dhauz | uma empresa QuantumRise

Inteligência artificial ética:

Como detectar e corrigir o viés nos modelos de IA

A inteligência artificial (IA) e os modelos de machine learning (ML) estão assumindo, cada vez mais, a responsabilidade da tomada de decisões nas organizações. Sistemas automatizados já influenciam quem recebe crédito, quais currículos são priorizados em processos seletivos, como riscos são avaliados em sistemas supostamente imparciais e até quais pacientes recebem mais atenção em tratamentos de saúde.

À primeira vista, essas tecnologias costumam ser percebidas como mais objetivas, consistentes e eficientes do que decisões tomadas exclusivamente por humanos. Mas e se esses sistemas aprendessem vieses presentes nos dados do mundo real?

A resposta, infelizmente, é que eles já aprenderam, já que **modelos de machine learning aprendem a partir do mundo como ele é, e não necessariamente como ele deveria ser.**

Na prática, algoritmos refletem os padrões presentes nos dados utilizados em seu treinamento. Quando esses dados carregam desigualdades históricas, distorções de representação ou decisões humanas enviesadas, os modelos tendem a reproduzir ou amplificar tais assimetrias. Assim, a promessa de algoritmos imparciais pode rapidamente se transformar em risco de discriminação automatizada, escalável e difícil de detectar.

Esse não é um risco teórico. Temos diversos casos reais que demonstraram como sistemas de IA podem produzir impactos sociais significativos. Um exemplo conhecido é o estudo da ProPublica (Angwin, 2016) que analisou o algoritmo COMPAS, utilizado no sistema judicial dos Estados Unidos para estimar o risco de reincidência criminal.

A investigação revelou que o modelo atribui pontuações de risco mais altas a pessoas negras com maior frequência, mesmo quando o histórico real não justificava essa diferença. Como essas previsões influenciavam decisões judiciais, o viés algorítmico acabava contribuindo para sentenças mais severas em determinados grupos.

Esse tipo de problema surge dos dados, dos métodos e das escolhas humanas em cada etapa do ciclo de aprendizado.

O viés que salva e ao mesmo tempo exclui

Casos semelhantes surgem em contextos ainda mais sensíveis, como na **área da saúde**. A jornalista inglesa **Josephine Lethbridge cita no artigo** AI is sexist – can we change that? | Les Glorieuses **que o uso de modelos de aprendizagem de máquina** para priorização de pacientes em listas de transplante, de forma não intencional, passaram a excluir mulheres. O problema não estava no algoritmo em si, mas nos dados históricos que o alimentaram por décadas com estudos clínicos que tratavam o corpo masculino como padrão biológico.

Como consequência, o modelo aprendeu padrões que favoreciam pacientes homens, institucionalizando uma desigualdade em decisões literalmente relacionadas à vida e à morte. Esse tipo de exemplo deixa claro que **a IA não cria vieses, ela apenas reproduz, com eficiência matemática, o mundo que a ensinamos a enxergar.**

Um reflexo ampliado dos nossos vieses

Os efeitos desse fenômeno não se limitam a decisões críticas ou explícitas. Em sistemas de linguagem, recomendação e busca, o viés se manifesta de forma mais sutil, porém persistente. Modelos utilizados para triagem automática de currículos já demonstraram associar maior compatibilidade a nomes masculinos, enquanto sistemas de geração de texto tendem a reforçar estereótipos de gênero e profissão.

Em **aplicações de visão computacional**, a falta de representatividade nos dados de treinamento já levou a desempenhos significativamente inferiores na identificação de condições médicas em tons de pele mais escuros. Esses exemplos evidenciam que o viés não está apenas nos números ou nas tabelas, mas também nas narrativas, imagens e padrões culturais presentes nos dados.

É nesse contexto que surge o debate sobre ***unbiased models*** (modelos imparciais). Mais do que um problema estritamente técnico, a ética em sistemas de IA deve ser entendida como um desafio sociotécnico, no qual escolhas de dados, métricas, algoritmos, processos organizacionais e objetivos de negócio se entrelaçam. Um modelo pode apresentar excelente desempenho estatístico e, ainda assim, produzir impactos desiguais sobre diferentes grupos sociais. Esse descompasso leva a preocupações éticas, legais e reputacionais, especialmente em setores regulados ou de alto impacto social.

Ao mesmo tempo, é importante reconhecer que *unbiased* não possui uma definição única e universal. Diferentes aplicações exigem diferentes noções de ética, e frequentemente existem *trade-offs* inevitáveis entre desempenho, interpretabilidade e equidade. Portanto, falar em modelos imparciais significa **adotar processos consistentes, mensuráveis e transparentes para identificar, avaliar e mitigar vieses ao longo do ciclo de vida dos modelos.**

1. Entendendo o viés

De acordo com Ntoutsi et al. (2020), o viés pode surgir em três níveis do ciclo de vida de IA:

Etapa	Fonte de Viés	Exemplo
I) Coleta de dados	Amostras não representativas	Banco de imagens com 80% de rostos brancos
II) Aprendizado do modelo	Métricas que priorizam acurácia global	Classificador de crédito que acerta 95% mas erra sempre para um grupo minoritário
II) Decisão e uso	Interpretação e feedback humano enviesado	Usuários reforçam estereótipos ao corrigir resultados “errados” do modelo

I. Coleta de dados

Os dados constituem a principal matéria-prima dos modelos de ML. Quando conjuntos de dados são não representativos, incompletos ou refletem desigualdades históricas, o modelo tende a internalizar essas distorções. Exemplos incluem sub-representação de determinados grupos (por exemplo, mulheres ou minorias étnicas) ou diferenças sistemáticas na qualidade das medições entre grupos.
Além disso, mesmo atributos aparentemente neutros podem funcionar como proxies para características sensíveis, como raça ou gênero. Código postal, nível educacional ou histórico profissional, por exemplo, podem estar fortemente correlacionados com atributos protegidos.

II. Aprendizado do modelo

O viés no aprendizado pode ocorrer durante o treinamento, pois modelos de ML são frequentemente otimizados com métricas globais de desempenho (Ex: acurácia, AUC etc.). Tal escolha pode mascarar desempenhos desiguais entre grupos. Um classificador pode atingir uma acurácia média alta, enquanto erra sistematicamente em um grupo minoritário, produzindo um impacto desproporcional.

III. Decisão e uso

Mesmo quando dados e modelos são projetados cuidadosamente, o uso prático do sistema pode introduzir novos vieses. Interpretações humanas enviesadas, feedback seletivo e decisões baseadas em recomendações automatizadas podem reforçar estereótipos existentes, criando ciclos de retroalimentação (*feedback loops*).

Esses vieses se manifestam em duas formas:

- **Desbalanceamento de classes** – certos grupos aparecem menos no treinamento (ex: mulheres em cargos de liderança).
- **Correlação espúria** – atributos neutros (como CEP ou universidade) acabam servindo de proxies para características sensíveis (raça, gênero, idade).

Remover colunas sensíveis do dataset, por si só, **não resolve**. Como mostram Zliobaite & Custers (2016), o modelo continua aprendendo relações indiretas. O viés apenas muda de nome.

2. Detectando o viés

Não existe uma definição matemática de "imparcialidade". A literatura lista mais de 20 métricas diferentes (Verma & Rubin, 2018), que podem ser utilizadas no contexto de imparcialidade em algoritmos de IA. As mais comuns são:

Tipo de Métrica	O que mede	Exemplo prático
Demographic Parity	Proporção de resultados positivos deve ser igual entre grupos	Taxa de aprovação de crédito entre gêneros
Equalized Odds	Taxas de falsos positivos/negativos semelhantes entre grupos	Erro de rejeição igual para homens e mulheres
Equal Opportunity	Igualdade apenas na taxa de verdadeiros positivos	Oportunidade de aprovação igual quando o cliente é realmente bom
Predictive Parity	Probabilidade predita deve refletir realidade de cada grupo	Se o modelo dá 0.8 de chance de sucesso, isso deve ser válido para todos os grupos
Calibration	Coerência entre previsão e ocorrência real	Modelo calibrado não "superestima" um grupo

Essas métricas frequentemente entram em conflito entre si, o que implica que satisfazer uma definição de *unbiased* pode violar outra. Portanto, a escolha da métrica adequada depende fortemente do contexto da aplicação e dos impactos desejados.
Ferramentas como Fairlearn (Microsoft, 2020) e AIF360 (IBM) facilitam essa análise, permitindo comparar métricas de desempenho e imparcialidade entre grupos definidos por atributos sensíveis, por exemplo: sexo, religião, raça, etnia etc.

3. Estratégias para correção do viés: abordagens técnicas

Ntoutsi et al. (2020) e Bird et al. (2020) organizam as estratégias de mitigação do viés em algoritmo de IA em três fases:

A. Pré-processamento: corrigir os dados antes do treino

Os métodos de pré-processamento atuam diretamente nos dados, buscando reduzir desigualdades antes do treinamento. O foco é *balancear ou transformar os dados* para reduzir desigualdades:

Reamostragem balanceada: aumenta a presença de grupos sub-representados (por oversampling ou reweighting).

Label flipping: altera rótulos de instâncias próximas à fronteira de decisão para equalizar distribuições.

Discretização ética: reagrupa valores de atributos correlacionados com variáveis sensíveis.

B. Em-processamento: mitigar o viés durante o treinamento

Nesta etapa a ideia de imparcialidade é incorporada diretamente no treinamento do modelo (função de custo):

Regularização: adiciona penalidades para desigualdade entre grupos (Zafar et al., 2017).

Adversarial debiasing: usa uma rede adversária para remover informação sobre atributos sensíveis.

Fair constraints: redefine o problema de otimização para garantir igualdade de oportunidades.

*Ferramentas: **Fairlearn** oferece o algoritmo Exponentiated Gradient, que busca o melhor equilíbrio entre acurácia e igualdade de resultados.*

O resultado é uma família de modelos que exploram o trade-off entre desempenho e imparcialidade.

C. Pós-processamento: ajustar o resultado do modelo pronto

Se o modelo já foi treinado, ainda é possível aplicar ajustes às saídas do modelo:

Threshold optimization: aplica limiares diferentes de decisão para cada grupo.

Equalized odds postprocessing: ajusta previsões para equilibrar taxas de erro (Hardt et al., 2016).

4. O desafio: trade-off entre imparcialidade e performance (acurácia)

Como destacam Bird et al. (2020), priorizar equidade exige aceitar pequenas perdas de acurácia para reduzir desigualdade de impacto. Por isso, a ferramenta **Fairlearn Dashboard** permite visualizar o trade-off entre performance e imparcialidade, facilitando decisões éticas e transparentes. Toda intervenção tem custo, ***unbiased* (imparcialidade) não é gratuita**.

A chave está em **documentar as escolhas**:

- Quais métricas éticas foram priorizadas?
- Qual impacto social se espera reduzir?
- Quais grupos participaram da definição do "imparcial"?

Sem essas respostas, o risco é cair no que a literatura chama de "*fairness washing*", parecer justo, sem de fato mudar nada.

5. IA ética como prática contínua

Nenhum modelo é neutro. Alertam Atari et al. (2023): mesmo as IAs mais avançadas refletem a visão cultural de seus criadores.
Isso significa que o caminho para a imparcialidade algorítmica **não é apenas técnico**, mas também social: diversidade nas equipes, inclusão nos dados e transparência nas decisões.
Construir uma IA ética não é um ponto de chegada; é um processo contínuo de auditoria, reflexão e responsabilidade.

Conclusão

Detecção e mitigação de viés são boas práticas em ciência de dados, e **requisitos éticos para sistemas que influenciam vidas humanas**.
Um modelo que decide crédito, saúde ou segurança precisa ser **ético e imparcial, explicável e auditável**.

O desafio é **transformar tudo isso em cultura organizacional**, onde cada linha de código e cada conjunto de dados carregam a pergunta: "Para quem esse modelo funciona, e para quem ele ainda falha?"

Quer ver como aplicar uma IA ética na prática?

Neste link, você tem acesso ao notebook desenvolvido para a Mitigação viés de um exemplo prático. Explore, teste e conte para a gente como foi a sua experiência!

Referências

Ntoutsi, E. et al. (2020). *Bias in Data-driven Artificial Intelligence Systems: An Introductory Survey. WIREs Data Mining and Knowledge Discovery*.

Bird, S. et al. (2020). *Fairlearn: A Toolkit for Assessing and Improving Fairness in AI*. Microsoft Research.

Atari, M. et al. (2023). *Which Humans?* Harvard University.

Lethbridge, Josephine. *AI is sexist – can we change that? Six experts share their thoughts*. AI is sexist – can we change that? | Les Glorieuses. Disponível em 06/01/2026.

Angwin, J. et. al. (2016). *Machine Bias - There's software used across the country to predict future criminals. And it's biased against blacks*. Machine Bias — ProPublica

Zliobaite, I. e Custers, B. (2016). *Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models*. Artificial Intelligence and Law, 24 (2), 183-201.

Verma, S. e Rubin, J. S. (2018). *Fairness Definitions Explained*. IEEE/ACM International Workshop on Software Fairness (FairWare), pag 1-7. https://api.semanticscholar.org/CorpusID:49561627

dhauz | uma empresa QuantumRise