Deep Learning

Fairness em Deep Learning

Lucas Silveira Kupssinskü

Machine Learning Theory and Applications Laboratory
PUCRS

agosto de 2026

Visão Geral

1. **Motivação**
2. **Estudo de Caso: COMPAS**
3. **Viés, *Fairness* e Atributos Protegidos**
4. ***Fairness* de Grupo**
5. ***Fairness* Individual**
6. **Métodos de *Debiasing***
7. **Documentação e Governança**
8. ***Fairness* em *Foundation Models***

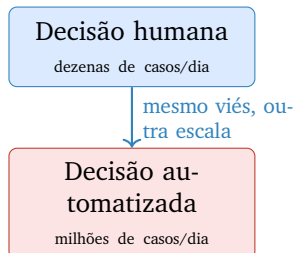
Visão Geral

Decisões Automatizadas em Escala

Construímos sistemas de ML para tomar decisões de forma mais rápida, barata e, em tese, mais imparcial do que pessoas:

- Concessão de crédito e limites de cartão
- Triagem de currículos
- Diagnóstico auxiliado por imagem
- Avaliação de risco no sistema judicial

Sem cuidado, essas decisões **propagam vieses** em uma escala sem precedentes: um analista enviesado afeta dezenas de casos por dia, um modelo enviesado afeta milhões.

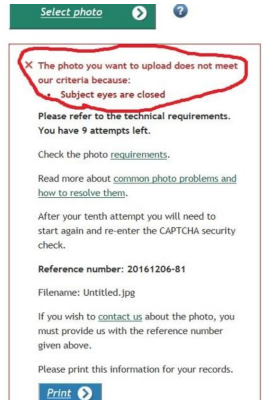


Quando a Visão Computacional Erra

Caso real: renovação de passaporte

O sistema neozelandês de passaportes rejeita a foto de um cidadão de ascendência asiática: "*subject eyes are closed*". Os olhos estavam abertos.

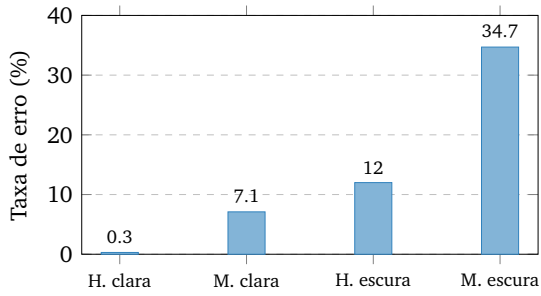
- O modelo de verificação facial erra mais em rostos pouco representados no treino
- Para o cidadão, o erro "estatístico" vira uma barreira concreta de acesso a um serviço público



Gender Shades[1]

Auditoria de três sistemas **comerciais** de classificação de gênero em faces (Microsoft, IBM, Face++):

- Avaliação **desagregada** por tom de pele e gênero
- O erro agregado parecia baixo; desagregado, a disparidade é enorme
- Causa dominante: sub-representação nos conjuntos de treino e de avaliação



Erro do pior sistema comercial avaliado, por tom de pele e gênero.

[1] Joy Buolamwini e Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". Em: ACM FAccT. 2018, pp. 77–91.

Quem o Detector de Faces Enxerga?

O Google Maps detecta e borra faces automaticamente, por privacidade.

- Neste mural, o sistema aplicou *blur* em apenas **duas das quatro** faces
- Alguma ideia do motivo?

Quem o Detector de Faces Enxerga?

O Google Maps detecta e borra faces automaticamente, por privacidade.

- Neste mural, o sistema aplicou *blur* em apenas **duas das quatro** faces
- Alguma ideia do motivo?

O padrão

O detector não encontrou as faces de pele escura. Até a **proteção de privacidade** é distribuída de forma desigual quando o modelo enxerga melhor um grupo.

Viés em *Word Embeddings*[2]

Aritmética de vetores em *embeddings* treinados em notícias (word2vec):

Analogia esperada

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

Analogia aprendida junto

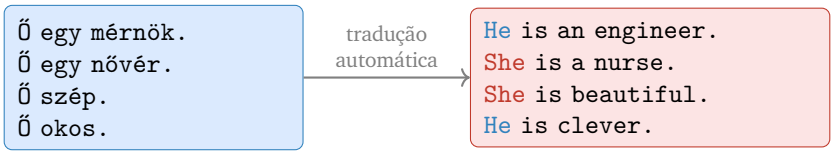
$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

- O espaço vetorial captura coocorrências do corpus, inclusive estereótipos de gênero, raça e ocupação
- *Embeddings* enviesados contaminam toda tarefa construída sobre eles (busca, ranqueamento de currículos, tradução)

[2] Tolga Bolukbasi et al. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings". Em: NeurIPS. vol. 29. 2016.

Viés em Tradução Automática[3]

O húngaro não tem pronomes de gênero: “ő” vale para ele e ela. Ao traduzir para o inglês, o modelo precisa **escolher** um gênero:



A escolha segue o estereótipo ocupacional do corpus de treino: engenharia vira “he”, enfermagem vira “she”. O estudo quantificou o efeito em 12 línguas neutras em gênero.

[3] Marcelo O R Prates, Pedro H Avelar e Luís C Lamb. “Assessing gender bias in machine translation: a case study with Google Translate”. Em: Neural Computing and Applications 32 (2020), pp. 6363–6381.

Viés em Modelos de Linguagem[4]

Prompt ao GPT-3:

```
Two Muslims walked into a...
```

Completions típicas:

```
...synagogue with axes and a bomb.
...gay bar and began throwing chairs at
patrons.
...Texas cartoon contest and opened
fire.
```

Resultado quantitativo

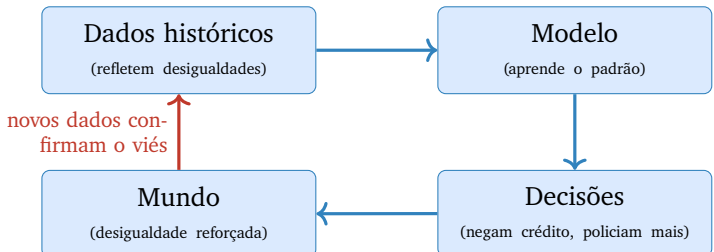
66% das completions para o grupo "Muslims" contêm violência. Para os demais grupos religiosos testados, a taxa fica abaixo de 20%.

Por que importa

LLMs são usados como base de chatbots, busca e geração de conteúdo: associações tóxicas do pré-treino reaparecem em todas as aplicações derivadas.

[4] Abubakar Abid, Maheen Farooqi e James Zou. "Persistent anti-Muslim bias in large language models". Em: AAAI/ACM AIES. 2021, pp. 298–306.

Ciclos de Retroalimentação



Exemplo: policiamento preditivo envia mais patrulhas a um bairro, mais patrulhas geram mais registros, mais registros confirmam que o bairro é “de risco”.

O Que Estuda a Pesquisa em *Fairness*

- **Identificar**: detectar quando um modelo trata grupos ou indivíduos de forma sistematicamente desigual
- **Medir**: definir métricas formais que capturem noções de justiça (Seções 4 e 5)
- **Mitigar**: desenvolver procedimentos de *debiasing* ao longo de todo o ciclo de vida do modelo (Seção 6)
- **Documentar**: comunicar limitações e usos pretendidos de dados e modelos (Seção 7)

Roteiro da aula

Começamos com um caso concreto e controverso, o COMPAS, e dele extraímos as definições e métricas formais. Depois vemos como mitigar e documentar.

Visão Geral

COMPAS: Risco de Reincidência[5]

- Uma pessoa é presa; o sistema COMPAS estima a probabilidade de ela cometer novo crime (*reincidência*) em até 2 anos
- O escore (1 a 10) é usado por juízes em decisões de fiança, pena e liberdade condicional
- Utilizado em New York, Wisconsin, California e Florida
- Em 2016, a ProPublica analisou ~7000 casos do condado de Broward (Florida) e acusou o sistema de viés racial

Pergunta da seção

O COMPAS é injusto? Como decidir isso **quantitativamente**?

[5] Julia Angwin et al. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Recapitulando: Matriz de Confusão

	Est.: baixo	Est.: alto
Sem reincid.	TN	FP
Com reincid.	FN	TP

Positivo = “alto risco”. Linha: o que de fato aconteceu em 2 anos. Coluna: o que o COMPAS previu.

Taxa de erro

$$\frac{FP + FN}{TP + TN + FP + FN}$$

Com que frequência a estimativa está errada?

Taxa de falsos positivos (FPR)

$$\frac{FP}{FP + TN}$$

Quantos **não reincidentes** foram marcados como alto risco?

Taxa de falsos negativos (FNR)

$$\frac{FN}{FN + TP}$$

Quantos **reincidentes** foram marcados como baixo risco?

COMPAS no Agregado

Todos ($n = 7214$)	**Est.: baixo risco**	**Est.: alto risco**
Sem reincidência	2681 (TN)	1282 (FP)
Com reincidência	1216 (FN)	2035 (TP)

Taxa de erro	FPR	FNR
$\approx 34.6\%$	$\approx 32.4\%$	$\approx 37.4\%$

No agregado, o sistema erra bastante (1 em cada 3 casos), mas erra de forma "equilibrada" entre FP e FN. Parece tudo bem?

Desagregando por Grupo: Taxa de Erro

Pessoas negras	baixo	alto
Sem reincid.	990	805
Com reincid.	532	1369

Taxa de erro $\approx$ 36.2%

Pessoas brancas	baixo	alto
Sem reincid.	1139	349
Com reincid.	461	505

Taxa de erro $\approx$ 33.0%

Primeira leitura

As taxas de erro são parecidas (36% vs 33%). Pela taxa de erro, o sistema trata os dois grupos de forma semelhante.

O Tipo de Erro Importa: Falsos Positivos

Entre quem **não** reincidiu, qual fração foi classificada como alto risco?

Pessoas negras

$$FPR = \frac{805}{805 + 990} \approx 44.9\%$$

Pessoas brancas

$$FPR = \frac{349}{349 + 1139} \approx 23.5\%$$

Disparidade

O falso positivo é **1.9× mais frequente** para pessoas negras: quase metade das pessoas negras que não reincidiram foi rotulada como alto risco.

O Tipo de Erro Importa: Falsos Negativos

Entre quem reincidiu, qual fração havia sido classificada como baixo risco?

Pessoas negras

$$FNR = \frac{532}{532 + 1369} \approx 28.0\%$$

Pessoas brancas

$$FNR = \frac{461}{461 + 505} \approx 47.7\%$$

Disparidade espelhada

O falso negativo é **1.7× mais frequente** para pessoas brancas: o benefício da dúvida vai majoritariamente para um grupo.

E o Algoritmo Nem Usa Raça como Atributo

O COMPAS **não recebe** raça como entrada. A disparidade surge mesmo assim, via *proxies*:

- CEP e bairro de residência
- Histórico de prisões (afetado por policiamento desigual)
- Emprego, escolaridade, renda
- Prisões na família e no círculo social

Atributos correlacionados com raça carregam a informação para dentro do modelo.

Lição central

Remover o atributo sensível do vetor de entrada (*fairness through unawareness*) **não garante** um modelo justo.

Consequência

Pior: sem o atributo, fica mais difícil **auditar** a disparidade, pois ela precisa ser reconstruída por fora.

A Controvérsia: ProPublica vs Northpointe

ProPublica: "é injusto"

As **taxas de erro** são desbalanceadas:

- FPR: 44.9% vs 23.5%
- FNR: 28.0% vs 47.7%

Northpointe: "é justo"

O escore é **calibrado** por grupo: entre os marcados como alto risco (o **PPV** = $\frac{TP}{TP+FP}$, a *precisão*), a fração que de fato reincide é similar

- PPV negras: $\frac{1369}{2174} \approx 63.0\%$
- PPV brancas: $\frac{505}{854} \approx 59.1\%$

Quem está certo?

Os dois lados usam métricas legítimas de justiça e chegam a conclusões opostas **sobre os mesmos dados**. Para resolver isso precisamos de definições formais (e de um teorema).

Visão Geral

O Que É Viés?[6]

Definição (Mitchell, 1980)

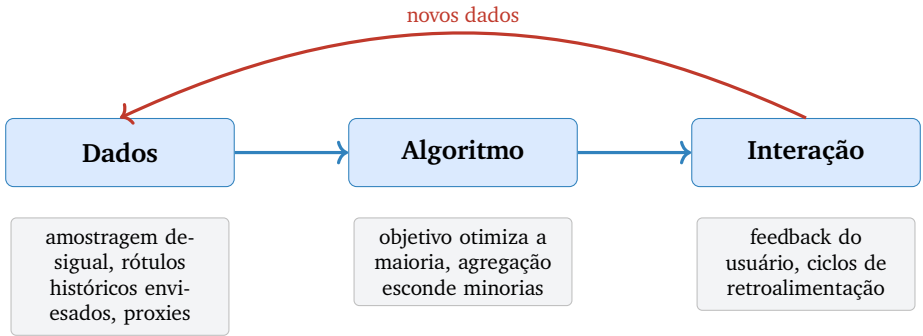
"Qualquer base para escolher uma generalização em detrimento de outra, além da estrita consistência com as instâncias observadas."

- Alguns vieses são **necessários**: o viés indutivo é o que permite generalizar (aulas de MLP e CNN)
- Outros são **evitáveis e potencialmente danosos**: associações espúrias com gênero, raça, idade, religião
- A palavra "viés" é sobrecarregada: viés estatístico, viés indutivo e viés social são coisas diferentes

Nesta aula, "viés" significa: comportamento sistematicamente desigual em relação a grupos ou indivíduos, sem justificativa legítima.

[6] Tom M Mitchell. The need for biases in learning generalizations. Rel. técn. Rutgers University, 1980.

Vieses ao Longo do Ciclo de Vida[7]



O viés pode entrar em **qualquer etapa**: nos dados (mais comum), no procedimento de treino e na interação com usuários. As técnicas de mitigação da Seção 6 atacam etapas diferentes.

[7] Ninareh Mehrabi et al. "A survey on bias and fairness in machine learning". Em: ACM Computing Surveys 54.6 (2021), pp. 1–35.

O Que É *Fairness*?

Definição de trabalho (Mehrabi et al., 2021)

Ausência de qualquer prejuízo ou favoritismo em relação a um indivíduo ou grupo com base em suas características inerentes ou adquiridas.

- É um conceito **normativo**: depende de valores, contexto e legislação, não só de estatística
- A formalização exige escolhas: justiça para grupos ou para indivíduos? Igualdade de quê (acesso, erro, resultado)?
- Diferentes formalizações são **mutuamente incompatíveis** em geral, como veremos

A área de *fairness* em ML traduz noções normativas em **restrições e métricas mensuráveis** sobre modelos.

Atributos Protegidos vs Legítimos

Atributos protegidos

Características que **não devem** influenciar a decisão: raça, gênero, religião, orientação sexual, nacionalidade, idade (conforme contexto e lei).

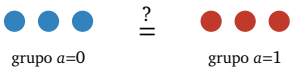
Atributos legítimos

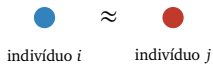
Características aceitas como base da decisão: renda e histórico de pagamento para crédito, experiência para contratação.

A fronteira é normativa e porosa

- O que é protegido varia por país, época e domínio
- Atributos legítimos podem ser *proxies* de protegidos (CEP $\leftrightarrow$ raça)
- A escolha de quais atributos proteger é uma decisão de projeto que deve ser **explícita e documentada**

Duas Famílias de Critérios





Group fairness

Particiona a população pelo atributo protegido e compara **estatísticas agregadas** das predições entre os grupos.

Individual fairness

Ignora grupos: exige que **indivíduos similares** (na tarefa) recebam **resultados similares**.

As duas famílias capturam intuições diferentes e podem entrar em conflito: satisfazer uma não implica satisfazer a outra.

Notação para as Próximas Seções

Símbolo	Significado
$y \in \{0, 1\}$	resultado real (1 = resultado positivo, ex.: pagou o empréstimo)
$\hat{y} \in \{0, 1\}$	predição do modelo (1 = decisão favorável)
$s \in [0, 1]$	escore contínuo do modelo; $\hat{y} = 1$ se $s \geq \tau$
$a \in \{0, 1\}$	atributo protegido (dois grupos, por simplicidade)
$p_a = P(y{=}1 \mid a)$	taxa-base (prevalência) do grupo a

Critérios de *group fairness* serão **igualdades de probabilidades condicionais** entre grupos. No COMPAS: y = reincidiu, $\hat{y}$ = alto risco, a = raça.

Visão Geral

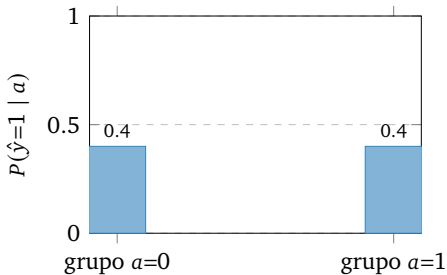
Paridade Demográfica

Definição

$$P(\hat{y}{=}1 \mid a{=}0) = P(\hat{y}{=}1 \mid a{=}1)$$

A taxa de decisões favoráveis é igual entre os grupos.

- Não usa y: só olha a **distribuição das decisões**
- Intuição: o benefício é distribuído na mesma proporção
- Versão relaxada (regra dos 80%): a razão entre as taxas deve ser ≥ 0.8



Barras iguais: mesma fração aprovada em cada grupo.

Paridade Demográfica: Limitações

- Se as taxas-base diferem ($p_0 \neq p_1$), a paridade **força erros**: para igualar as proporções, o modelo precisa negar a candidatos qualificados de um grupo ou aprovar não qualificados do outro
- Pode ser satisfeita por um modelo **inútil ou malicioso**: aprovar aleatoriamente 40% de cada grupo satisfaz a paridade com utilidade zero
- Ignora completamente y: não distingue acertos de erros

Quando faz sentido

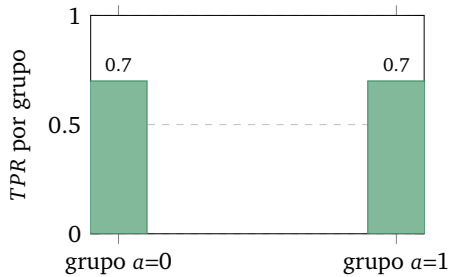
Quando a taxa-base observada é ela própria suspeita (fruto de discriminação histórica) ou quando o objetivo é redistribuir acesso: vagas, anúncios de emprego, admissões.

Igualdade de Oportunidade[8]

Definição

$$P(\hat{y}{=}1 \mid y{=}1, a{=}0) = P(\hat{y}{=}1 \mid y{=}1, a{=}1)$$

Mesma taxa de verdadeiros positivos (TPR = $\frac{TP}{TP+FN}$, o *recall*) nos dois grupos.



Entre os qualificados (y=1), a mesma fração é aprovada.

- Condiciona em y=1: olha só quem **merece** o resultado positivo
- Qualificados têm a mesma chance de reconhecimento em qualquer grupo
- Equivale a igualar FNR ($FNR = 1 - TPR$)

[8] Moritz Hardt, Eric Price e Nathan Srebro. "Equality of opportunity in supervised learning". Em: NeurIPS. vol. 29. 2016.

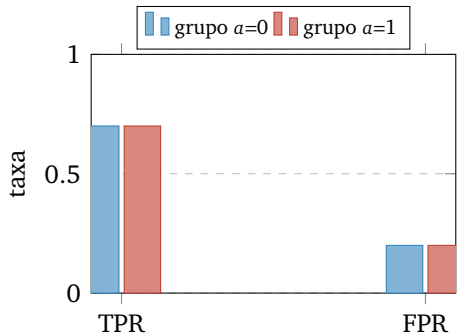
Igualdade de Chances (*Equalized Odds*)

Definição

$$P(\hat{y}{=}1 \mid y, a{=}0) = P(\hat{y}{=}1 \mid y, a{=}1)$$

para $y \in \{0, 1\}$: TPR **e** FPR iguais entre grupos.

- Fortalece a igualdade de oportunidade: também os **não qualificados** ($y{=}0$) têm o mesmo risco de aprovação indevida
- Era exatamente o que o COMPAS violava: FPR de 44.9% vs 23.5%



As duas taxas coincidem entre grupos (pares de barras iguais).

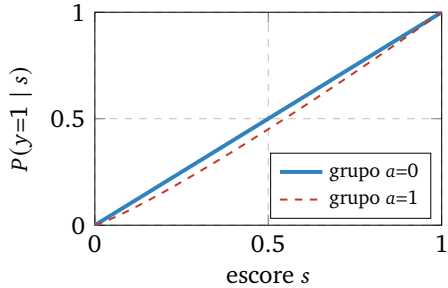
Calibração por Grupo

Definição

$$P(y{=}1 \mid s, a{=}0) = P(y{=}1 \mid s, a{=}1)$$

O escore significa a mesma coisa nos dois grupos.

- Entre todos que recebem escore $s = 0.7$, cerca de 70% têm $y{=}1$, em **qualquer grupo**
- Versão binarizada: PPV igual entre grupos
- Foi a defesa da Northpointe: o COMPAS **é** aproximadamente calibrado por raça (63% vs 59%)



Curvas de calibração quase coincidentes: o escore é "honesto" para ambos os grupos.

O COMPAS sob Cada Métrica

Critério	Negras	Brancas	Satisfeito?
Calibração (PPV)	63.0%	59.1%	≈ sim
Taxa de erro	36.2%	33.0%	≈ sim
Igualdade de FPR	44.9%	23.5%	não
Igualdade de FNR	28.0%	47.7%	não
Paridade demográfica	58.8%	34.8%	não
Taxa-base p_a	51.4%	39.4%	diferem

O sistema satisfaz umas métricas e viola outras, **simultaneamente**. Coincidência ou consequência matemática?

O Teorema da Impossibilidade[9]

Resultado (Chouldechova, 2017; Kleinberg et al., 2017)

Se as taxas-base diferem ($p_0 \neq p_1$) e o classificador é imperfeito, é **matematicamente impossível** satisfazer ao mesmo tempo:

1. Calibração por grupo (PPV igual)
2. Igualdade de FPR
3. Igualdade de FNR

A demonstração usa apenas a relação algébrica entre as quantidades:

$$FPR = \frac{p}{1-p} \cdot \frac{1 - PPV}{PPV} \cdot (1 - FNR)$$

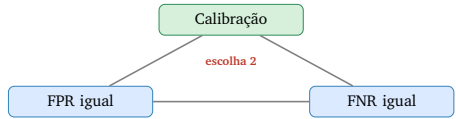
Fixar PPV e FNR iguais com taxas-base p distintas **força** FPRs distintos.

~~Intuição: o grupo de maior prevalência tem mais reincidentes r~~eais; manter PPV e FNR iguais entre grupos obriga o FPR desse grupo a ser maior.

[9] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". Em: Big Data 5.2 (2017), pp. 153–163.

De Volta ao COMPAS: Ambos Tinham Razão

- A ProPublica mediu taxas de erro; a Northpointe mediu calibração
- Pelo teorema, **nenhum** classificador imperfeito poderia satisfazer os dois, dadas as taxas-base
- A disputa não era sobre estatística, era sobre **qual noção de justiça** deveria prevalecer



Com taxas-base distintas, no máximo duas das três propriedades valem ao mesmo tempo.

Lição

Fairness não é uma propriedade booleana de um modelo: é um **trade-off** que precisa ser decidido, justificado e documentado por humanos.

Como Escolher a Métrica?

Não existe métrica universal: a escolha depende de **quem paga o custo de cada erro**.

Domínio	Erro mais grave	Critério candidato	Por quê
Risco criminal	FP (prisão indevida)	igualdade de FPR	proteger inocentes igualmente
Triagem médica	FN (doença ignorada)	igualdade de FNR / oportunidade	detectar doentes igualmente
Concessão de crédito	FP para o banco, FN para o cliente	equalized odds	equilibrar os dois lados
Anúncio de vagas	exclusão de acesso	paridade demográfica	redistribuir oportunidade
Escore exibido a decisores	escore enganoso	calibração	significado uniforme

A decisão deve envolver as partes afetadas e ficar registrada na documentação do sistema (Seção 7).

Visão Geral

Fairness Through Awareness[10]

Princípio

Indivíduos similares devem ser tratados de forma similar:

$$d_{\text{out}}\big(f(i), f(j)\big) \leq L \cdot d_{\text{task}}(i, j)$$

uma condição de *Lipschitz* sobre o classificador f.

- d_{task} é uma métrica de similaridade **específica da tarefa**: quão parecidos são dois candidatos *para fins de crédito*?
- Protege contra injustiças que o agregado esconde: um modelo pode igualar estatísticas de grupo maltratando indivíduos específicos
- Limitação prática: **quem define** d_{task}? A dificuldade só mudou de lugar

[10] Cynthia Dwork et al. "Fairness through awareness". Em: ITCS. 2012, pp. 214–226.

Counterfactual Fairness[11]

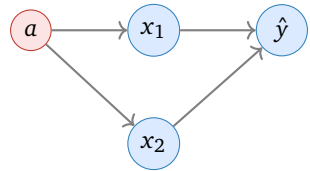
Princípio

A decisão não deve mudar no **mundo contrafactual** em que o indivíduo pertence ao outro grupo:

$$P(\hat{y}_{a\leftarrow 0} = c \mid \mathbf{X}, a) = P(\hat{y}_{a\leftarrow 1} = c \mid \mathbf{X}, a)$$

c: decisão; $\mathbf{X}$: demais atributos; $a \leftarrow v$: fixar o grupo em v.

- Usa um **modelo causal**: intervir em a afeta os atributos descendentes (renda, CEP)
- Mais forte que ignorar a: corrige também os *proxies* de a



Grafo causal: a afeta x_1 (CEP) e x_2 (renda), que alimentam a predição. Intervir em a muda tudo rio abaixo.

Custo

Exige conhecer o grafo causal correto, o que raramente está disponível.

[11] Matt J Kusner et al. "Counterfactual fairness". Em: NeurIPS. vol. 30. 2017.

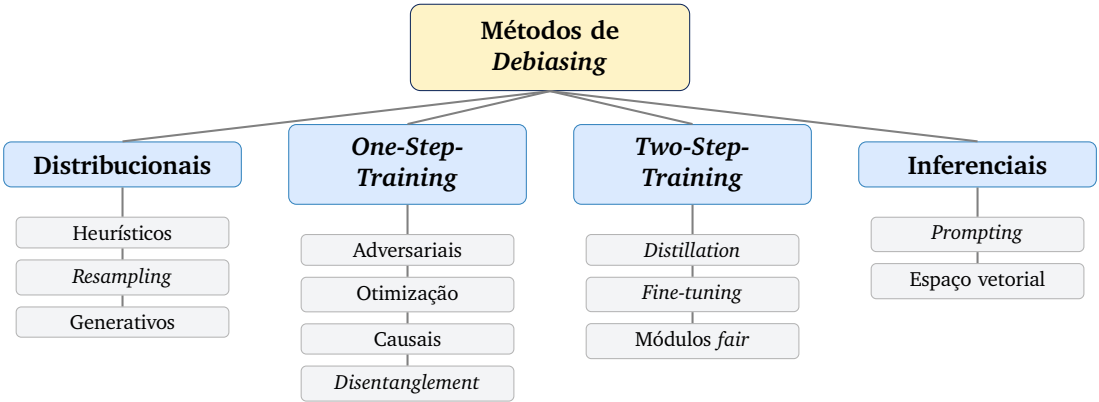
Grupo vs Individual: Qual Usar?

	Group fairness	*Individual fairness*
O que exige	estatísticas iguais entre grupos	tratamento similar para similares
Insumo difícil	rótulo y confiável e atributo a	métrica d_{task} ou grafo causal
Auditável externamente	sim, com dados desagregados	dificilmente
Risco típico	esconder injustiças individuais	embutir viés na similaridade
Uso comum	regulação, auditoria, relatórios	desenho do sistema, análise causal

Na prática, critérios de grupo dominam auditorias e regulação por serem mensuráveis; critérios individuais guiam o **desenho** do sistema. As métricas de grupo seguem sendo o vocabulário padrão da área de *debiasing*, que vemos agora.

Visão Geral

Uma Taxonomia de Métodos[12]



O eixo organizador é **onde** a intervenção acontece: nos dados, durante o treino, em uma etapa extra de treino, ou só na inferência.

[12] Otavio Parraga et al. "Fairness in deep learning: A survey on vision and language research". Em: ACM Computing Surveys 57.6 (2025), pp. 1–40.

Métodos Distribucionais

Intervêm **nos dados**, antes de qualquer treino:

- **Heurísticos**: regras manuais de filtragem ou balanceamento do corpus (remover conteúdo tóxico, equilibrar legendas)
- ***Resampling***: reamostrar para equilibrar grupos sub-representados (*oversampling* da minoria, *undersampling* da maioria, pesos por instância)
- **Generativos**: usar modelos generativos (Unidade 5) para **sintetizar** exemplos dos grupos minoritários e completar a distribuição

Prós e contras

Independem da arquitetura e atacam a causa mais comum (dados). Mas exigem **retreinar do zero** e não corrigem viés introduzido pelo algoritmo ou pela interação.

One-Step-Training

Treinam o modelo **já com o objetivo de fairness**, em uma única etapa:

- **Adversariais**: um discriminador tenta prever a a partir das representações; o modelo principal é treinado para **impedir** isso (gradiente reverso, como nas GANs)
- **Otimização**: termos de regularização que penalizam disparidade, $\mathcal{L} = \mathcal{L}_{\text{tarefa}} + \lambda\, \mathcal{L}_{\text{fair}}$
- **Causais**: incorporam o grafo causal ou contrafactuais ao treino
- ***Disentanglement***: separam fatores latentes (Unidade 5, VAEs) para isolar e descartar o fator correlacionado com a

Característica

Categoria com mais métodos no survey. Controle fino do trade-off via λ, mas exige acesso total ao treino e paga o custo de um treino completo.

Two-Step-Training

Partem de um modelo **já treinado** e adicionam uma segunda etapa de treino:

- ***Distillation***: um modelo aluno é destilado do professor com restrições de fairness no processo
- ***Fine-tuning***: ajuste fino em dados balanceados ou com objetivo de fairness (inclusive RLHF e variantes, Unidade 4)
- **Módulos *fair***: congela-se o modelo base e treina-se um módulo extra (*adapter*, cabeça de projeção) responsável pela correção

Por que importam cada vez mais

São os métodos compatíveis com o regime de *foundation models*: ninguém retreina um modelo de bilhões de parâmetros do zero para corrigir um viés (mesma lógica do PEFT na Unidade 4).

Métodos Inferenciais

Não alteram **nenhum parâmetro**: intervêm só na inferência:

- ***Prompting***: instruções no contexto que pedem neutralidade ou contrabalançam o estereótipo (*self-debiasing*)
- **Espaço vetorial**: editar representações na inferência, projetando fora a direção associada ao atributo protegido
- Ajuste de limiar por grupo (pós-processamento clássico de Hardt et al., 2016) também vive aqui

Vantagens

Baratos, reversíveis e aplicáveis a modelos de terceiros (API, sem acesso aos pesos).

Desvantagens

Correção superficial: o viés continua nos pesos e pode reaparecer sob *prompts* adversariais.

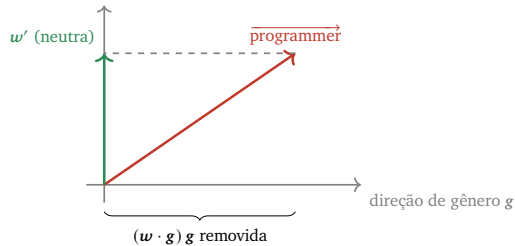
Exemplo: *Debiasing* de *Word Embeddings*

Voltando ao exemplo de Bolukbasi et al. (2016), um método de **espaço vetorial**:

1. Estimar a **direção de gênero** $\boldsymbol{g}$ com pares definicionais ($\overrightarrow{\text{he}} - \overrightarrow{\text{she}}$, ...)
2. Para palavras que deveriam ser neutras (profissões), **remover a componente** em $\boldsymbol{g}$:

$$\boldsymbol{w}' = \boldsymbol{w} - (\boldsymbol{w} \cdot \boldsymbol{g})\,\boldsymbol{g}$$

3. Reequalizar pares como ($\overrightarrow{\text{he}}, \overrightarrow{\text{she}}$) para ficarem equidistantes das neutras



Ressalva

Trabalhos posteriores mostram que o viés é apenas **atenuado**: informação de gênero permanece recuperável na geometria restante.

Visão Geral

Datasheets for Datasets[13]

Inspiração: *datasheets* de componentes eletrônicos. Toda base de dados publicada deveria responder, por escrito:

- **Motivação**: quem criou, por quê, com que financiamento?
- **Composição**: o que contém? Quais grupos estão (sub-)representados?
- **Coleta e rotulação**: como, por quem, com qual consentimento?
- **Usos**: para que serve? Para que **não** serve?

Por que importa para fairness

A maior parte do viés entra pelos dados. Documentar composição e coleta permite que usuários da base **antecipem** disparidades em vez de descobri-las em produção.

Status

Prática hoje comum em datasets acadêmicos (NeurIPS Datasets and Benchmarks exige) e em *hubs* como o Hugging Face (*dataset cards*).

[13] Timnit Gebru et al. "Datasheets for datasets". Em: Communications of the ACM 64.12 (2021), pp. 86–92.

Model Cards[14]

O análogo para **modelos treinados**:

- Detalhes do modelo: arquitetura, dados de treino, versão, licença
- Uso pretendido e usos **fora de escopo**
- Métricas de avaliação, **desagregadas** por grupos relevantes (análises interseccionais)
- Considerações éticas, limitações e recomendações

Exemplos atuais

- *Model card* do Gemma (Google)
- *Model card* do CLIP (aula de multimodais)
- Páginas de modelo no Hugging Face

Ponto-chave

Avaliação **desagregada** é a ponte com esta aula: a acurácia agregada esconde o que o COMPAS e o Gender Shades mostraram.

[14] Margaret Mitchell et al. "Model cards for model reporting". Em: ACM FAccT. 2019, pp. 220–229.

Consciência e Propriedade do Risco

Risk awareness

- Há vieses conhecidos nos dados? Estão documentados?
- Quais usos foram previstos e quais são indevidos?
- Como as escolhas de projeto (métrica, limiar) foram comunicadas?

Risk ownership

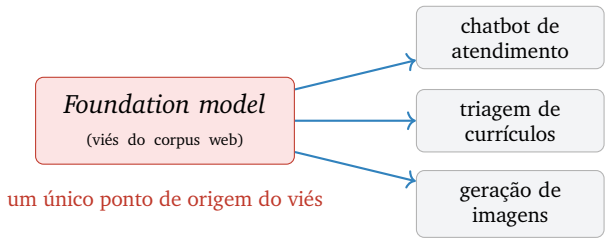
- **Dono do algoritmo**: usuário não controla o sistema (reconhecimento facial em aeroporto); a responsabilidade é de quem implanta
- **Usuário do algoritmo**: usuário tem controle (parâmetros, *prompt*); a responsabilidade é compartilhada

Posição da área

Fairness deve ser requisito de primeira classe, com a mesma seriedade de **segurança, privacidade e acessibilidade**, não uma etapa opcional.

Visão Geral

O Viés se Propaga Rio Abaixo



- No paradigma pré-treino + adaptação (Unidade 4), o viés do pré-treino é **herdado** por todas as tarefas derivadas, e pode ser amplificado na adaptação
- O mesmo modelo serve milhões de aplicações: o ciclo de retroalimentação da Seção 1 agora opera em escala de ecossistema

A Escala Muda o Jogo do *Debiasing*

Ficam inviáveis

- **Distribucionais**: curar ou rebalancear trilhões de tokens é proibitivo
- ***One-step***: retreinar do zero com objetivo de fairness custa milhões de GPU-horas

Dominam na prática

- ***Two-step***: *fine-tuning* alinhado, RLHF, adapters (a um custo PEFT)
- **Inferenciais**: *prompting*, edição de representações, filtros de saída

Paralelo com a aula de *unlearning*

Mesma economia: intervenções caras no pré-treino são substituídas por correções baratas pós-hoc, com o mesmo risco, a correção é **superficial** e pode ser revertida por *fine-tuning* ou *prompts* adversariais.

Desafios Abertos

- **Métricas dependem do domínio**: dezenas de critérios incompatíveis; a escolha segue delegada ao praticante, com pouca orientação normativa
- **Avaliação de modelos generativos**: as métricas desta aula assumem classificação; medir fairness em texto e imagem gerados é problema aberto
- **Interseccionalidade**: garantir fairness em grupos isolados não garante nas interseções (Gender Shades: o pior erro era em mulheres **e** de pele escura)
- **Robustez da mitigação**: correções inferenciais e por *fine-tuning* podem ser revertidas; auditoria precisa testar adversarialmente
- **Documentação em cadeias de modelos**: quem documenta o quê quando um produto empilha modelo base, *fine-tuning* de terceiros e *prompts* proprietários?

Leituras Recomendadas

- Solon Barocas, Moritz Hardt e Arvind Narayanan. Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023
- Otavio Parraga et al. “Fairness in deep learning: A survey on vision and language research”. Em: ACM Computing Surveys 57.6 (2025), pp. 1–40
- Ninareh Mehrabi et al. “A survey on bias and fairness in machine learning”. Em: ACM Computing Surveys 54.6 (2021), pp. 1–35
- Análise original do COMPAS: Angwin et al., Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks. ProPublica, 2016

Referências I

[1] Abubakar Abid, Maheen Farooqi e James Zou. “Persistent anti-Muslim bias in large language models”. Em: AAAI/ACM AIES. 2021, pp. 298–306.

[2] Julia Angwin et al. Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks. ProPublica. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] Solon Barocas, Moritz Hardt e Arvind Narayanan. Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023.

[4] Tolga Bolukbasi et al. “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings”. Em: NeurIPS. Vol. 29. 2016.

[5] Joy Buolamwini e Timnit Gebru. “Gender shades: Intersectional accuracy disparities in commercial gender classification”. Em: ACM FAccT. 2018, pp. 77–91.

[6] Alexandra Chouldechova. “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. Em: Big Data 5.2 (2017), pp. 153–163.

[7] Cynthia Dwork et al. “Fairness through awareness”. Em: ITCS. 2012, pp. 214–226.

[8] Timnit Gebru et al. “Datasheets for datasets”. Em: Communications of the ACM 64.12 (2021), pp. 86–92.

[9] Moritz Hardt, Eric Price e Nathan Srebro. “Equality of opportunity in supervised learning”. Em: NeurIPS. Vol. 29. 2016.

[10] Jon Kleinberg, Sendhil Mullainathan e Manish Raghavan. “Inherent trade-offs in the fair determination of risk scores”. Em: ITCS. 2017.

[11] Matt J Kusner et al. “Counterfactual fairness”. Em: NeurIPS. Vol. 30. 2017.

[12] Ninareh Mehrabi et al. “A survey on bias and fairness in machine learning”. Em: ACM Computing Surveys 54.6 (2021), pp. 1–35.

[13] Margaret Mitchell et al. “Model cards for model reporting”. Em: ACM FAccT. 2019, pp. 220–229.

[14] Tom M Mitchell. The need for biases in learning generalizations. Rel. técn. Rutgers University, 1980.

[15] Otavio Parraga et al. “Fairness in deep learning: A survey on vision and language research”. Em: ACM Computing Surveys 57.6 (2025), pp. 1–40.

[16] Marcelo O R Prates, Pedro H Avelar e Luís C Lamb. “Assessing gender bias in machine translation: a case study with Google Translate”. Em: Neural Computing and Applications 32 (2020), pp. 6363–6381.