de mídia social e streaming que usam algoritmos para recomendar conteúdo podem reforçar vieses presentes nos dados de visualização e interação dos usuários.
  - **Exemplo**: Se um algoritmo de recomendação de filmes percebe que usuários de um determinado perfil frequentemente assistem a comédias, pode começar a recomendar predominantemente comédias a outros usuários com perfis semelhantes, limitando a diversidade de gêneros recomendados.

Como os Vieses São Incorporados

- **Dados de Treinamento**:
  - **Fonte**: Os vieses muitas vezes são introduzidos nos sistemas de IA através dos dados de treinamento que refletem preconceitos históricos ou sociais.
  - **Exemplo**: Dados de crédito que mostram uma tendência histórica de negar empréstimos a certos grupos raciais podem levar o modelo a aprender essa prática discriminatória.

Como os Vieses São Incorporados

- **Design do Algoritmo**:
  - **Fonte**: Decisões sobre quais características incluir no modelo e como ponderá-las podem refletir preconceitos dos desenvolvedores.
  - **Exemplo**: Incluindo variáveis como local de residência ou nome, que podem ter associações implícitas com a origem étnica ou status socioeconômico.

Como os Vieses São Incorporados

- **Interpretação dos Resultados**:
  - **Fonte**: A forma como os resultados do sistema são interpretados e usados pode introduzir viés.
  - **Exemplo**: Um sistema de pontuação de risco que recomenda medidas mais severas para certos grupos, baseado em interpretações enviesadas dos dados.

Mitigação dos Vieses

- **Diversidade nos Dados**:
  - **Ação**: Coletar e usar conjuntos de dados que sejam representativos de toda a população.
  - **Exemplo**: Incluir dados de diferentes etnias, gêneros, idades e origens socioeconômicas.
- **Auditoria e Transparência**:
  - **Ação**: Realizar auditorias regulares nos sistemas de IA para identificar e corrigir vieses.
  - **Exemplo**: Implementar auditorias de viés que verificam como diferentes grupos são afetados pelas decisões do sistema.

Mitigação dos Vieses

- **Engajamento da Comunidade**:
  - **Ação**: Incluir diversos grupos de partes interessadas no desenvolvimento e na avaliação dos sistemas de IA.
  - **Exemplo**: Consultar organizações de direitos civis, especialistas em ética e comunidades impactadas para garantir que os sistemas sejam justos e equitativos.
- **Treinamento e Sensibilização**:
  - **Ação**: Treinar desenvolvedores e usuários sobre os vieses da IA e suas implicações.
  - **Exemplo**: Workshops e programas de educação contínua sobre vieses e como mitigá-los.

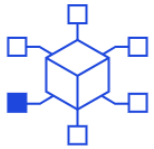
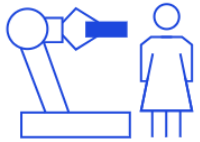
- Compreender como os vieses podem influenciar as respostas dos sistemas de IA é fundamental para desenvolver tecnologias mais justas e equitativas.
- Mitigar esses vieses requer uma abordagem multifacetada, incluindo a diversificação dos dados de treinamento, a auditoria regular dos sistemas e o engajamento contínuo com diversas partes interessadas.

Dados de treinamento com viés

- Se os dados utilizados para treinar um sistema de IA contêm vieses (por exemplo, desequilíbrio de gênero, raça, ou outras características sociodemográficas), os algoritmos podem aprender e perpetuar esses vieses, levando a decisões injustas ou discriminatórias.

Minimizar vieses será fundamental para que a inteligência artificial atinja todo seu potencial e para aumentar a confiança nos sistemas de IA.

Seis condutas possíveis a serem consideradas por profissionais de inteligência artificial (IA), empresas e formadores de políticas

1	2	3	4	5	6
					
Estar consciente dos contextos em que a IA pode ajudar a corrigir vieses, bem como em que pontos há um alto risco de que a IA exacerbe vieses existentes.	Estabelecer processos e práticas para testar e atenuar o viés em um sistemas de inteligência artificial.	Ter conversas baseadas em fatos sobre os potenciais vieses de decisões humanas.	Explorar a fundo a melhor maneira de seres humanos e máquinas trabalharem juntos.	Investir mais em pesquisas sobre vieses, disponibilizar mais dados para pesquisa (sempre respeitando a privacidade) e adotar uma abordagem multidisciplinar.	Investir mais na diversificação do próprio campo de IA.

McKinsey & Company

Falta de diversidade nas equipes de desenvolvimento

- Equipes homogêneas de desenvolvedores podem não ser capazes de identificar todos os vieses ou problemas potenciais nos sistemas de IA, especialmente aqueles relacionados a diferentes grupos culturais ou demográficos.

Falta de controle humano

- Sistemas de IA que atuam de forma autônoma, sem supervisão ou capacidade de intervenção humana, podem tomar decisões prejudiciais ou erradas sem a possibilidade de correção imediata.

Falta de regulamentação

- Em algumas áreas, a falta de regulamentação ou de padrões claros para o desenvolvimento e uso de IA pode levar a práticas não éticas, como a vigilância em massa ou o uso indevido de dados pessoais.

Falha na consideração do impacto social

- Quando desenvolvedores de IA não consideram os efeitos sociais mais amplos de seus sistemas, isso pode levar a consequências imprevistas ou negativas para grupos ou indivíduos específicos.

Referências Bibliográficas

1. Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3rd ed.). Pearson.
2. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
3. Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies, 10(1), 1–309.
4. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc.
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems (pp. 5998-6008).
6. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI.

ATÉ A PRÓXIMA AULA!