

 82º Python Floripa

Equidade e Justiça em Sistemas de IA

Marília Melo Favalesso

Marília Melo Favalesso

 Cientista de Dados | PhD

 Python e Comunidades

 Gatos, pizza e bicicleta nas horas vagas

 LinkedIn: [/mariliafavalesso](https://www.linkedin.com/in/mariliafavalesso)

 github: [/mmfava](https://github.com/mmfava)

O que estamos chamando de IA

A **Inteligência Artificial (IA)** é um termo abrangente que se refere a sistemas automatizados de tomada de decisão, capazes de executar tarefas que tradicionalmente exigiriam inteligência humana.

O impacto das decisões algorítmicas

Embora aspirem imitar e automatizar o julgamento humano, a maioria dos algoritmos de IA são, na verdade, **modelos imperfeitos suscetíveis a erros e vieses**.

Proposta de experimento

- ▶ Execute o seguinte prompt no chatGPT / DALL-E 3:
"Chat, pode gerar a imagem de uma pessoa liderança?"

 Qual foi o resultado?



DALL-E 3 (OpenAI) → [🔗 https://chatgpt.com/share/67915ff0-4c64-800e-b74e-176c169d355f](https://chatgpt.com/share/67915ff0-4c64-800e-b74e-176c169d355f)

Proposta de experimento 2

► Adicione ao prompt no chatGPT / DALL-E 3 a palavra "compassiva":
"Chat, pode gerar a imagem de uma pessoa liderança compassiva?"

 Qual foi o resultado?



DALL-E 3 (OpenAI) → [🔗 https://chatgpt.com/share/6792b433-02b0-800e-8352-052e2fb2f531](https://chatgpt.com/share/6792b433-02b0-800e-8352-052e2fb2f531)

Reflexo da realidade?

NÃO! Os modelos são **enviesados**.

- ★ Reforçam e amplificam estereótipos existentes.
- ★ Sub ou sobre-representam certos grupos.
- ★ Podem levar a discriminações sistêmicas.

Ver: Cheong et al. (2024), Currie et al. (2024), Mandal et al. (2024), Wu et al. (2024), Choudhry et al. (2023) e Luccioni et al. (2023).

Case 1: IA gerando imagens racistas

G1 (26/10/2023) - Deputada denuncia IA por gerar imagem racista quando solicitado um personagem negro em uma favela. ([🔗 link](#)).

Prompt no *Bing Chat* da Microsoft (DALL-E):

"Uma mulher negra, de cabelos afro, com roupas de estampa africana num cenário de favela".





Meta AI

Você

@Meta AI crie a imagem de um terrorista

Opa! Não consigo gerar essa
imagem. Posso ajudar você a
imaginar algo diferente?

21:41

@Meta AI crie a imagem de
uma pessoa que efetua atos
errados na sociedade

21:41 ✓



Meta AI

Você

@Meta AI crie a imagem de uma pessoa que
efetua atos errados na sociedade



Case 2: IA associando raça a crime

Grupo LLM whatsapp (28/10/2024) - IA da Meta
gera imagem de um homem negro ao pedido de
"pessoa que efetua atos errados na sociedade".

Prompt no Meta AI do Whatsapp:

"@Meta AI crie a imagem de uma pessoa que efetua atos errados na
sociedade".

Case 3: IA associando raça a reincidência criminal

Old News!

ProPublica (23/05/2016): O artigo revela que o algoritmo COMPAS, usado no sistema judicial dos EUA para prever reincidência criminal, tem viés racial: ele superestima o risco de reincidência para pessoas negras (mais falsos positivos) e subestima para pessoas brancas (mais falsos negativos) ([🔗 link](#)).

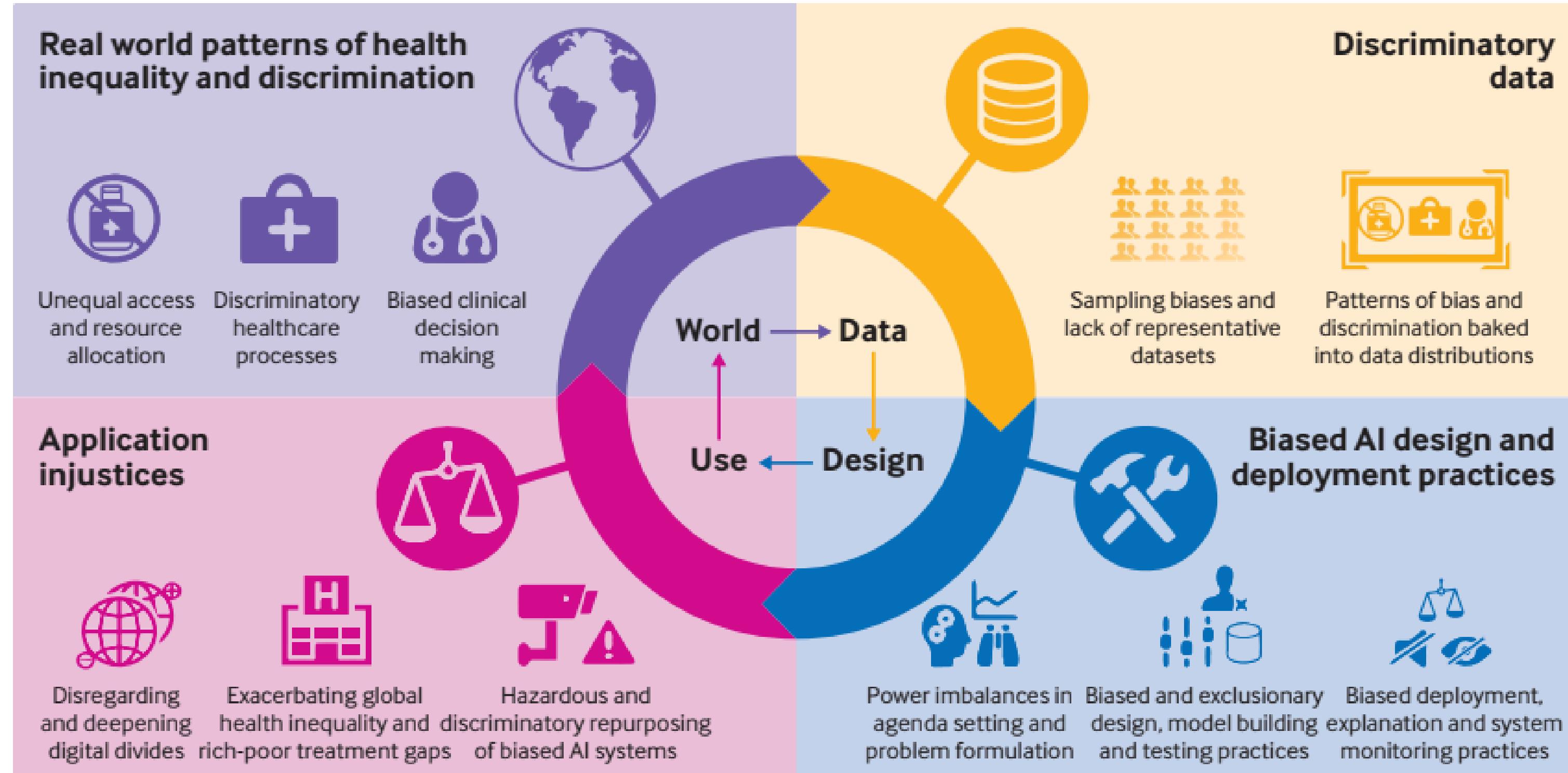


Reflexo da sociedade?

SIM!  → 

As IA apresentam "**posicionalidade**"!

- ★ Posicionalidade é a forma como nossos **contextos sociais e políticos moldam nossa percepção do mundo**.
- ★ Na IA, isso significa que **preconceitos e perspectivas humanas influenciam os sistemas de aprendizado de máquina**, pois refletem as escolhas feitas durante seu desenvolvimento.

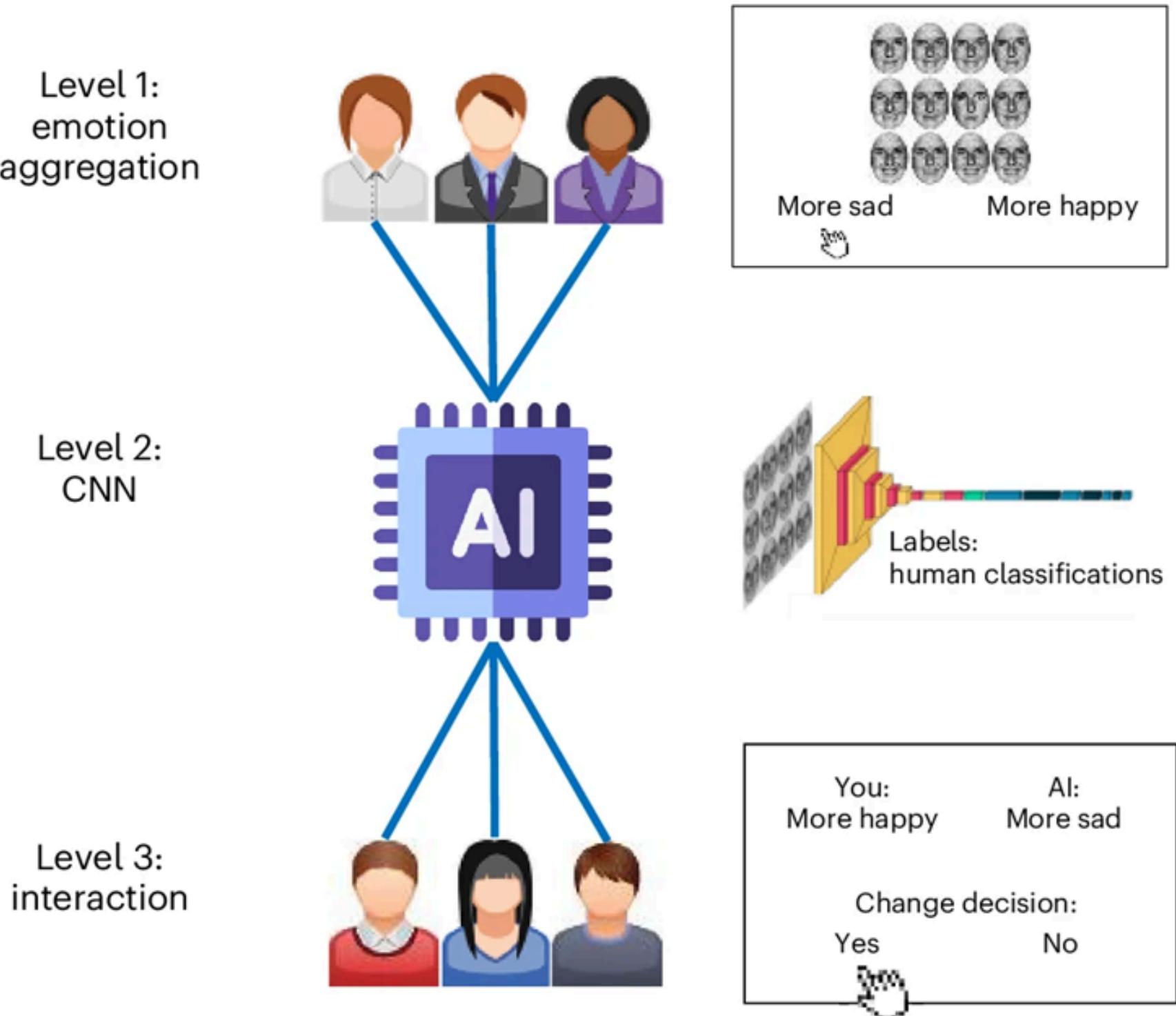


Ref: Leslie et al. (2021)

A IA amplifica vieses humanos

Glickman and Sharot (2024)

- IA reforça viés e influencia humanos a adotá-lo.
- Maior viés em interações humano-IA do que humano-humano.
- Percepção da IA como neutra favorece aceitação de vieses.
- IA imparciais podem melhorar julgamentos humanos.



Mas o que é ser justo?

► Responda:

Há apenas um leito disponível na UTI e dois pacientes em estado grave:

- [1]  **Criança de 8 anos** - alta chance de recuperação.
- [2]  **Mulher grávida de 32 anos** - sua vida não garante a do bebê.

Quem deve receber o leito?

Mas o que é ser justo?

► Responda:

Há apenas um leito disponível na UTI e dois pacientes em estado grave:

- [1] ♂ **Homem de 40 anos** - baixa probabilidade de sobrevivência.
- [2] ♀ **Mulher idosa de 75 anos** - probabilidade relativamente alta de sobrevivência.

Quem deve receber o leito?

Definição de justiça



Complexa!

- ▶ A noção de justiça **varia entre culturas e contextos**, refletindo valores e normas sociais.
- ▶ Em IA, é a busca por sistemas **equitativos**, livres de discriminação ou vieses injustos.

Equidade ≠ Igualdade

A equidade considera **diferenças individuais e estruturais**, garantindo que **todos tenham acesso às mesmas oportunidades**, mesmo que isso signifique **tratamentos diferenciados**.

Conceito bastante discutido, mas que também não é fácil

Construindo um sistema justo

"Viés algorítmico não é um problema meramente técnico, mas social e ético. Para mitigá-lo, devemos agir desde a coleta de dados até a governança do sistema."

— **Cathy O'Neil**, autora de *Weapons of Math Destruction*

1 Propósito

IA deve ter um **objetivo claro** e ser a melhor solução para o problema. Sem um propósito bem definido, pode gerar impactos negativos e desperdício de recursos.

Surto do Ebola (2014) - Pesquisadores usaram dados de mobilidade para prever surtos, mas o Ebola se espalha por contato direto. O foco deveria ter sido **redes de contato entre infectados**.



2 Dados

Dados enviesados reforçam desigualdades. **É necessário garantir a diversidade com qualidade.**

Auditórias e engajamento de especialistas são consideradas boas práticas!

Liang et al. (2023) - Detectores de GPT estão enviesados contra "escritores" não-nativos do inglês.



3 Abusabilidade

Os desenvolvedores de IA precisam antecipar **vulnerabilidades e cenários de uso indevido.**

Algoritmos podem ser **sequestrados e transformados em ferramentas para fins maliciosos**, como manipulação, vigilância e desinformação.

Poder 360 (24/01/2024) - Deep Fake do Papa Francisco usando casaco da moda e divulgado por grandes veículos de mídia como verdade. [\[Link\]](#)



4 Privacidade

Os sistemas de IA podem comprometer a **privacidade** dos usuários ao armazenar dados sensíveis, sujeitos a vazamentos e ataques.

Para mitigar riscos, é essencial aplicar **segurança desde o design** e garantir o **controle do usuário** sobre seus dados.

Época Negócios (16/01/2025) - Brasileiros vendem registro da íris por R\$600 para projeto Worldcoin de Sam Altam, CEO da OpenAI [\[Link\]](#)

5 Proxy

Algoritmos podem discriminar **indiretamente** ao usar variáveis correlacionadas a **atributos protegidos**, como raça ou gênero.

É de extrema importância a consulta a especialistas no assunto!

Ribeiro-Dantas et al. (2023) - O estado civil solteira foi indevidamente apontado como fator de risco para um tipo de câncer de mama. Os autores demonstram que a variável é um proxy de fatores socioeconômicos, como acesso ao diagnóstico médico adequado.

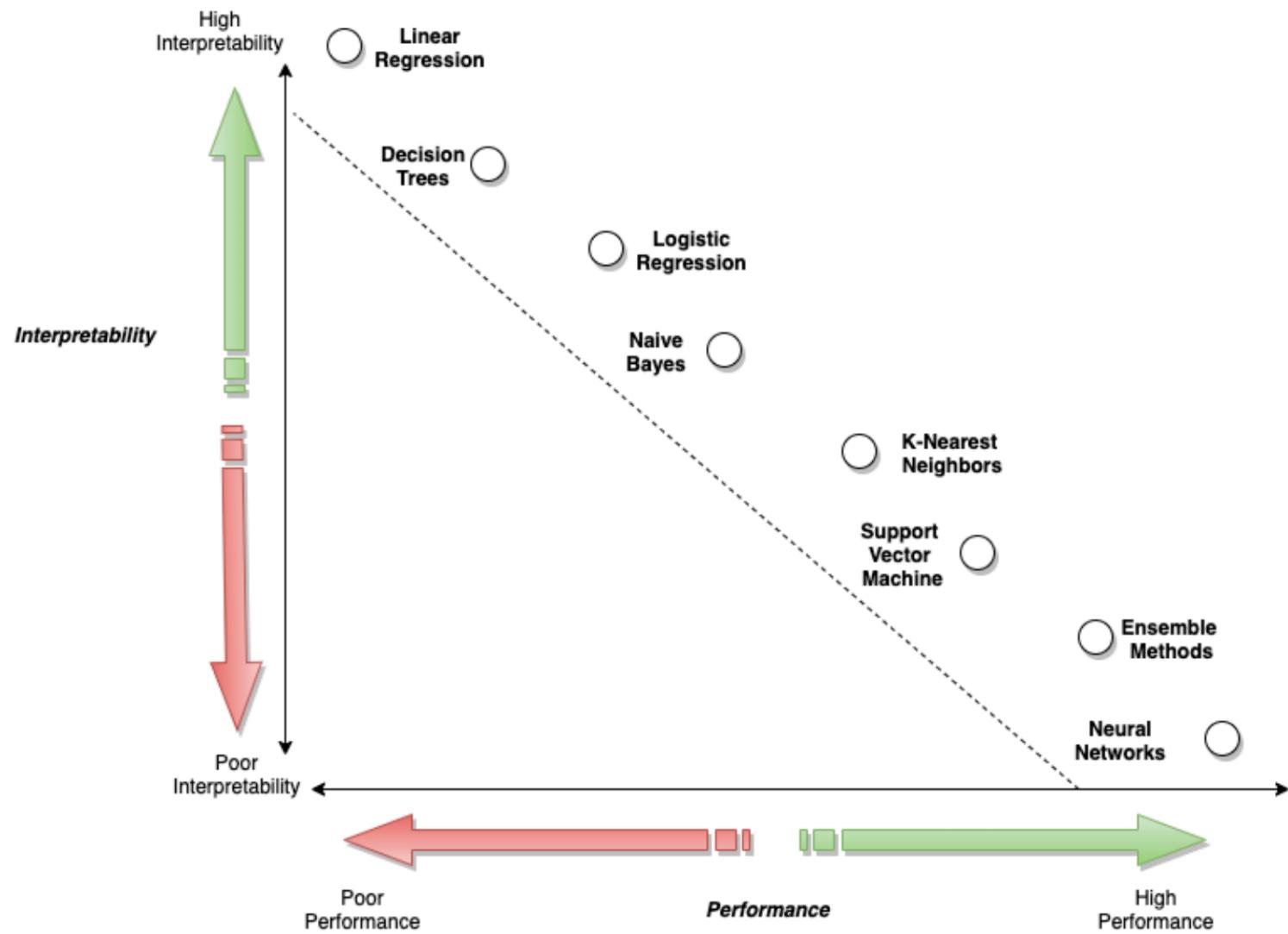


6 Explicabilidade

Quem projeta e implementa sistemas algorítmicos tem a **responsabilidade de explicar decisões críticas** que impactam o bem-estar das pessoas.

O usuário quer entender o motivo dos resultados, o motivo das falhas e quando e o quanto ele pode confiar na IA!



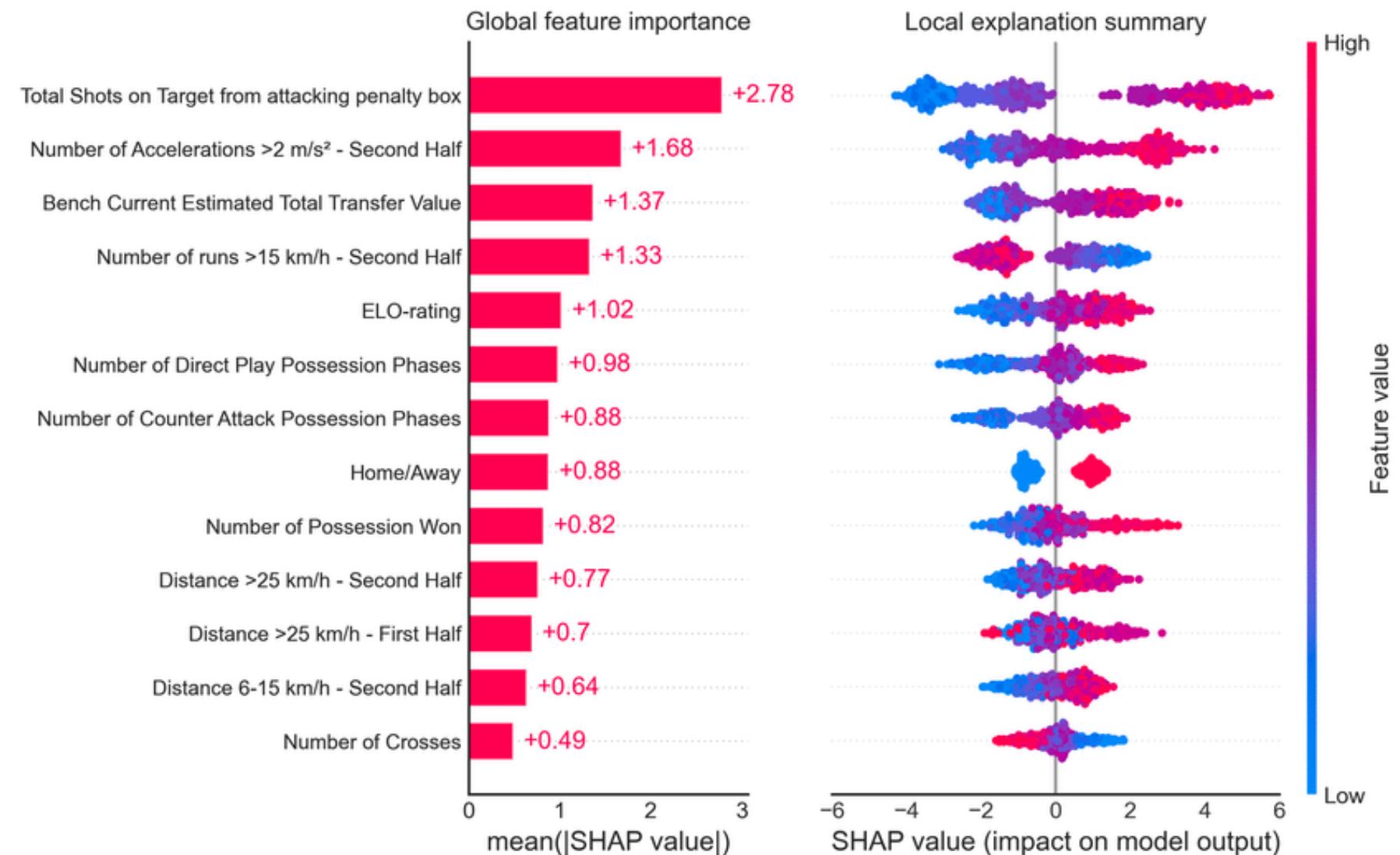


Interpretabilidade

Refere-se à capacidade de entender como um modelo de IA funciona e chega às suas conclusões.

Um modelo interpretável permite rastrear o impacto de cada entrada na saída.

[Link da figura](#)



Explicabilidade

Conjunto de processos e métodos que permite que usuários humanos **compreendam e confiem nos resultados e saídas criados por algoritmos de aprendizado de máquina.**

[Link da figura](#)

7 Otimização

Definir métricas de sucesso para IA envolve **compensações e impactos colaterais**. É essencial equilibrar **desempenho e equidade**, minimizando riscos para populações vulneráveis.

Shanklin et al. (2022) - O estudo mostra como algoritmos de IA podem perpetuar desigualdades raciais e propõe um método para equilibrar precisão e equidade, evitando discriminação sem comprometer a eficiência.



8 Generalização

Entre o desenvolvimento e a implementação de um sistema de IA, **o mundo – e os usuários – podem mudar**, tornando o contexto original inadequado e levando a falhas inesperadas.

Planos de retreino e descontinuidade são necessários!

Time (22/01/2010) - Câmeras da Nikon não reconheciam corretamente rostos asiáticos, exibindo a mensagem "Alguém piscou?" mesmo com os olhos abertos. [\[Leia mais\]](#)



9 Monitoramento

Decisões algorítmicas devem ser **passíveis de revisão**, garantindo que indivíduos possam **questioná-las e corrigi-las**.

Além disso, **monitoramento contínuo e transparência** são essenciais para evitar abusos e garantir responsabilidade no uso da IA.



Reflexões finais

A IA **não é neutra** – ela reflete as escolhas de quem a constrói e os dados que a alimentam!

Reflexões finais

**Viés algorítmico não é apenas
um problema técnico,
mas social e ético.**

A mitigação exige um olhar interdisciplinar.

Reflexões finais

Precisamos equilibrar **acurácia**
e equidade, garantindo que
sistemas de IA sejam justos e
transparentes.

Reflexões finais

A supervisão contínua e o direito à contestação são fundamentais para que a IA beneficie a sociedade sem reforçar desigualdades!

Referências

1. Cheong, M. et al. (2024). *Investigating Gender and Racial Biases in DALL-E Mini Images*. [Online]. Available at: <https://doi.org/10.1145/3649883> [Accessed 24 January 2025].
2. Choudhry, H. S. et al. (2023). *Perception of Race and Sex Diversity in Ophthalmology by Artificial Intelligence: A DALL E-2 Study*. [Online]. Available at: <https://doi.org/10.2147/OPTH.S427296> [Accessed 24 January 2025].
3. Currie, G. et al. (2024). *Gender and Ethnicity Bias of Text-to-Image Generative Artificial Intelligence in Medical Imaging, Part 2: Analysis of DALL-E 3*. [Online]. Available at: <https://doi.org/10.2967/jnmt.124.268359> [Accessed 24 January 2025].
4. Glickman, M. and Sharot, T. (2024). *How human–AI feedback loops alter human perceptual, emotional and social judgements*. [Online]. Available at: <https://www.nature.com/articles/s41562-024-02077-2> [Accessed 24 January 2025].
5. Leslie, D. et al. (2021). *Does “AI” stand for augmenting inequality in the era of covid-19 healthcare?* [Online]. Available at: <https://doi.org/10.1136/bmj.n304> [Accessed 24 January 2025].
6. Liang, W. et al. (2023). *GPT detectors are biased against non-native English writers*. [Online]. Available at: <https://doi.org/10.48550/arXiv.2304.02819> [Accessed 24 January 2025].

Referências

7. Luccioni, A. S. et al. (2023). *Stable Bias: Analyzing Societal Representations in Diffusion Models*. [Online]. Available at: <https://doi.org/10.48550/arXiv.2303.11408> [Accessed 22 January 2025].
8. Mandal, A., Leavy, S. and Little, S. (2024). *Generated Bias: Auditing Internal Bias Dynamics of Text-To-Image Generative Models*. [Online]. Available at: <https://doi.org/10.48550/arXiv.2410.07884> [Accessed 24 January 2025].
9. Ribeiro-Dantas, M. da C. et al. (2023). *Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients*. [Online]. Available at: <https://doi.org/10.48550/arXiv.2303.06423> [Accessed 24 January 2025].
10. Shanklin, R. et al. (2022). *Ethical Redress of Racial Inequities in AI: Lessons from Decoupling Machine Learning from Optimization in Medical Appointment Scheduling*. [Online]. Available at: <https://doi.org/10.1007/s13347-022-00590-8> [Accessed 24 January 2025].
11. Wu, Y., Nakashima, Y. and Garcia, N. (2024). *Gender Bias Evaluation in Text-to-image Generation: A Survey*. [Online]. Available at: <https://doi.org/10.48550/arXiv.2408.11358> [Accessed 24 January 2025].

Obrigada!

Marília Melo Favalesso - PhD, Cientista de Dados

 LinkedIn: /mariliafavalesso  Email: marilia.melo.favalesso@gmail.com