Submission date: 07/12/2021
Resubmission date: 09/05/2022
Camera ready submission date: 18/10/2022

1st round notification: 07/04/2022
2nd round notification: 27/06/2022
Available online: 21/10/2022
Publication date: 30/12/2022

Section: regular article

Mitigando Vieses no Aprendizado de Máquina: Uma Análise Sociotécnica

Mitigating Bias in Machine Learning: A Socio-technical Analysis

Lívia Ruback[1], **Denise Carvalho**[2], **Sandra Avila**[3]

[1]Departamento de Computação – Universidade Federal Rural do Rio de Janeiro (UFRRJ), Seropédica, RJ – Brasil

[2]Programa de Pós-graduação em Estudos da Mídia (PPgEM) – Universidade Federal do Rio Grande do Norte (UFRN), Natal, RN – Brasil

[3]Instituto de Computação – Universidade Estadual de Campinas (Unicamp) Campinas, SP – Brasil

{liviaruback,denisecarvalho.mail}@gmail.com, sandra@ic.unicamp.br

Abstract. *This work presents a socio-technical analysis of biases included in the machine learning process. We describe in this work four types of biases: historical bias, data bias, model bias, and human interpretation bias, and how they can occur during the learning process, together with their social and cultural implications. We also bring strategies to mitigate those biases, including computation solutions, such as balancing the data used for the training and alternative metrics for the model evaluation, non-computational solutions, regulatory efforts, and initiatives to promote diversity in the tech industry and academy.*

Keywords. *Bias; Machine Learning; Algorithm bias*

Resumo. *Este artigo apresenta uma análise sociotécnica sobre os vieses inseridos durante o aprendizado de máquina. Descrevemos aqui quatro tipos de vieses: vieses históricos, vieses nos dados, vieses no modelo e vieses de interpretação humana; apontamos como eles podem ser inseridos nos modelos durante o processo de aprendizado e suas implicações sociais e culturais. Apontamos também as direções para mitigar estes vieses, que incluem soluções computacionais, como o balanceamento das bases de dados utilizadas para o treinamento dos modelos e das métricas alternativas para avaliar estes modelos, até soluções não computacionais, regulação do uso dos modelos e políticas para promover a diversidade na tecnologia e na academia.*

Palavras-Chave. *Vieses; Aprendizado de máquina; Discriminação algorítmica*

Cite as: Ruback, L., Carvalho, D. & Avila, S. (2022). Mitigating Bias in Machine Learning: A Socio-technical Analysis (Mitigando Vieses no Aprendizado de Máquina: Uma Análise Sociotécnica). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 15(1), 23:1-23:31. DOI: 10.5753/isys.2022.2396

1. Introdução

As tecnologias que usam algoritmos de aprendizado de máquina tem feito cada vez mais parte das nossas vidas, seja quando escolhemos um filme para assistir em um catálogo online, quando nos guiamos pelo trajeto sugerido por um sistema que recomenda rotas ou quando fazemos uma compra online. Muitas das decisões que são tomadas por seres humanos estão sendo cada vez mais automatizadas: decidir quem será contratado para uma vaga de emprego — ou demitido [Tavares 2021], decidir se ao cliente será concedido um empréstimo [Vilarino e Vicente 2020], identificar suspeitos [Castelvecchi 2020], determinar por quanto tempo um criminoso permanecerá na prisão [Maybin 2016], entre outros.

Estes sistemas são treinados com algoritmos de aprendizado de máquina que aprendem a partir de um grande volume de dados e geram modelos. Tais modelos são capazes de predizer a classe de novos exemplos não rotulados, a partir do que o seu algoritmo foi capaz de aprender com o conjunto de dados utilizado para o treinamento. Ao projetar tais sistemas, acredita-se na existência de modelos matemáticos perfeitos e imparciais e que a objetividade da matemática e da estatística seriam capazes de abarcar todas as realidades [O'Neil 2020]. Mas não é o que temos visto na prática nos últimos anos [Broussard 2018].

Muitos destes sistemas têm apresentado falhas e vieses, que resultam em casos de discriminação e preconceito, favorecendo determinados grupos e prejudicando outros. Vemos casos de discriminação algorítmica em sistemas de recrutamento que favorecem um único perfil de profissionais [Raghavan et al. 2020], serviços de tradução que reforçam preconceitos de gênero [Tomalin et al. 2021] e sistemas que auxiliam juízes a predizer as chances de reincidência de pessoas que cometeram crimes, desfavorecendo certos grupos étnicos [Washington 2018]. Nos sistemas de visão computacional, como aplicativos de rotulagem automática de fotos e sistemas de detecção e de reconhecimento facial, temos acompanhado casos de maiores taxas de falhas ao identificar grupos sub-representados — como o de mulheres e pessoas negras [Buolamwini e Gebru 2018; Raji et al. 2020].

Neste artigo, fazemos uma análise sociotécnica dos vieses inseridos no processo de aprendizado de máquina. Este trabalho pode servir como um guia sobre os principais vieses no aprendizado de máquina e traz as seguintes contribuições:

1. Descrevemos em detalhes — a partir de exemplos — todas as etapas do processo de aprendizado de máquina, para contextualizar os vieses que podem ser inseridos em cada etapa do processo.
2. Fazemos uma análise sociotécnica das principais implicações (morais, sociais e éticas) de diferentes vieses na sociedade.
3. Sumarizamos algumas direções para mitigar estes vieses, que passam por soluções computacionais (para lidar com vieses nos dados e no modelo) e por soluções não computacionais (para lidar com os vieses históricos e de interpretação humana).

Os estudos sociotécnicos fazem parte de um campo interdisciplinar sobre tecnologias digitais que possibilita a implementação de uma abordagem analítica qualitativa amparada em teorias e métodos da Sociologia para a análise de tecnologias digitais [Henriques 2018]. Neste trabalho, fazemos uso da análise sociotécnica por consi-

derarmos esta dimensão essencial para a compreensão do processo de tomada de decisões e do modo pelo qual o fator humano influencia no uso de tecnologias e métodos computacionais, como os utilizados no aprendizado de máquina, que abordamos aqui [Petukhov e Steshina 2018]

O que trazemos neste trabalho concilia de uma abordagem fundada em métodos computacionais e também na Sociologia. Tal metodologia permite que, diante da multiplicidade de fenômenos que ocorrem, seja possível a compreensão sobre questões específicas de uma forma ampla. Assim, questões que poderiam ser observadas como problemáticas de ordem individual, quando vistas sob a ótica sociológica, são compreendidas como questões de ordem coletiva e, consequentemente, essenciais para o convívio social em um nível macro de análise [Giddens 2008]. A observação dos processos sociais e culturais presentes nos diversos fenômenos que compõem a sociedade permitem a percepção das dinâmicas que estruturam a vida social e dos impactos culturais destas estruturas da vida social contemporânea no âmbito das tecnologias digitais [Nascimento 2020].

Neste trabalho, nós referenciamos os principais autores que fornecem definições e métricas para avaliar justiça neste contexto. [Bellamy et al. 2018; Verma e Rubin 2018; Suresh e Guttag 2019; Caton e Haas 2020; Mehrabi et al. 2021]. Muitas destas abordagens propõem novas formas matemáticas de se representar vieses, justiça e/ou discriminação.

O artigo está organizado da seguinte forma. Na Seção 2, trazemos uma visão geral do aprendizado de máquina, descrevendo as suas etapas, com exemplos. Na Seção 3, apresentamos quatro dos potenciais vieses que podem ser inseridos durante estas etapas. Na Seção 4, apontamos algumas direções para mitigar tais vieses. Finalmente, na Seção 5, concluímos e apresentamos os trabalhos futuros.

2. Etapas do Aprendizado de Máquina

O aprendizado de máquina surgiu como um subcampo da Inteligência Artificial que projeta algoritmos que aprendem a partir de grandes quantidades de exemplos — ou dados — relacionados a um determinado fenômeno [Mitchell 1997]. Hoje, o aprendizado de máquina é utilizado em uma infinidade de sistemas, desde sistemas de recomendação, sistemas de recrutamento, sistemas de tradução automática, de liberação de créditos, sistemas prisionais e em serviços de vigilância pública por reconhecimento facial. Nesta seção, apresentamos alguns conceitos importantes relacionados ao aprendizado de máquina e detalhamos as suas etapas.

O aprendizado de máquina, ou aprendizagem de máquina, é classificado em tipos, de acordo com os objetivos dos algoritmos. O mais frequentemente usado na prática é o aprendizado *supervisionado*, realizado a partir de um conjunto de dados (*dataset*, em Inglês) contendo exemplos *rotulados*. O principal objetivo do aprendizado supervisionado é aprender a partir de dados treinados e rotulados (ou seja, respostas corretas) de forma a ser capaz de gerar um *modelo* que faça novas predições, para dados não rotulados (não treinados) [Raschka 2015]. Os demais aprendizados — *não supervisionado* e *por reforço* — não utilizam dados rotulados e não serão abordados neste trabalho. É importante

mencionar que, neste trabalho, quando nos referimos ao aprendizado de máquina, nos referimos ao aprendizado de máquina supervisionado.

O aprendizado de máquina supervisionado — baseado em rótulos de exemplos aprendidos e capaz de prever rótulos para exemplos futuros — é subdividido em duas categorias: *classificação* e *regressão*. Ambas as tarefas aprendem a partir de dados rotulados, porém, para tarefas de classificação, como tarefas de classificação de e-mails como spam, o rótulo predito é um valor categórico, como sim ou não. Já para as tarefas de regressão, o valor predito é um valor contínuo, como tarefas de previsão do tempo ou de previsão de custos [Raschka 2015].

A Figura 1 mostra as etapas do processo de aprendizado de máquina supervisionado. O processo apresentado representa uma nova versão de um *framework* apresentado em um trabalho anterior [Ruback et al. 2021]. O processo pode ser visto como um *pipeline*, onde a saída de uma etapa representa a entrada da etapa subsequente: se inicia pela coleta de dados, passa pelo pré-processamento, pela criação do modelo e termina no pós-processamento. Estas etapas são detalhadas a seguir.

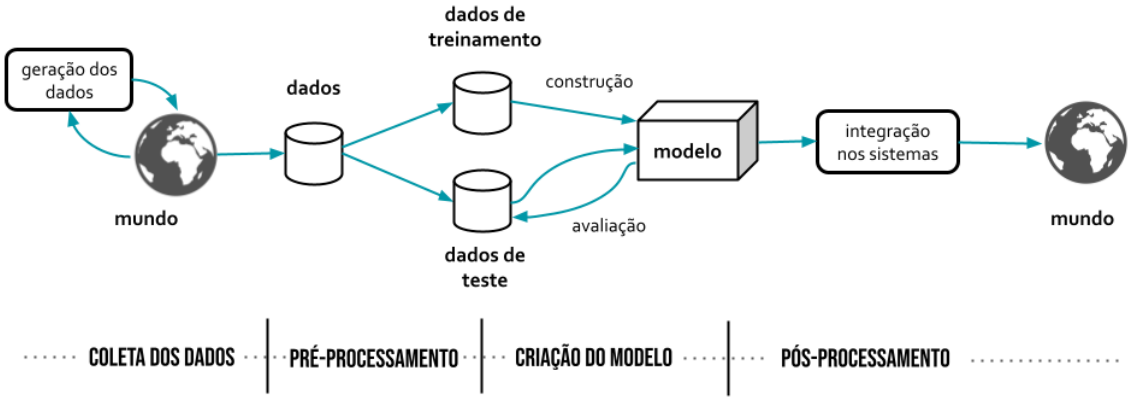


Figura 1. Etapas do aprendizado de máquina supervisionado.

2.1. Coleta dos Dados

O processo de aprendizado de máquina começa tipicamente com a coleta de dados. Tais dados são compostos por duas partes: entrada e saída. A entrada pode conter quaisquer características dos exemplos a serem treinados e a saída é o *rótulo* a ser predito. Por exemplo, para um sistema que irá detectar automaticamente e-mails que são spam, a entrada é um conjunto de mensagens de e-mail, já rotulado como "spam" ou "não spam", e a saída é o mesmo rótulo, predito para mensagens de e-mail ainda não conhecidas [Burkov 2019]. Já para um sistema que aprende a detectar faces em imagens, as entradas são as imagens já rotuladas e a saída é o rótulo "sim" (quando o modelo detecta a face) ou o rótulo "não" (caso contrário). Em alguns sistemas que detectam faces, a saída pode ser, também, as coordenadas do retângulo (*bounding box*, do Inglês) contornando o rosto detectado.

Os sistemas de reconhecimento facial podem ser usados tanto para verificar a identidade da pessoa, como no desbloqueio automático de smartphones, quanto para identificar uma pessoa em meio a muitas outras, a partir de um escaneamento "um-para-muitos".

Nestes sistemas, é feita uma busca em um banco imagens até que haja uma correspondência com a face desejada, a partir das características dos pontos nodais de cada face [Castelvecchi 2020]. Para estes sistemas, as *entradas* são basicamente um grande volume de imagens coletadas e os modelos são treinados para reconhecer a geometria das faces, considerando pontos que conectam, por exemplo, os olhos, o nariz, a boca e características como tamanho do queixo, distância entre olhos, entre outras, visando criar uma "impressão facial" (*faceprint*, do Inglês).

Intuitivamente, e de uma maneira geral, quanto maior for o conjunto de dados para o treinamento, menos provável é que o modelo erre em detectar faces em novas imagens. Se os exemplos utilizados durante o treinamento forem selecionados aleatoriamente, quando o modelo tentar reconhecer um rosto em meio a muitos outros presentes nos dados coletados, estatisticamente, é mais provável que a face seja corretamente reconhecida caso compartilhe características semelhantes com as outras presentes no conjunto de dados usado para treinar o modelo.

As pessoas programadoras e engenheiras que constroem modelos de aprendizado de máquina, na prática, frequentemente não geram tais dados, mas utilizam dados disponíveis gratuitamente online. Para sistemas de reconhecimento facial, por exemplo, alguns projetos que disponibilizam dados com imagens de rostos para download[1].

2.2. Pré-processamento

Os dados brutos passam então pela etapa de pré-processamento (Figura 1), que tem como objetivo transformar os dados brutos no conjunto de dados utilizado para o aprendizado. Esta tarefa geralmente é chamada de engenharia de características (*feature engineering*, em Inglês) e tem como objetivo selecionar — ou criar — as características (ou atributos) mais *informativas*, com alto poder preditivo, que permitem construir um modelo que dê boas predições (com uma taxa maior de acertos). Algumas vezes, porém, principalmente quando o conjunto de dados é produzido manualmente, estas características podem estar ausentes e então é necessária a remoção ou substituição de dados incompletos ou inválidos.

Os dados pré-processados são embaralhados e particionados em dois conjuntos: o conjunto de treinamento e o conjunto de teste. O conjunto de treinamento geralmente é maior e é utilizado para construir (ou treinar) o modelo; já o conjunto de teste é utilizado para avaliar o modelo antes de liberá-lo para ser integrado nos sistemas. Tipicamente, o conjunto de treinamento representa 70% dos dados e o de teste representa 30% dos dados [Burkov 2019]. Por exemplo, se um sistema que usa dados de treinamento para detecção de faces coleta 1000 imagens aleatórias de rostos, seguindo tal proporção, 700 delas são utilizadas para o treinamento e 300 para teste.

É importante reforçar que o aprendizado só funciona porque o modelo *aprende* com um conjunto de dados (o de treinamento) e é *avaliado* com um outro conjunto de dados (o de teste). A avaliação, neste contexto, é um processo incremental nas duas direções: o modelo é testado e otimizado até que atinja um desempenho considerado bom o suficiente. Dessa forma, caso o modelo gerado tenha um bom desempenho na sua

[1] https://www.face-rec.org/databases

avaliação, ele será bom em predizer novos exemplos que o algoritmo de aprendizagem ainda não conheceu, ao invés simplesmente "memorizar" os exemplos de treinamento.

Na prática, há ainda um terceiro conjunto: o conjunto de validação. Este conjunto de dados é geralmente utilizado para escolher o algoritmo para o aprendizado e encontrar os melhores valores das variáveis que definem o modelo matemático aprendido pelo algoritmo (chamados de hiperparâmetros do algoritmo). Neste trabalho, porém, nos referimos a ambos os conjuntos (validação e teste) somente como conjunto de teste, para simplificar o entendimento.

2.3. Criação e Avaliação do Modelo

Nesta etapa, um modelo estatístico — que utiliza algum algoritmo de aprendizado — é escolhido e então, baseado nos dados de treinamento aprendidos, é capaz de predizer rótulos para novos exemplos. Esta não é uma tarefa simples e os algoritmos de aprendizagem podem ser implementados em diferentes linguagens (como exemplo, para a linguagem Python, a biblioteca `scikit-learn` tem muitos algoritmos disponíveis para serem utilizados[2]). Para cada um dos algoritmos, são testados diferentes métodos de otimização e, então, a configuração que garante um melhor desempenho para o modelo é escolhida.

O modelo tem então o seu desempenho *avaliado*, ou seja, é verificado se o modelo *generaliza bem* classificando os exemplos do conjunto de testes (Figura 1). Para as tarefas de classificação, que predizem valores categóricos, uma tabela chamada de *matriz de confusão* é utilizada. A Figura 2 mostra um exemplo hipotético de matriz de confusão para um modelo de reconhecimento facial. A matriz resume o quão bem sucedido é o modelo ao ser utilizado para identificar novas faces. O eixo vertical se refere ao rótulo real do exemplo e o eixo horizontal se refere ao rótulo predito pelo modelo/classificador.

Naturalmente, as métricas que avaliam o desempenho dos modelos consideram os seus acertos e seus erros. Estas métricas estão descritas na literatura e são amplamente utilizadas [Lorena et al. 2021; Burkov 2019; Murphy 2022] Os acertos, nos sistemas de reconhecimento facial, são os casos em que o sistema fez o reconhecimento corretamente — seja reconhecendo um rosto ou não. Os erros são os casos em que o modelo erra — ao reconhecer erroneamente um rosto ou ao não reconhecer um rosto que existe nos dados. Os acertos do modelo são os "verdadeiros": VP (verdadeiros positivos) e VN (verdadeiros negativos). Os erros dos modelos são os "falsos": FP (falsos positivos) e FN (falsos negativos). No exemplo da Figura 2, temos um total de 160 predições, entre acertos e erros, de um modelo de reconhecimento facial.

Os verdadeiros positivos indicam quantos foram preditos como positivos e são de fato positivos, ou seja, quantos indivíduos foram corretamente reconhecidos pelo modelo. Já os falsos positivos indicam quantos foram preditos como positivos, mas não eram positivos, ou seja, casos que o modelo reconheceu o rosto em meio aos dados, mas não se tratava da mesma pessoa. De um total de 100 reconhecidos (casos positivos), seriam 70 os casos onde o modelo reconheceu corretamente a face fornecida como entrada e 30 os casos em que o modelo fez uma correspondência incorreta da face fornecida como entrada

[2] https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

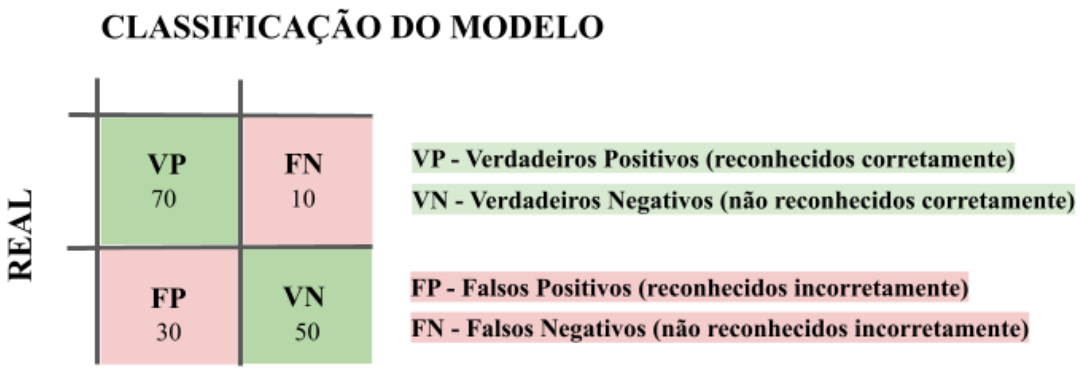


Figura 2. Exemplo hipotético de matriz de confusão para reconhecimento facial.

com uma outra face do conjunto de dados, ou seja, identificou como indivíduo procurado aquele que não é o correto.

Os verdadeiros negativos indicam quantos foram preditos como negativos e são de fato negativos, ou seja, casos em que o modelo não reconheceu o rosto e de fato ele não estava presente nos dados. Já os falsos negativos indicam quantos foram preditos como negativos, mas eram positivos, ou seja, casos em que o modelo não reconheceu o rosto, mas ele existia nos dados e o modelo não conseguiu fazer a correspondência. De um total de 60 não reconhecidos (negativos), seriam 50 os casos onde o modelo não fez a correspondência e de fato a face não existia nos dados e 10 casos em que o modelo não fez a correspondência da face, mas que havia uma imagem correspondente no conjunto de dados.

A matriz de confusão apresentada na Figura 2 é o principal mecanismo para se avaliar se o modelo teve um bom desempenho, pois a partir dos valores que ela apresenta, são calculadas diferentes *métricas*. A Tabela 1 apresenta as principais métricas utilizadas para avaliar o desempenho de classificadores. Apresentamos, juntamente com a fórmula utilizada para o cálculo, a descrição de cada métrica, exemplos de casos e a pontuação de cada métrica para o nosso exemplo.

A métrica mais intuitiva para avaliar os modelos é aquela que considera a proporção geral de acertos em relação ao total de predições e é chamada de *acurácia* (*accuracy*, do Inglês). No exemplo da matriz de confusão do modelo de reconhecimento facial da Figura 2, a acurácia do modelo seria de 120 (total de acertos) / 160 (total de predições), indicando que o modelo acertou 75% das vezes, portanto, teve uma acurácia de 75%.

Muitas vezes, porém, a acurácia sozinha para avaliar um modelo não é suficiente. Em sistemas que usam reconhecimento facial para a identificação de criminosos, por exemplo, os falsos positivos representam casos em que o modelo reconhece incorretamente uma face, o que pode levar a prisões de pessoas acusadas injustamente. Um outro exemplo de falsos positivos que pode trazer muitos prejuízos são os modelos de detecção de spam (um falso positivo em tais modelos representa um e-mail que foi classificado

como spam, mas não era spam). Para estes casos, a métrica de precisão (*precision*, do Inglês) é mais indicada. A precisão é a proporção entre os verdadeiros positivos e o total de positivos. Quanto mais falsos positivos, menor será a precisão (fórmula da Tabela 1), ou seja, um modelo com poucos falsos positivos tem uma boa precisão. No exemplo da Figura 2, a precisão seria de 70 (verdadeiros positivos) / 100 (total de positivos), indicando que o modelo teve uma precisão de 70%, ou seja, a probabilidade de um indivíduo reconhecido pelo modelo de reconhecimento facial ser culpado é de 70%.

Tabela 1. Métricas tradicionais para avaliar os modelos.

Métrica	Fórmula	Descrição	Pontuação Exemplo
Acurácia	$\frac{VP+VN}{VP+FN+FP+VN}$	Fração de acertos em relação a todas as previsões. Ex.: Probabilidade do modelo acertar a previsão	$\frac{70+50}{70+10+30+50}$ $0.75 = 75\%$
Precisão	$\frac{VP}{VP+FP}$	Fração de verdadeiros positivos em relação a todos os preditos positivos. Ex.: Probabilidade de um indivíduo reconhecido ser culpado	$\frac{70}{70+30}$ $0.7 = 70\%$
Revocação ou Taxa de Verdadeiros Positivos (TVP)	$\frac{VP}{VP+FN}$	Fração de verdadeiros positivos em relação a todos os casos que eram de fato positivos. Ex.: Probabilidade de um culpado ser corretamente reconhecido.	$\frac{70}{70+10}$ $0.875 = 87,5\%$
Taxa de Falsos Positivos (TFP)	$\frac{FP}{FP+VN}$	Fração de falsos positivos em relação a todos os casos que eram de fato negativos. Ex.: Probabilidade de um culpado não ser reconhecido	$\frac{30}{30+50}$ $0.37 = 37\%$

Em outras vezes, queremos evitar ao máximo os falsos negativos, como por exemplo, em modelos de previsão de diagnósticos de doenças (um diagnóstico negativo de uma doença existente diminui as suas chances de tratamento). Para estes casos, utilizamos a métrica *revocação* (*recall*, do Inglês). A revocação é a proporção entre os verdadeiros positivos e o total de exemplos que são de fato verdadeiros. Ou seja, quanto mais falsos negativos, menor será a revocação. No exemplo da Figura 2, a revocação seria de 70 (verdadeiros positivos) / 80 (total de exemplos que são de fato positivos), indicando que o modelo teve uma revocação de 87%, ou seja, a probabilidade de um culpado ser cor-

retamente reconhecido pelo modelo de reconhecimento facial é de $87,5\%$. A revocação também é conhecida como Taxa de Verdadeiros Positivos (TVP).

Em outras vezes, queremos analisar a taxa de falsos positivos, ou seja, a proporção de falsos positivos em relação a todos os casos que eram de fato negativos. Esta métrica é conhecida como Taxa de Falsos Positivos (TFP). Nos modelos de reconhecimento facial, por exemplo, esta métrica avalia a probabilidade de um culpado não ser reconhecido pelo modelo. Quanto mais verdadeiros negativos, menor será a Taxa de Falsos Positivos (TFP) (fórmula da Tabela 1). A pontuação da métrica para o nosso exemplo seria 30 (falsos positivos) / 80 (total de exemplos que são de fato negativos), ou seja, o modelo teve uma Taxa de Falsos Positivos (TFP) de 37%. Dessa forma, a probabilidade de um culpado não ser reconhecido pelo modelo é de 37%.

2.4. Pós-processamento

Finalizado o modelo de aprendizado de máquina, algumas outras etapas podem ser necessárias para que o modelo seja integrado nos sistemas e de fato usado nas aplicações. É preciso interpretar as saídas do modelo, de acordo com o propósito pelo qual o sistema foi construído. Para os casos em que o modelo gera como saída uma pontuação para cada classe, esta pontuação muitas vezes deve ser convertida nas classes desejadas, através da escolha de um *limiar*.

Por exemplo, se o algoritmo escolhido para o modelo de reconhecimento facial gera como saída uma porcentagem que representa a probabilidade da face detectada na foto ser de determinada pessoa, e o sistema espera uma saída binária (sim ou não) é preciso escolher um limiar para converter tal probabilidade em uma classificação binária. Um limiar escolhido de 90%, por exemplo, indica que a correspondência da face de entrada com uma outra face seria de 90%. Nestes casos, para as correspondências a partir de $0,9$, o rosto procurado é identificado. Já para as correspondências cuja saída do modelo é um número menor do que $0,9$, o rosto buscado não corresponde à imagem analisada.

O processo de desenvolvimento de um modelo de aprendizado de máquina geralmente é incremental, de forma que eles são retroalimentados com *feedbacks*. Os erros de predição do modelos servem como aprendizado para correção das falhas. Dessa forma, os modelos aprendem com os próprios erros, o que permite a melhora contínua do seu desempenho.

3. Vieses Inseridos no Aprendizado de Máquina

Os vieses podem surgir nas várias das etapas do aprendizado de máquina, detalhadas na seção anterior. Nesta seção, apresentamos alguns dos vieses que podem ser inseridos durante este processo. Na literatura, encontramos inúmeras terminologias e classificações de vieses que podem estar presentes neste processo [Bellamy et al. 2018; Verma e Rubin 2018; Suresh e Guttag 2019; Caton e Haas 2020; Mehrabi et al. 2021; Ruback et al. 2021] .

Neste trabalho, consideramos quatro deles: (1) viés histórico, (2) viés nos dados, (3) viés no modelo e (4) viés de interpretação humana. Consideramos dois destes vieses como *vieses computacionais*, gerados diretamente por escolhas na construção

e preparação dos dados e dos modelos (viés nos dados e viés no modelo) e os outros dois como *vieses não computacionais*, vieses relacionados a fatores sociais, culturais e éticos que vão além do uso dos dados, algoritmos e métricas (viés histórico e viés de interpretação humana). Na prática, porém, não consideramos que tais vieses — computacionais e não computacionais — estão dissociados, pelo contrário, não consideramos ser possível, na prática, mensurar qual deles tem maior importância ou peso para o problema como um todo. A seguir detalhamos cada um deles.

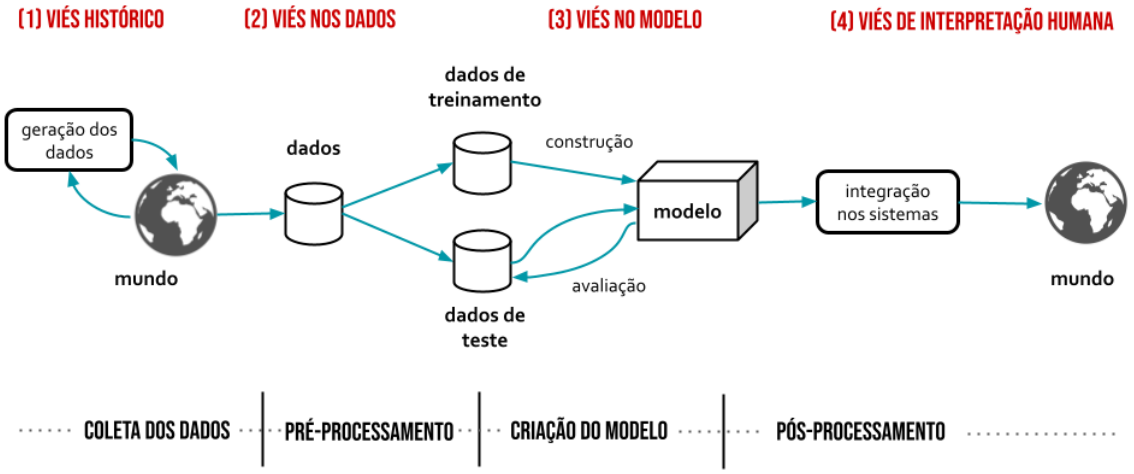


Figura 3. Vieses inseridos no aprendizado de máquina.

3.1. Viés Histórico

Os vieses históricos acontecem na etapa anterior à coleta de dados e podem se propagar por todo o *pipeline* do aprendizado de máquina. Quando vieses sociais e culturais inerentes nos dados que refletem resultados passados são discriminatórios, os modelos acabam perpetuando vieses de forma não intencional [Tecs USP 2018]. Quando o mundo como é — ou foi — leva a um modelo com vieses que reforça julgamentos e preconceitos dos indivíduos e instituições, como o racismo e os preconceitos de gênero, tais modelos reforçam estereótipos, que se refletem em casos de discriminação algorítmica.

A compreensão dos vieses históricos vai além da reflexão apresentada por O'Neil [O'Neil 2020] de que a compreensão sobre os padrões identificados nos modelos matemáticos que aparecerão no futuro, foram preestabelecidos no passado. De modo semelhante como ocorre com os modelos matemáticos apresentados aqui, é exatamente um retorno ao passado histórico da sociedade brasileira que auxilia no entendimento de que os dados algorítmicos incorporam e expressam sistemas de representação [Hall 2016], ou seja, a expressão de linguagens, sons, imagens e demais elementos culturalmente dotados de significado e carregados de sentido.

De acordo com Gonzalez [Gonzalez 2018], o racismo é fruto de uma construção ideológica cujas práticas tornam-se concretas por meio da perpetuação de diversos processos de discriminação e de desigualdade racial e de gênero, de acordo com os interesses

de grupos beneficiados por estes processos. A sociedade brasileira tem sido institucionalizada sobre as bases de um sistema historicamente herdado do sistema colonial escravagista [Carvalho 2021] que reproduz um processo de construção social destes corpos por meio de um racismo estrutural, que, intermediado pela organização política e econômica da sociedade brasileira [Almeida 2019], racializa, invisibiliza e desumaniza os corpos negros [Kilomba 2020; Gonzalez 2018; Ribeiro 2018]. O racismo atravessa os corpos desses indivíduos por meio da estigmatização anatômica advinda de premissas provenientes do racismo científico, largamente difundido nos espaços acadêmicos e políticos do século XIX [Almeida 2019].

Podemos compreender o viés histórico na prática quando nos deparamos com casos de racismo algorítmico. Muitos são os casos reportados de discriminação algorítmica. A Linha do Tempo do Racismo Algorítmico[3], desenvolvida por Tarcízio Silva, apresenta casos, reportagens e reações ao racismo algorítmico. Tarcízio Silva também organizou o livro 'Comunidades, algorítmos e ativismos digitais: olhares afrodiaspóricos', que "busca combater uma lacuna na academia brasileira: reflexões sobre a relação entre raça, racismo, negritude e branquitude com as tecnologias digitais como algoritmos, mídias sociais e comunidades online" [Silva 2020]. A seguir apresentamos exemplos de discriminação algorítmica em dois contextos diferentes.

Exemplo 1. Sistemas de contratação que utilizam modelos de aprendizado de máquina para selecionar candidatos para uma indústria predominantemente masculina e branca tendem a priorizar candidatos com estas características, sem considerar outros aspectos como diversidade e inclusão. Tais sistemas frequentemente recomendam desproporcionalmente a contratação de mais homens brancos para tais cargos, uma vez que se enquadram mais com a cultura já existente nestas empresas [D'Ignazio e Klein 2020]. Sem dados diversos para treinar estes modelos, tais ferramentas de contratação carregam os mesmos preconceitos que existiam na contratação de profissionais de tecnologia desde os anos 80[4].

Exemplo 2. Em um outro exemplo, no contexto de visão computacional, que ficou conhecido em 2016, um jovem mostrou a diferença ao fazer buscas no Google Imagens por "*three black teenagers*" (três adolescentes negros) e por "*three white teenagers*" (três adolescentes brancos)[5]. Para a busca por adolescentes negros, a plataforma exibia majoritariamente imagens usadas nas fichas policiais de jovens afro-americanos, enquanto para a busca por adolescentes brancos, as imagens predominantes exibiam jovens sorridentes, aproveitando seu tempo livre.

Reconhecer o viés histórico requer uma compreensão retrospectiva de como a opressão se manifestou em um determinado contexto ao longo do tempo [Suresh e Guttag 2019], como o racismo. Silvio de Almeida [Almeida 2019] define racismo como "uma forma sistemática de discriminação que tem a raça como fundamento, e que se manifesta por meio de práticas conscientes ou inconscientes que culminam em desvantagens ou privilégios para os indivíduos, a depender do

[3] https://tarciziosilva.com.br/blog/destaques/posts/racismo-algoritmico-linha-do-tempo

[4] https://cio.com.br/tendencias/por-que-a-inteligencia-artificial-na-contratacao-pode-ser-prejudicial

[5] https://brasil.elpais.com/brasil/2016/06/10/tecnologia/1465577075_876238.html

grupo racial ao qual pertençam". Os vieses históricos são, portanto, assim como o racismo estrutural, sistêmicos por natureza.

3.2. Viés nos Dados

Chamados também de vieses de representação, vieses de amostra ou vieses de seleção, os vieses nos dados são inseridos durante a coleta dos dados (Figura 3). A coleta e o pós-processamento de dados enviesados leva a modelos que não generalizam bem para determinados grupos populacionais. Muitos dos casos de preconceito algorítmico que vem emergindo nos últimos anos estão relacionados diretamente com as estratégias adotadas para coleta e tratamento dos dados nestas etapas. Os vieses nos dados podem se manifestar de diferentes maneiras. Detalhamos duas delas, a seguir.

Quando os dados coletados não são representativos da população a ser modelada. Quando a amostra coletada não é representativa da população, de forma balanceada, o modelo irá errar muito mais ao predizer rótulos para os grupos sub-representados.

Exemplo. Joy Buolamwini, pesquisadora do MIT, realizou uma das primeiras pesquisas que tratam de vieses em sistemas de reconhecimento facial. Boulamwini e Gebru [Buolamwini e Gebru 2018] analisaram o desempenho de modelos de classificação de gênero (feminino / masculino) por sistemas reconhecimento facial de alguns sistemas comerciais (Microsoft, da IBM e do Face++). A pesquisa concluiu que, no geral, homens e pessoas brancas foram melhor classificados pelos modelos do que os outros grupos. Uma visão interseccional da pesquisa revelou que todos os classificadores avaliados tiveram um pior desempenho ao classificar especificamente mulheres negras, explicitando a algoritmização de desigualdades socialmente estruturadas de raça e de gênero [Carrera 2020]. Análises interseccionais nos permitem compreender melhor as desigualdades, visto que as formas de preconceito — como o racismo e o sexismo — muitas vezes se sobrepõem, de forma que um mesmo indivíduo pode ser discriminado de várias formas [Collins e Bilge 2021; Akotirene 2018; Gonzalez 2018; Noble 2022; Carneiro 2003]. Pioneira na definição do conceito de *interseccionalidade*, Crenshaw [Crenshaw 2002] demarca este fenômeno como a captura das consequências estruturais resultantes do atravessamento e/ou das dinâmicas de interação entre dois ou mais eixos de opressão. A ótica interseccional permite uma compreensão aprofundada com relação ao modo como desigualdades como o patriarcalismo, o racismo e a opressão de classe estruturam as representações e posições sociais de mulheres, classes, raças e etnias, por exemplo [Crenshaw 2002]. Tais resultados são um exemplo de viés nos dados potencialmente presentes em modelos de reconhecimento facial, que se baseiam em dados de treinamento desbalanceados, treinados majoritariamente em rostos de homens e pessoas brancas.

A Figura 4 mostra a distribuição por gênero e tipo de pele em dois conjuntos de dados desbalanceados utilizados em sistemas de reconhecimento facial: Adience e IJB-A [Buolamwini e Gebru 2018]. Nos dados da Adience, enquanto homens de pele clara representam 41,6% do total, as mulheres negras representam 7,4%. Já nos dados do IJB-A,

a diferença é ainda maior: homens de pele clara representam 59,4% do total e mulheres negras representam somente 4,4% do total. Este desbalanceamento invariavelmente gera maiores taxas de erros — entre falsos positivos e falsos negativos — ao identificar os grupos sub-representados.

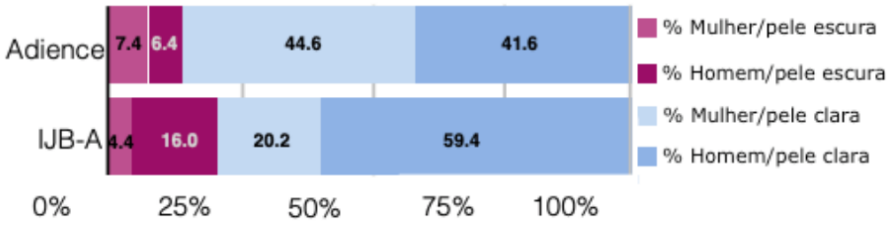


Figura 4. Exemplos de dados com vieses dos dados [Buolamwini e Gebru 2018].

Quando os dados coletados não refletem a população alvo do sistema. Quando os dados coletados não refletem corretamente a população que será utilizada para fazer as predições, seja no aspecto social, geográfico ou temporal, também podemos inserir vieses nos dados.

Exemplo 1. Quando a amostra coletada de um modelo contém, majoritariamente, dados representativos da cidade do Rio de Janeiro, e o modelo treinado será utilizado para a população da cidade de São Paulo, ele também irá apresentar um viés nos dados. De forma semelhante, dados coletados há 30 anos — independente da região — provavelmente não irão refletir a população que será alvo do modelo atualmente.

Exemplo 2. No diagnóstico de diabetes, por exemplo, a dosagem de hemoglobina A1C é um dos instrumentos mais importantes para controlar a doença. Porém, estudos mostram que os níveis de A1C variam de formas complexas entre diferentes etnicidades e gênero [Herman e Cohen 2012]. Portanto, a coleta dos dados usados para o treinamento do modelo que não considera as particularidades de cada uma das subpopulações analisadas, irá gerar um viés nos dados que pode prejudicar qualquer um dos diferentes grupos populacionais.

Na Seção 4, descrevemos algumas das estratégias para mitigar os vieses inseridos durante esta etapa.

3.3. Viés no Modelo

Chamados também de vieses no algoritmo, vieses de avaliação [Ruback et al. 2021] ou vieses no aprendizado [Suresh e Guttag 2019], os vieses no modelo que aparecem após o pós-processamento (Figura 3), durante a criação e avaliação do modelo.

É importante destacar que, neste trabalho, nós abordamos os vieses presentes no processo de aprendizado de máquina em geral. O modelo, em si, é gerado durante o processo e pode ser utilizado em outros sistemas. Portanto, o viés no modelo é um dos vieses que abordamos aqui. Os vieses no modelo podem se manifestar de diferentes maneiras. Detalhamos duas delas a seguir.

Vieses relacionados ao funcionamento interno do algoritmo. Alguns algoritmos utilizados pelos modelos assumem que correlação implica causalidade. A correlação (co-ocorrência frequente de dois fatores) é muitas vezes utilizada como indicador de uma relação causal, o que leva a saídas discriminatórias do modelo [Tecs USP 2018; Vigen 2015].

Exemplo Alguns sistemas da área financeira utilizam modelos para automatizar a decisão de concessão de crédito a clientes. Tais sistemas se baseiam em pontuações de crédito — processo de atribuição de pontos às variáveis de decisão de crédito de um indivíduo ao longo da sua vida como consumidor. Vilarino e Vicente [Vilarino e Vicente 2020] detalharam um modelo experimental de pontuação de crédito desenvolvido com dados reais no Brasil e demonstraram como o uso de informações de localização introduz preconceito racial. Os pesquisadores fizeram algumas simulações e observaram que, ao "mover" potenciais consumidores de regiões com uma maior população branca, como o estado de São Paulo, para regiões com uma maior população não branca, como o estado da Bahia, em 99,8% dos casos, as suas pontuações de crédito reduziram, embora todos os outros atributos, como idade e histórico financeiro não tivessem sido alterados. O modelo, dessa forma, assumiu que a correlação (localização geográfica e probabilidade de inadimplência) implica causalidade (a localização geográfica teve um papel determinante no risco de inadimplência).

Os pesquisadores também reforçam a importância de dados de censo para a pesquisas em aprendizado de máquina. Esta pesquisa demonstra como o preconceito racial algorítmico pode facilmente emergir em um modelo de pontuação de crédito com dados brasileiros.

Vieses relacionados à avaliação do modelo criado. Após a criação do modelo, os especialistas em aprendizado de máquina escolhem algumas métricas de avaliação de desempenho para os seus modelos (Tabela 1). Ao escolher, por exemplo, uma métrica geral como a acurácia para avaliar o modelo, podemos esconder disparidades entre os diferentes subgrupos [Suresh e Guttag 2019].

Por exemplo, um modelo de reconhecimento facial pode ter uma precisão geral de 80%, mas se formos considerar a precisão dentro do grupo que inclui mulheres negras, a precisão cai para 60%, enquanto que a precisão dentro do grupo que corresponde a homens brancos, a precisão sobe para 90%. Na Seção 4, descrevemos algumas das métricas alternativas para avaliar os modelos de forma a mitigar os vieses inseridos durante esta etapa.

3.4. Viés de Interpretação Humana

Os vieses de interpretação humana podem ser inseridos na etapa de pós-processamento, durante a integração dos sistemas (Figura 3). A saída do modelo, por exemplo, a identificação de um suspeito, nos sistemas de reconhecimento facial, deve ser sempre interpretada por seres humanos, de forma a evitar consequências injustas. Estes vieses ocorrem quando há uma incompatibilidade entre o problema que o modelo se propôs a resolver e a forma em que ele é usado na prática [Suresh e Guttag 2019]. Nos sistemas

de reconhecimento facial, o uso indiscriminado de modelos que já tiveram a sua eficácia questionada em determinados grupos, como de pessoas negras, sem oportunidade de uma avaliação humana para validar e/ou questionar os dados pode perpetuar problemas e desigualdades sociais e culturais. Muitas vezes, quando acontece o reconhecimento facial de um suspeito por um sistema, as autoridades que buscam a punição criminal de alguém já consideram a saída do modelo como prova da prática do crime, sem realizar uma análise mais aprofundada ou dar continuidade às investigações.

Rosa [Rosa et al. 2020], em seu artigo "Neutralidade tecnológica: reconhecimento facial e racismo", aborda políticas criminais de reconhecimento facial e suas aplicações e investiga "como a instrumentalização de dispositivos tecnológicos configura e compõe uma política criminal capaz de acentuar e multiplicar, em outros planos, o racismo que atravessa a sociedade". Os pesquisadores também destacam que "o acúmulo de informações sobre os indivíduos elevou a capacidade de países de vigiar, controlar, organizar, fiscalizar e punir por meio de dispositivos tecnológicos".

A opacidade destes modelos, ou seja, a dificuldade — ou impossibilidade — de se compreender claramente como estes modelos geraram as previsões, juntamente com a crença em sua objetividade, tornam o processo de interpretação humana ainda mais passível de ser enviesado. Neste contexto, há uma área de pesquisa dentro da computação, conhecida como *interpretabilidade*, que vem crescendo bastante nos últimos, e buscando estratégias para que seja possível "abrir" as caixas opacas (modelos) e assim interpretá-los [Arrieta et al. 2020].

De acordo com Osoba e Welser IV [Osoba e Welser IV 2017], este entendimento opaco a respeito das dinâmicas algorítmicas também tende a desencadear uma opacidade no entendimento acerca de suas deficiências. Sob esta ótica destes autores, uma compreensão mais aprofundada acerca da dinâmica dos algoritmos é essencial, principalmente em decorrência da atmosfera de objetividade e de imparcialidade algorítmica culturalmente construída sobre os algoritmos e os modelos de aprendizado de máquina. Especificamente, em relação à utilização de modelos de aprendizado de máquina no sistema de justiça criminal, o exemplo implementado em território norte-americano revela que esta prática — apesar de promover um alívio nas demandas relacionadas à gestão de um sistema tão grande e complexo — resultou na identificação de vieses, erros e desigualdades de caráter cumulativo e sistemático por meio destas ferramentas [Osoba e Welser IV 2017]. Com relação a estes aspectos, Lyon [Zuboff e Bruno 2018] explicita que as tecnologias de vigilância fornecem condições para que possíveis visões de mundo estruturadas socialmente sejam reproduzidas no interior dos seus próprios sistemas.

Uma reflexão sobre os vieses de interpretação humana precisa antes ser demarcada pela evidência de que, como afirmam Zuboff e Bruno [Zuboff e Bruno 2018], os processos de informação não são apenas compostos pela imposição de conjuntos de comandos programados, mas também pela produção de informações. E as informações produzidas podem estar relacionadas a significados socialmente construídos nas relações cotidianas e em espaços midiáticos como a internet, o cinema ou os veículos de informação jornalística [Zuboff e Bruno 2018], entre outros. Como exemplo, vemos várias situações na

visão computacional, como quando, em 2016, para a busca por "três adolescentes negros", a plataforma Google Imagens exibia majoritariamente imagens usadas nas fichas policiais de jovens afro-americanos, enquanto para a busca por "três adolescentes brancos", as imagens predominantes exibiam jovens sorridentes, aproveitando seu tempo livre. A partir destas questões, Lyon [Zuboff e Bruno 2018] ressalta a necessidade de que fatores como a implementação de normas regulatórias e de reflexões sobre o aspecto ético dessas práticas sejam constantemente levadas em consideração.

4. Mitigando Vieses Inseridos no Aprendizado de Máquina

Os vieses inseridos no processo de aprendizado de máquina, apresentados na Seção 3, representam um problema complexo e que exige esforços multidisciplinares. Tais vieses são incorporados no processo de aprendizado de forma consciente ou inconsciente, portanto, a sua eliminação completa se torna uma tarefa muito difícil. Utilizamos, portanto, neste trabalho o termo "mitigar", entendendo que para este problema não existe uma solução única, que funciona sempre e para todos os casos.

Para os *vieses computacionais* apresentados na Seção 3 (vieses nos dados e nos modelos), apresentamos *soluções computacionais*, que envolvem ajustes nos dados (Seção 4.1) e nos modelos (Seção 4.2). Para os *vieses não computacionais* (vieses históricos e de interpretação humana), descrevemos algumas *soluções não computacionais* (Seção 4.3), que vão além de soluções envolvendo o uso de dados, novos algoritmos e métricas. Vale relembrar neste ponto que, na prática, não consideramos que os vieses computacionais e não computacionais estão dissociados, pelo contrário, consideramos ser impossível, na prática, mensurar qual deles tem maior importância ou peso para o problema. Os esforços para mitigar os vieses devem cobrir ambas as esferas, que se complementam.

Neste ponto, surgem muitos questionamentos a respeito dos critérios para avaliar o quanto um sistema pode ser "justo". O termo *fairness*, em Inglês, traduzido muitas vezes como "equidade" (ou "justiça"), tem ganhado espaço na comunidade acadêmica e no mundo corporativo. Porém, trabalhar com a ideia de equidade não é simples, pois a própria ideia de equidade (ou de justiça) não é restrita a área de aprendizado de máquina, sendo parte também de múltiplas e diversas reflexões e teorias das ciências humanas [Cortiz 2020].

Apontamos nesta seção algumas direções para mitigar os vieses apresentados neste trabalho, que podem tornar os sistemas de aprendizado de máquina mais justos e inclusivos. Apresentamos também outras estratégias (que também podem ser compreendidas como soluções não computacionais, ou esforços não computacionais) para lidar com o problema, que vão desde a regulação do uso destes sistemas até as ações afirmativas e políticas públicas para promover diversidade e inclusão na área de tecnologia.

4.1. Mitigando Vieses nos Dados

Uma grande parte dos casos de preconceito algoritmo que desfavorecem um determinado grupo em relação a outros está diretamente relacionado ao processo de construção dos dados a serem utilizados pelo modelo para fazer as predições (Seção 3.2).

Os vieses nos dados podem ser gerados por várias razões. Em uma grande parte dos casos, a razão é o próprio desbalanceamento nos dados — quando a amostra coletada não é representativa da população, de forma balanceada. Para lidar com este problema, algumas medidas podem ser tomadas. Uma dessas medidas é a construção de dados de treinamento representativos.

Um outro problema relacionado aos vieses nos dados é quando os dados coletados não refletem a população alvo do sistema, como nos casos onde o modelo é construído com dados desatualizados ou é utilizado em uma região geográfica diferente daquele em que os dados foram coletados (Seção 3.2). A seguir apresentamos algumas iniciativas para mitigar os vieses nos dados.

Para modelos utilizados em sistemas de reconhecimento facial, alguns conjuntos de dados inclusivos vêm sendo construídos para gerar modelos mais justos. Estes dados podem ser utilizados tanto como dados de treinamento para modelos aprenderem quanto como dados de referência (*benchmarks*). Os *benchmarks* são conjuntos de dados utilizados como referência para comparar o desempenho de modelos com a mesma finalidade que vêm sendo propostos.

Um exemplo de conjunto de dados balanceado, que pode ser usado para treinar modelos de visão computacional, como os utilizados no reconhecimento facial, é o *FairFace*[6] [Karkkainen e Joo 2021], contendo mais de cem mil imagens curadas para mitigar vieses raciais, coletadas do Flicker e englobando 7 grupos raciais: Brancos, Negros, Indianos, Asiáticos do leste, Asiáticos do sudeste, Oriente Médio e Latinos.

Um outro conjunto de dados de referência representativo é o PPB (Pilot Parliaments Benchmark) (Figura 5), proposto por Joy Buolamwini e Timnit Gebru, no projeto *Gender Shades*[7], criado para prover uma melhor representação interseccional de gênero e raça em sistemas de visão computacional. Os dados contém 1270 imagens de rostos de pessoas, incluindo três países africanos e três países europeus, que pode ser usado tanto como dados de treinamento quanto como dados de teste e de referência para modelos que implementam reconhecimento facial. As pesquisadoras, sozinhas, conseguiram gerar dados mais precisos do que os oferecidos por algumas gigantes de tecnologia. O impacto das pesquisas iniciadas pelas pesquisadoras foi tão grande que, em meados de 2020, a IBM encerrou as suas pesquisas em reconhecimento facial, e se posiciona, hoje, contra o uso da tecnologia para monitoramento em massa e vigilância[8].

Além de se construir conjuntos de dados balanceados, estamos vendo algumas iniciativas para descontinuar o uso de alguns não balanceados. Um das maiores bibliotecas de aprendizado de máquina em Python, a scikit-learn, descontinuou, por questões éticas, um conjunto de dados descrevendo moradias em Boston[9]. Este conjunto de dados vem sendo bastante utilizado para demonstrações práticas em estudos e cursos de aprendizado de máquina e é usado como demonstração para treinar um modelo que prevê preços de

[6] https://paperswithcode.com/dataset/fairface

[7] http://gendershades.org

[8] https://g1.globo.com/economia/tecnologia/noticia/2020/06/09/ibm-encerra-area-de-pesquisa-em-reconhecimento-facial-e-pede-reforma-da-policia.ghtml

[9] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html

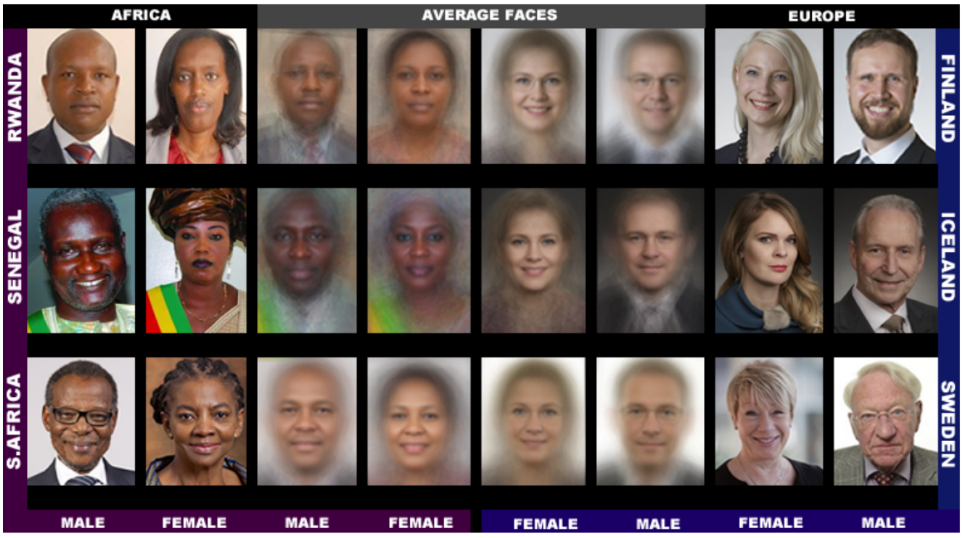


Figura 5. Dados inclusivos, com rostos de parlamentares de 6 países [Buolamwini e Gebru 2018].

imóveis em Boston, baseado em algumas características nas regiões. Um dos atributos utilizados para treinar o modelo representa, porém, a proporção de negros em cada uma das regiões e o modelo assume que a segregação racial tem um impacto positivo no valor dos imóveis, perpetuando o racismo sistêmico neste contexto[10].

Tais soluções apresentadas para lidar com os vieses nos dados demandam, porém, além de soluções técnicas, um tipo diferente de sabedoria de cientistas de dados e criadores de algoritmos, em sua maioria homens brancos, que envolve questões sociais e de políticas públicas a qual estes profissionais têm pouca exposição [Silva 2020].

4.2. Mitigando Vieses no Modelo

A discriminação no aprendizado de máquina pode acontecer quando algumas características de indivíduos usadas nos modelos acabam privilegiando um grupo em detrimento de outros. Os vieses no modelo podem ser inseridos, em geral, por duas razões: (a) pelo funcionamento interno do algoritmo utilizado e (b) pelas estratégias utilizadas para a criação e avaliação do modelo (Seção 3.3).

Os vieses relacionados ao funcionamento interno do algoritmo podem ser gerados, por exemplo, quando alguns modelos assumem que correlação implica causalidade e utilizam características de forma indevida nos algoritmos para treinar o modelo. Muitas abordagens para mitigar este tipo de viés consideram que alguns dos atributos, chamados de atributos protegidos — ou atributos sensíveis, não deveriam ser utilizados no treinamento para o aprendizado, pois definem, muitas vezes, aspectos dos dados que são potencialmente perigosos para as previsões, sob o ponto de vista sociocultural [Caton e Haas 2020].

Como exemplos de atributos protegidos, podemos citar: raça, gênero, etnia/nacionalidade, religião, idade, deficiência, entre outros [Bellamy et al. 2018]. Porém,

[10] https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8

a noção de atributo protegido vai além dessas dimensões, podendo abranger qualquer outra característica que envolva indivíduos. Neste ponto, surge uma dúvida muito comum na produção de modelos de aprendizado de máquina justos: Quais atributos devem ser protegidos? Alguns atributos são compreendidos universalmente como protegidos (ou sensíveis), como raça, gênero e etnia. Outros são explicitamente definidos em regulamentações específicas que dispõem sobre o tratamento de dados pessoais, com o objetivo de proteger os direitos de privacidade dos indivíduos. Ainda há casos de atributos que não são rigorosamente protegidos, mas que tem uma relação com atributos protegidos [Vilarino e Vicente 2020]. A Tabela 2 mostra alguns exemplos de atributos protegidos e equivalentes que são considerados correlacionados.

Tabela 2. Atributos protegidos e atributos co-relacionados Tabela adaptada de [Caton e Haas 2020].

Atributo protegido	**Atributos correlacionados**
Gênero	Nível de escolaridade Renda Profissão Dados criminais Conteúdo gerado pelas redes sociais Jornada de trabalho
Estado Civil	Nível de escolaridade Renda
Raça	Dados criminais Conteúdo gerado pelas redes sociais CEP
Deficiência	Dados de teste de personalidade

Na prática, mesmo que os modelos não utilizem os atributos protegidos listados na Tabela 2, por limitações impostas por regulações, por exemplo, muitas vezes utilizam outros atributos que consideram correlacionados para realizar as predições.

Alguns sistemas conhecidos como "avaliadores de risco" são utilizados para determinar a pena de condenados em alguns estados nos Estados Unidos como apoio aos juízes para proferir sentenças. Estes modelos indicam a probabilidade de uma pessoa que cometeu um crime reincidir, com base em mais de 100 atributos, incluindo idade, sexo e histórico criminal [Tecs USP 2018]. O sistema mais famoso de sentenciamento é o COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) [Freeman 2016; Washington 2018], que atribui aos réus pontuações de 1 a 10 para tal probabilidade. Embora a raça não seja explicitamente utilizada pelo modelo para fazer previsões (por ser um atributo considerado *protegido*), o questionário que alimenta o modelo do COMPAS inclui perguntas como "alguém da família foi preso?", "vive em uma área com alto índice de criminalidade?" ou "tem amigos que fazem parte de gangues?", assim como perguntas relacionadas ao seu histórico profissional e escolar[11].

[11] https://www.bbc.com/portuguese/brasil-37677421

Sabemos, porém, que algumas análises de causa-efeito se baseiam em correlação entre atributos considerados protegidos — por exemplo, quando estudos apontam que a renda, o gênero ou a raça influencia no rendimento escolar. Tais análises devem ser acompanhadas, além da análise estatística, de uma cautelosa análise sociológica, que considere aspectos sociais, culturais e demográficos da população analisada. Caso contrário, pode-se gerar análises baseadas em correlações espúrias [Geirhos et al. 2020] [12].

Além de proteger alguns atributos do "aprendizado", como descrito até aqui, os vieses no modelo podem ser mitigados também através da utilização de novas métricas de avaliação que levem em consideração não apenas requisitos técnicos, mas também aspectos sociais [Cortiz 2020], ao invés de simplesmente utilizar as métricas tradicionais para avaliar os modelos, como a acurácia, precisão e a revocação, apresentadas na Seção 2.3.

Na literatura, diferentes autores fornecem várias definições e métricas de justiça para avaliar os modelos [Bellamy et al. 2018; Verma e Rubin 2018; Suresh e Guttag 2019; Caton e Haas 2020; Mehrabi et al. 2021]. A maior parte dessas abordagens propõem formas matemáticas de se representar vieses, justiça e/ou discriminação e se baseiam em classificadores binários (Seção 2).

Estas métricas alternativas utilizam outros conceitos e definições, que apresentaremos a seguir. Quando a predição favorece um determinado indivíduo/grupo, representando uma vantagem em relação aos demais, podemos dizer que este indivíduo/grupo recebeu um *rótulo favorável* [Bellamy et al. 2018], como por exemplo, ter um crédito aprovado, ser contratado para um emprego e não ser preso injustamente (para os casos de falha nos sistemas de reconhecimento facial). Quando estes vieses indesejados fornecem uma vantagem sistemática para um grupo, este grupo é chamado de *grupo privilegiado*, e quando fornecem uma desvantagem sistemática, o grupo é chamado de *grupo não privilegiado*.

Nesta seção, descrevemos quatro métricas de justiça definidas por Bellamy et al. e utilizadas pela ferramenta *AI Fairness Toolkit*[13], desenvolvida pela IBM, para investigar e mitigar discriminação e vieses inseridos por modelos de aprendizado de máquina [Bellamy et al. 2018]. A ferramenta, de código aberto, oferece casos de demonstração utilizando alguns dados e vários algoritmos para mitigação dos vieses[14].

Para descrever estas métricas, retomamos o cenário do reconhecimento facial, que utiliza datasets com imagens de rostos para o treinamento. Consideramos também um recorte interseccional raça-gênero em 4 grupos: (1) mulheres brancas, (2) mulheres não brancas, (3) homens brancos e (4) homens não brancos. Consideramos aqui somente os atributos raça e gênero, por simplificação, mas é importante lembrar que os vieses podem prejudicar indivíduos pertencentes a outros grupos não contemplados aqui. Neste exemplo, consideramos o grupo com imagens compostas por homens brancos como grupo privilegiado e o grupo composto por imagens de mulheres negras como grupo não privilegiado. A Tabela 3 apresenta algumas das métricas que podem mitigar os vieses descritos neste trabalho, listadas a seguir.

[12] https://www.tylervigen.com/spurious-correlations
[13] https://aif360.mybluemix.net
[14] https://aif360.mybluemix.net/resources#guidance

Tabela 3. Métricas alternativas para mitigar vieses [Bellamy et al. 2018].

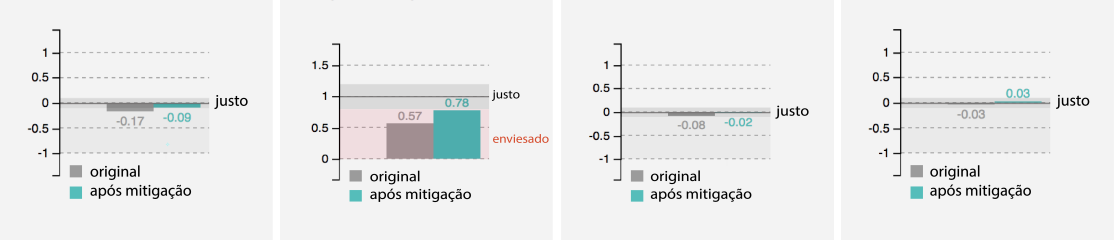
Métrica	Descrição	Pontuação	Exemplo grupos	Ex. Pontuação métrica
Diferença de paridade estatística (*statistical parity difference*)	Diferença entre a probabilidade de rótulos favoráveis entre grupos não privilegiados e privilegiados	0: não há diferença < 0 (menor do que 0): grupo privilegiado tem vantagem > 0 (maior do que 0): grupo não privilegiado tem vantagem	Probab. rótulo favorável grupo homens brancos (privilegiado): 85% Probab. rótulo favorável grupo mulheres negras (não privilegiado): 68%	$0,68 - 0,85 = -0,17$ Diferença de paridade estatística < 0: Favorece grupo privilegiado
Impacto desproporcional (*disparate impact*)	Razão entre a probabilidade de rótulos favoráveis entre grupos não privilegiados e privilegiados	1: não há impacto < 1 (menor do que 1): grupo privilegiado tem vantagem > 1 (maior do que 1): grupo não privilegiado tem vantagem	Probab. rótulo favorável grupo homens brancos (privilegiado): 91% Probab. rótulo favorável grupo mulheres negras (não privilegiado): 52%	$0,52/0,91 = 0,57$ Impacto desproporcional < 1: Favorece grupo privilegiado
Diferença de probabilidade média (*average odds difference*)	Diferença entre as taxas de falsos positivos (TFP) (especificidade) entre grupos não privilegiados e privilegiados	0: não há diferença < 0 (menor do que 0): grupo privilegiado tem vantagem > 0 (maior do que 0): grupo não privilegiado tem vantagem	Taxa de falsos positivos (TFP) grupo homens brancos: 60% Taxa de falsos positivos (TFP) grupo mulheres negras: 90%	$0,6 - 0,9 = -0,3$ Diferença de probabilidade média < 0: Favorece grupo privilegiado
Diferença de igualdade de oportunidade (*equal opportunity difference*)	Diferença entre as taxas de verdadeiros positivos (TVP) (revocação) entre grupos não privilegiados e privilegiados	0 : não há diferença < 0 (menor do que 0): grupo privilegiado tem vantagem > 0 (maior do que 0): grupo não privilegiado tem vantagem	Taxa de verdadeiros positivos (TVP) grupo homens brancos: 80% Taxa de verdadeiros positivos (TVP) grupo mulheres negras: 70%	$0,7 - 0,8 = -0,1$ Diferença de igualdade de oportunidade < 0: Favorece grupo privilegiado

Diferença de paridade estatística (*statistical parity difference*, do Inglês):

A diferença de paridade estatística considera a probabilidade de rótulos favoráveis entre grupos não privilegiados e privilegiados. Uma pontuação de 0 indica que não há diferença, já um valor menor do 0 que o indica que o grupo privilegiado tem uma maior vantagem e um valor maior do que 0 indica que o grupo não privilegiado tem uma vantagem.

No reconhecimento facial, por exemplo, considerando um modelo com uma probabilidade de 85% de um homem branco ser corretamente reconhecido (e, portanto, não correr o risco de ser preso injustamente) e de 68% no grupo de mulheres negras, o impacto desproporcional considerando estes dois grupos seria de $0,68 - 0,85$, portanto, $-0,17$. Neste caso, de acordo com esta métrica, tal modelo não seria justo e estaria fornecendo uma vantagem para o grupo privilegiado.

A Figura 6(a) mostra um exemplo de gráfico gerado, para esta métrica, antes e depois da aplicação de um algoritmo de mitigação [Bellamy et al. 2018]. A primeira barra horizontal, em cinza, representa a pontuação da métrica antes da mitigação e a segunda barra horizontal, em verde, representa a pontuação da métrica após a mitigação. No exemplo, a pontuação foi de $-0,17$ para $-0,09$, se aproximando mais de zero (pontuação máxima para esta métrica) e tornando o modelo um pouco mais justo, de acordo com esta métrica.



(a) (b) (c) (d)

Figura 6. Exemplos de pontuação para as métricas antes e após a mitigação. (a) Diferença de paridade estatística. (b) Impacto desproporcional. (c) Diferença de probabilidade média. (d) Diferença de igualdade de oportunidade. Figura adaptada de [Bellamy et al. 2018].

Impacto desproporcional (*disparate impact*, do Inglês):

O impacto desproporcional considera a razão entre a probabilidade de rótulos favoráveis entre grupos privilegiados e não privilegiados. Uma pontuação de 1 indica que ambos os grupos têm a mesma probabilidade de rótulos favoráveis, isto é, os grupos têm a mesma vantagem. Uma pontuação menor do que 1 indica uma maior vantagem para o grupo privilegiado e um valor maior do que 1 indica uma maior vantagem para o grupo não privilegiado.

No reconhecimento facial, por exemplo, considerando um outro modelo, com uma probabilidade de 91% de um homem branco ser corretamente reconhecido (e portanto, não correr o risco de ser preso injustamente) e de 52% no grupo de mulheres negras, o impacto desproporcional considerando estes dois grupos seria de $0,91/0,52$, portanto, 57% (ou $0,57$). Neste caso, tal modelo não seria justo, de acordo com a métrica impacto desproporcional.

A Figura 6(b) mostra um exemplo de gráfico gerado, para esta métrica, antes e depois da aplicação de um algoritmo de mitigação [Bellamy et al. 2018]. A primeira barra horizontal, em cinza, representa a pontuação da métrica antes da mitigação $(0,57)$ e a segunda barra horizontal, em verde, representa a pontuação após a mitigação $(0,78)$. A mitigação neste caso, também promoveu uma aproximação do um, tornando o modelo um pouco mais justo, de acordo com esta métrica.

Diferença de probabilidade média (*average odds difference*, do Inglês):

A diferença de probabilidade média considera a diferença entre as taxas de falsos positivos (TFP) (Tabela 1) entre grupos não privilegiados e privilegiados. Uma pontuação de 0 indica que não há diferença, já um valor menor do 0 que o indica que o grupo privilegiado tem uma maior vantagem e um valor maior do que 0 indica que o grupo não privilegiado tem uma vantagem.

No reconhecimento facial, por exemplo, considerando um modelo com uma taxa de falsos positivos (TFP) de 60% para o grupo composto por homens brancos e uma taxa de falsos positivos (TFP) de 90% para o grupo composto por mulheres negras, a diferença de probabilidade média considerando estes dois grupos seria de $0,6-0,9$, portanto, $-0,3$.

Neste caso, tal modelo não seria justo e estaria fornecendo uma vantagem para o grupo privilegiado.

A Figura 6(c) mostra um exemplo de gráfico gerado, para esta métrica, antes e depois da aplicação de um algoritmo de mitigação [Bellamy et al. 2018]. No exemplo, a pontuação foi de $-0,08$ para $-0,02$, se aproximando mais de zero e tornando o modelo um pouco mais justo, de acordo com esta métrica.

Diferença de igualdade de oportunidade (*equal opportunity difference*, do Inglês):

A diferença de igualdade de oportunidade considera a diferença entre as taxas de verdadeiros positivos (TVP) entre grupos não privilegiados e privilegiados. Uma pontuação de 0 indica que não há diferença, já um valor menor do 0 que o indica que o grupo privilegiado tem uma maior vantagem e um valor maior do que 0 indica que o grupo não privilegiado tem uma vantagem.

No reconhecimento facial, por exemplo, considerando um modelo com uma taxa de verdadeiros positivos (TVP) de 80% para o grupo composto por homens brancos e uma taxa de verdadeiros positivos (TVP) de 70% para o grupo composto por mulheres negras, a diferença de igualdade de oportunidade considerando estes dois grupos seria de $0,7 - 0,8$, portanto, $-0,1$. Neste caso, tal modelo não seria justo e estaria fornecendo uma vantagem para o grupo privilegiado.

A Figura 6(d) mostra um exemplo de gráfico gerado, para esta métrica, antes e depois da aplicação de um algoritmo de mitigação [Bellamy et al. 2018]. No exemplo, a pontuação foi de $-0,03$ para $0,03$, indicando que antes apresentava uma leve vantagem para o grupo privilegiado e, após a mitigação, uma leve vantagem para o grupo não privilegiado.

As quatro métricas descritas aqui representam, portanto, maneiras alternativas de se avaliar os modelos em relação à justiça e estão disponíveis na ferramenta *AI Fairness Toolkit*, desenvolvida pela IBM, que oferece casos de demonstração utilizando alguns dados e vários algoritmos para mitigação dos vieses[15]. Não fornecemos aqui, porém, uma lista exaustiva das métricas alternativas, mas elas podem ser encontradas na literatura [Bellamy et al. 2018; Verma e Rubin 2018; Suresh e Guttag 2019; Caton e Haas 2020; Mehrabi et al. 2021].

4.3. Mitigando Vieses Históricos e de Interpretação Humana

Os vieses históricos e de interpretação humana (Seções 3.1 e 3.4) não são vieses inseridos diretamente por um escolha na construção dos dados de treinamento ou do algoritmo utilizado para treinar o modelo, ou seja, não demandam soluções computacionais. Esses vieses são um reflexo das nossas estruturas sociais e culturais enviesadas, e demandam, antes de tudo, o reconhecimento e conscientização do problema por parte das pessoas analistas e programadoras envolvidas no processo de criação de tais modelos.

A discussão acerca da implementação de estratégias para mitigar estes vieses deve incluir, portanto, uma busca pelo despertar de uma consciência acerca das consequências sociais resultantes do racismo e do sexismo que ainda se fazem presentes na sociedade

[15] https://aif360.mybluemix.net/resources#guidance

brasileira [Gonzalez 2018; Saffioti 2004]. Esta consciência contribui para o conhecimento acerca do esquecimento, da alienação, do encobrimento, da rejeição e do desconhecimento resultantes do racismo [Gonzalez 2018]. A compreensão acerca do entrelaçamento entre estes dois vieses auxilia no entendimento de como desigualdades já existentes na sociedade são transmutadas para o espaço digital, perpetuando os sistemas sociais já existentes, reiterando as estruturas de poder e punindo os grupos sociais mais vulneráveis [Noble 2022; O'Neil 2020].

As possibilidades de mitigação destes vieses também precisam envolver um entendimento acerca de algumas dinâmicas socioculturais e que precisam ser evitadas o quanto possível, como a estereotipagem. Hall [Hall 2016] atesta que o processo de produção de significados, especificamente com relação à representação da diferença racial, por vezes é perpassado por práticas representacionais chamadas de estereotipagem. *Apropriando-se* de poucas características de uma pessoa, o estereótipo *reduz* esta pessoa a estes traços que, posteriormente, são *exagerados* e *simplificados* [Hall 2016, grifos do autor]. Assim, na visão do autor, a estereotipagem atua no processo de redução, essencialização, naturalização e fixação da diferença. Além disso, a prática da estereotipagem também implementa uma estratégia de cisão, que tipifica e separa as diferenças em duas categorias: o normal e aceitável e o anormal e inaceitável. Como consequência, Hall [Hall 2016] afirma que a estereotipagem, exclui tudo o que é diferente, tudo o que não está delimitado dentro dos limites de normalidade estabelecidos.

Fanon [Fanon 2008] ainda ressalta que as interações sociais que circundam o mundo são perpassadas por modos de criação e de vivência socialmente construídos por percepções advindas do racismo e do colonialismo. É sob estas duas bases que os negros são *construídos como negros* na sociedade [Fanon 2008]. Estes mecanismos sociais de racialização das pessoas negras muitas vezes suscitam comportamentos humanos discriminatórios, com base em reproduções estereotipadas, simplificando matematicamente os indivíduos em *scores*, reproduzindo assimetrias, perpetuando invisibilidades e amplificando as exclusões.

Neste contexto, é importante destacar o impacto da adoção de políticas de inclusão e/ou cotas no país no combate à desigualdade e ao racismo estrutural. Ao implementar tais estratégias, as bases produzidas nestes novos contextos naturalmente já incluem alguns dos grupos historicamente marginalizados, o que pode reduzir significativamente os vieses históricos nestas bases. Replicar estratégias como estas é essencial para a adoção de tecnologias digitais dissociadas do contexto de relações de poder presentes no cenário global com base em fatores raciais, por exemplo, como afirma Noble [Noble 2022]. Mas, vale ressaltar que esta estratégia também precisa ser atravessada pela observação de marcos que envolvam também a dimensão interseccional para reduzir desigualdades de classe, raça, gênero e das demais formas de opressão [Collins e Bilge 2021].

Por fim, os vieses históricos e de interpretação humana também precisam ser mitigados por meio da implementação de políticas de letramento com relação a questões étnico-raciais direcionadas à sociedade brasileira em geral, para reduzir o distanciamento entre a população brasileira e as heranças culturais construídas pelos povos originários do Brasil e pela população afro-diaspórica. O alcance da história e das culturas afro-brasileira

e indígena nos currículos pedagógicos regulamentados pelas Leis nº. 10.639/2003[16] e 11.645/2008[17], por exemplo, possuem um papel essencial no reconhecimento do valor cultural e da ressignificação dos saberes dos povos originários do Brasil e dos povos afro-diaspóricos como fontes de conhecimento legítimas e fundamentais para a formação da identidade cultural brasileira.

4.4. Legislação

Na literatura, muitas abordagens voltadas para mitigar os vieses no aprendizado de máquina teorizam sobre os aspectos legais, sociais e éticos relacionados à discriminação algorítmica. As primeiras pesquisas sobre vieses, discriminação e justiça foram impulsionadas pela lei americana dos direitos civis de 1964, que tornou ilegal a discriminação com base em raça, cor, religião, sexo ou origem nacional. No Brasil, a lei que trata de tais crimes surgiu somente em 1989[18], assinada pelo então presidente da República José Sarney.

O Regulamento Geral de Proteção de Dados da União Europeia[19] (*General Data Protection Regulation*, em Inglês) foi aprovado em meados de 2016 e entrou em vigor em meados de 2018 em toda a União Europeia. A lei representou um marco regulatório em relação à privacidade e proteção de dados pessoais e sensíveis de cidadãos europeus. Inspirada na lei europeia, a brasileira Lei Geral de Proteção de Dados Pessoais (LGPD)[20] foi aprovada em meados de 2018 e sancionada em outubro de 2020. O artigo 5 da LGPD considera três grandes grupos de dados:

I - **dado pessoal**: informação relacionada a pessoa natural identificada ou identificável;
II - **dado pessoal sensível**: dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural;
III - **dado anonimizado**: dado relativo a titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento.

A LGPD estabelece usos específicos para cada um destes grupos de dados, sendo estes mais restritivos para os dados sensíveis (Artigo 11) do que para os dados pessoais (Artigo 7). Os dados biométricos processados por sistemas de reconhecimento facial podem ser reconhecidos, pela LGPD, como dados pessoais, já que eles nos permitem extrair

[16] Lei nº. 10.639, de 9 de janeiro de 2003. Altera a Lei nº. 9.394, de 20 de dezembro de 1996, que estabelece as diretrizes e bases da educação nacional, para incluir no currículo oficial da Rede de Ensino a obrigatoriedade da temática "História e Cultura Afro-Brasileira", e dá outras providências.

[17] Lei nº. 11.645, de 10 de março de 2008. Altera a Lei nº. 9.394, de 20 de dezembro de 1996, modificada pela Lei no 10.639, de 9 de janeiro de 2003, que estabelece as diretrizes e bases da educação nacional, para incluir no currículo oficial da rede de ensino a obrigatoriedade da temática "História e Cultura Afro-Brasileira e Indígena".

[18] https://ambitojuridico.com.br/cadernos/direito-penal/o-brasil-e-o-preconceito-uma-analise-teorica-e-critica-da-lei-7-716-89-frente-a-realidade-brasileira

[19] https://gdpr-info.eu

[20] http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

informações que podem identificar uma pessoa. Mas também podem ser considerados dados sensíveis, pela sua própria definição, que consta na lei [Carvalho et al. 2021].

Na prática, porém, os sistemas de reconhecimento facial são muitas vezes utilizados sem restrições, gerando muitas das consequências já apresentadas neste artigo. San Francisco, em 2019, foi a primeira cidade dos Estados Unidos a banir o uso de reconhecimento facial na segurança pública pelo governo[21], o que inspirou outras cidades pelo mundo. No Rio de Janeiro, por exemplo, está em tramitação na Câmara Municipal do Rio de Janeiro o Projeto de Lei nº. 824/2021[22], que tem como objetivo proibir o uso de tecnologias de reconhecimento facial pelo poder executivo municipal.

No âmbito da LGPD é importante ressaltar que embora alguns de seus aspectos sejam satisfatórios com relação ao sigilo dos dados pessoais e à proteção da identidade, para-além da regulamentação normativa dos objetivos, dos fundamentos e de demais categorizações, faz-se necessária a promulgação de novas regulamentações normativas que protejam as pessoas de forma eficaz e efetiva.

Os elementos demonstrados na justificativa para a implementação do Projeto de Lei nº. 824/2021 apontam para alguns fatores que precisam ser levados em conta com relação aos vieses que podem se fazer presentes nas tecnologias de reconhecimento facial: os riscos de aprofundar desigualdades históricas, assim como de reproduzir estigmas e exclusões; a aplicação de tecnologias inadequadas e ineficazes que ampliam o risco de discriminação, erros e confusões; a implementação de uma vigilância opressiva e contínua sobre rostos de indivíduos específicos; a aplicação de recursos financeiros públicos em tecnologias caras e com baixa eficiência em troca do investimento em setores públicos mais necessitados; o desperdício de recursos humanos em trabalhos extras evitáveis para a abordagem de casos falsos positivos; a ameaça à segurança da identidade de pessoas na faixa-etária infanto-juvenil; e a invisibilidade com relação aos rostos de grupos sociais vulneráveis como pessoas negras, mulheres, pessoas transgênero e não-binárias.

Estes argumentos demonstram que os processos automatizados e semi-automatizados que compõem o aprendizado de máquina podem atuar como formas de controle social que afetam diretamente liberdades e direitos fundamentais, como a dignidade, a liberdade de ir e vir, a privacidade e a preservação da imagem e da honra das pessoas. Questões como estas, apresentadas no Projeto de Lei nº. 824/2021, explicitam que, no contexto brasileiro, a discussão sobre resoluções jurídicas para mitigar estes vieses necessita de uma avaliação crítica que leve em consideração as experiências anteriores vivenciadas em outros locais, como a Europa e os EUA que, por exemplo, colocaram em pauta a possibilidade de banimento das tecnologias de reconhecimento facial em seus territórios.

4.5. Diversidade e Inclusão na Tecnologia

A criação de modelos de aprendizado de máquina mais justos demanda, além das soluções computacionais e não computacionais mencionadas até aqui, a garantia de que estes siste-

[21] https://epocanegocios.globo.com/Tecnologia/noticia/2019/05/centro-da-revolucao-tecnologica-sao-francisco-bane-o-uso-de-reconhecimento-facial-pelo-governo.html

[22] http://aplicnt.camara.rj.gov.br/APL/Legislativos/scpro2124.nsf/8446f2be3d9bb8730325863200569352/33b9222f266e43710325872700723005?OpenDocument

mas reflitam os valores das pessoas ou da população para quem vão servir. A diversidade e a inclusão, tanto no mercado de trabalho quanto na esfera acadêmica, que discute estes problemas e aponta soluções, são essenciais como estratégia para tornar estes sistemas mais justos.

Este cenário, porém, está longe de ser alcançado, tanto no contexto mundial quanto no Brasil. Mulheres e outros grupos seguem sub-representados na área. Algumas pesquisas e relatórios vêm apontando que a crise de diversidade é sistêmica na área de tecnologia.

Em se tratando da crise de diversidade de gênero na esfera acadêmica, o relatório *Gender Diversity in AI Research*[23], produzido pela fundação de inovação britânica *Nesta*, aponta que as mulheres seguem sub-representadas em todas as subáreas dentro da Ciência da Computação. O relatório analisou publicações em Inteligência Artificial no *arXiv*, repositório amplamente utilizado pela comunidade acadêmica. A pesquisa aponta que as mulheres representam 13,8% dos autores de artigos na área. Em se tratando de artigos de autoria única, representam somente 6,72% dos artigos.

No mundo corporativo, o cenário é também desafiador. Segundo o *Relatório Global de Equidade de Gênero*[24], do Fórum Econômico Mundial de 2021, as mulheres representam, globalmente, 32% da força de tarefa nas áreas de Inteligência Artificial e Ciência de Dados.

Já quando se trata de diversidade racial na tecnologia, o cenário é tão invisibilizado que faltam inclusive dados precisos para mensurar o problema. Segundo Silvana Bahia, diretora de projetos do Olabi[25], organização social que trabalha para democratizar a produção de tecnologia, a ausência de mulheres negras e indígenas na tecnologia está ligada diretamente a dois fatores: acesso e falta de referências. Segundo a diretora, os custos relacionados à formação na área são muito caros, muitas vezes inacessíveis, em Inglês, e são raras as políticas (públicas ou privadas) destinadas ao ingresso e permanência de mulheres negras nestes espaços. Além disso, a ausência de referências positivas sobre mulheres negras na área é uma questão social que vai além do mundo da tecnologia e atinge os mais variados campos profissionais e de poder.

Enfrentar a crise de diversidade na área da tecnologia, se configura, portanto, como uma estratégia fundamental para mitigar os vieses apresentados neste trabalho, juntamente a regulação do uso dos modelos e das soluções computacionais e não computacionais apresentadas aqui.

5. Conclusão

Os modelos de aprendizado de máquina, que aprendem a partir de exemplos, vêm sendo usados em cada vez mais sistemas para automatizar processos, como em sistemas de recrutamento, de diagnóstico de doenças e de reconhecimento facial. Estes sistemas vêm apresentando falhas e vieses, favorecendo determinados grupos em relação a outros, o que mostra que eles não são modelos imparciais, como se acredita muitas vezes.

[23] https://media.nesta.org.uk/documents/Gender_Diversity_in_AI_Research.pdf
[24] https://www3.weforum.org/docs/WEF_GGGR_2021.pdf
[25] https://www.pretalab.com

Neste trabalho, apresentamos uma análise sociotécnica dos vieses inseridos durante o processo de aprendizado de máquina. Apresentamos, portanto, além dos principais aspectos técnicos para compreender como e quando podem ser inseridos os vieses neste processo, uma análise sobre as suas implicações sob o ponto de vista histórico-social.

Primeiro, descrevemos as etapas nas quais podem ser inseridos quatro vieses comumente relatados na literatura: vieses históricos, vieses nos dados, vieses nos modelos e vieses de interpretação humana e, então, apresentamos exemplos de implicações sociais e culturais em modelos utilizados em vários domínios, como no reconhecimento facial, nos modelos que predizem as chances de reincidência de pessoas que cometeram crimes, entre outros.

Apontamos algumas estratégias que vêm sendo utilizadas para mitigar estes vieses. Para os vieses nos dados, abordamos a construção de bases de treinamento representativas. Para mitigar os vieses nos modelos, apresentamos métricas alternativas para avaliá-los em relação à justiça e considerando grupos privilegiados e não privilegiados. Para os vieses históricos e de interpretação humana, exemplificamos como eles são reflexo das nossas estruturas sociais e culturais enviesadas.

As direções que apresentamos aqui englobam não somente soluções computacionais, mas também não computacionais, como o reconhecimento e conscientização do problema por parte das pessoas envolvidas no processo de criação destes modelos. Apresentamos também a importância da regulação através de projetos de lei e a necessidade de políticas para enfrentar a crise de diversidade na tecnologia.

Agradecimentos

D. Carvalho é financiada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. S. Avila é financiada parcialmente pelo CNPq-PQ2 315231/2020-3, FAPESP 2013/08293-7, 2020/09838-0, H.IAAC e Google LARA 2021.

Referências

[Akotirene 2018] Akotirene, C. (2018). O que é interseccionalidade? *Belo Horizonte: Letramento*. 23:12

[Almeida 2019] Almeida, S. (2019). *Racismo estrutural*. Pólen Produção Editorial LTDA. 23:11

[Arrieta et al. 2020] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115. 23:15

[Bellamy et al. 2018] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*. 23:3, 23:9, 23:18, 23:20, 23:21, 23:22, 23:23

[Broussard 2018] Broussard, M. (2018). Machine learning: The DL on ML. In *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press. 23:2

[Buolamwini e Gebru 2018] Buolamwini, J. e Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency*, pages 77–91. 23:2, 23:12, 23:13, 23:18

[Burkov 2019] Burkov, A. (2019). *The hundred-page machine learning book*, volume 1. Andriy Burkov Canada. 23:4, 23:5, 23:6

[Carneiro 2003] Carneiro, S. (2003). Mulheres em movimento. *Estudos avançados*, 17:117–133. 23:12

[Carrera 2020] Carrera, F. (2020). A raça e o gênero da estética e dos afetos: algoritmização do racismo e do sexismo em bancos contemporâneos de imagens digitais. *MATRIZes*, 14(2):217–240. 23:12

[Carvalho 2021] Carvalho, D. (2021). O legado do sistema colonial escravagista como base para a gênese do sistema capitalista no brasil: a persistência do racismo no cotidiano da população negra. *Cadernos Cemarx*, 14:e021006–e021006. 23:11

[Carvalho et al. 2021] Carvalho, L. P., Oliveira, J., Santoro, F. M., e Cappelli, C. (2021). Social network analysis, ethics and lgpd, considerations in research. *iSys - Brazilian Journal of Information Systems*, 14(2):28–52. 23:26

[Castelvecchi 2020] Castelvecchi, D. (2020). Is facial recognition too biased to be let loose? *Nature*, 587(7834):347–349. 23:2, 23:5

[Caton e Haas 2020] Caton, S. e Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*. 23:3, 23:9, 23:18, 23:19, 23:20, 23:23

[Collins e Bilge 2021] Collins, P. H. e Bilge, S. (2021). *Interseccionalidade*. Boitempo Editorial. 23:12, 23:24

[Cortiz 2020] Cortiz, D. (2020). Inteligência artificial: equidade, justiça e consequências, 2020. https://www.cetic.br/media/docs/publicacoes/6/20200626161010/panorama_setorial_ano-xii_n_1_inteligencia_artificial_equidade_justi%C3%A7a.pdf. 23:16, 23:20

[Crenshaw 2002] Crenshaw, K. (2002). Documento para o encontro de especialistas em aspectos da discriminação racial relativos ao gênero. *Revista estudos feministas*, 10:171–188. 23:12

[D'Ignazio e Klein 2020] D'Ignazio, C. e Klein, L. F. (2020). *Data feminism*. MIT press. 23:11

[Fanon 2008] Fanon, F. (2008). Pele negra, máscaras brancas (r. silveira, trad.). *Salvador, BA: EdUFBA*. 23:24

[Freeman 2016] Freeman, K. (2016). Algorithmic injustice: How the wisconsin supreme court failed to protect due process rights in state v. loomis. *North Carolina Journal of Law & Technology*, 18(5):75. 23:19

[Geirhos et al. 2020] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., e Wichmann, F. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673. 23:20

[Giddens 2008] Giddens, A. (2008). *Sociologia*. 6a. edição. Tradução: Alexandra Figueiredo, Ana Patrícia Duarte Baltazar, Catarina Lorga da Silva, Patrícia Matos, Vasco Gil. Fundação Calouste Gulbenkian. 23:3

[Gonzalez 2018] Gonzalez, L. (2018). *Primavera para as rosas negras: Lélia Gonzalez em primeira pessoa.* Editora Filhos da África. 23:10, 23:11, 23:12, 23:24

[Hall 2016] Hall, S. (2016). Cultura e representação. *PUC-Rio: Apicuri*. 23:10, 23:24

[Henriques 2018] Henriques, T. S. (2018). A concepção sociotécnica quatro perspectivas francesas sobre a articulação entre tecnologia e sociedade. *Revista Habitus*, 16(2). 23:2

[Herman e Cohen 2012] Herman, W. H. e Cohen, R. M. (2012). Racial and ethnic differences in the relationship between hba1c and blood glucose: implications for the diagnosis of diabetes. *The Journal of Clinical Endocrinology & Metabolism*, 97(4):1067–1072. 23:13

[Karkkainen e Joo 2021] Karkkainen, K. e Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558. 23:17

[Kilomba 2020] Kilomba, G. (2020). *Memórias da plantação: episódios de racismo cotidiano*. Editora Cobogó. 23:11

[Lorena et al. 2021] Lorena, A., Faceli, K., Almeida, T., de Carvalho, A., e Gama, J. (2021). *Inteligência Artificial: uma abordagem de Aprendizado de Máquina (2a edição)*. 23:6

[Maybin 2016] Maybin, S. (2016). Sistema de algoritmo que determina pena de condenados cria polêmica nos eua. *BBC News, São Paulo, out*. 23:2

[Mehrabi et al. 2021] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., e Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6). 23:3, 23:9, 23:20, 23:23

[Mitchell 1997] Mitchell, T. M. (1997). Machine learning. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877. 23:3

[Murphy 2022] Murphy, K. P. (2022). *Probabilistic Machine Learning: An introduction*. MIT Press. 23:6

[Nascimento 2020] Nascimento, L. F. (2020). *Sociologia digital: uma breve introdução*. Salvador: EDUFBA. 23:3

[Noble 2022] Noble, S. U. (2022). *Algoritmos da Opressão: Como os mecanismos de busca reforçam o racismo*. Editora Rua do Sabão. 23:12, 23:24

[Osoba e Welser IV 2017] Osoba, O. A. e Welser IV, W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation. 23:15

[O'Neil 2020] O'Neil, C. (2020). Algoritmos de destruição em massa: como o big data aumenta a desigualdade e ameaça a democracia. *Tradução Rafael Abrahan. Santo André, SP: Editora Rua do Sabão*. 23:2, 23:10, 23:24

[Petukhov e Steshina 2018] Petukhov, I. e Steshina, L. (2018). Decision-making problems in sociotechnical systems. *Management of Information Systems*. 23:3

[Raghavan et al. 2020] Raghavan, M., Barocas, S., Kleinberg, J., e Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Conference on Fairness, Accountability, and Transparency*, page 469–481. 23:2

[Raji et al. 2020] Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., e Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151. 23:2

[Raschka 2015] Raschka, S. (2015). *Python machine learning*. Packt publishing ltd. 23:3, 23:4

[Ribeiro 2018] Ribeiro, D. (2018). *Quem tem medo do feminismo negro?* Editora Companhia das Letras. 23:11

[Rosa et al. 2020] Rosa, A., Pessoa, S. A., e Lima, F. S. (2020). Neutralidade tecnológica: reconhecimento facial e racismo. *REVISTA V! RUS*, 21. http://www.nomads.usp.br/virus/virus21/?sec=4&item=9&lang=pt. 23:15

[Ruback et al. 2021] Ruback, L., Avila, S., e Cantero, L. (2021). Vieses no aprendizado de máquina e suas implicações sociais: Um estudo de caso no reconhecimento facial. In *Anais do II Workshop sobre as Implicações da Computação na Sociedade*, pages 90–101. SBC. 23:4, 23:9, 23:13

[Saffioti 2004] Saffioti, H. I. B. (2004). Gênero, patriarcado, violência. In *Gênero, patriarcado, violência*, pages 151–151. 23:24

[Silva 2020] Silva, T. (2020). *Comunidades, algoritmos e ativismos digitais: Olhares afrodiaspóricos*. LiteraRua. 23:11, 23:18

[Suresh e Guttag 2019] Suresh, H. e Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2. 23:3, 23:9, 23:11, 23:13, 23:14, 23:20, 23:23

[Tavares 2021] Tavares, M. (2021). Plataforma usa inteligência artificial para conseguir ḿatchéntre vaga de emprego e candidato. https://g1.globo.com/economia/pme/pequenas-empresas-grandes-negocios/noticia/2021/02/07/plataforma-usa-inteligencia-artificial-para-conseguir-match-entre-vaga-de-emprego-e-candidato.ghtml. 23:2

[Tecs USP 2018] Tecs USP (2018). Inteligências artificiais, preconceitos reais. https://medium.com/tecs-usp/intelig%C3%AAncias-artificiais-preconceitos-reais-f30c018cb2dd. 23:10, 23:14, 23:19

[Tomalin et al. 2021] Tomalin, M., Byrne, B., Concannon, S., Saunders, D., e Ullmann, S. (2021). The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology*, pages 1–15. 23:2

[Verma e Rubin 2018] Verma, S. e Rubin, J. (2018). Fairness definitions explained. In *International Workshop on Software Fairness (FairWare)*, pages 1–7. 23:3, 23:9, 23:20, 23:23

[Vigen 2015] Vigen, T. (2015). *Spurious correlations*. Hachette UK. 23:14

[Vilarino e Vicente 2020] Vilarino, R. e Vicente, R. (2020). Dissecting racial bias in a credit scoring system experimentally developed for the brazilian population. *arXiv preprint arXiv:2011.09865*. 23:2, 23:14, 23:19

[Washington 2018] Washington, A. L. (2018). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colo. Tech. LJ*, 17:131. 23:2, 23:19

[Zuboff e Bruno 2018] Zuboff, S. e Bruno, F. (2018). Tecnopolíticas da vigilância: Perspectivas da margem. *Organização de Fernanda Bruno et al. Tradução de Heloísa Cardoso Mourão et al. 1ª. ed. São Paulo: Boitempo*. 23:15, 23:16