Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação

Fabio Gagliardi Cozman
Dora Kaufman

Em palestra proferida em 1985, Richard Feynman, Prêmio Nobel (1965) e um dos mais reconhecidos físicos teóricos, debate temas críticos do campo da inteligência artificial (IA)[1]. O diálogo com o público

1 Disponível em: https://www.programmersought.com/article/88885940764/. Acesso em: 26/7/2021. Vídeo da palestra disponível em: https://www.cantorsparadise.com/richard-feynman-on-artificial-general-intelligence-2c1b9d8aae31. Acesso em: 25/7/2021.

FABIO GAGLIARDI COZMAN é professor da Escola Politécnica da Universidade de São Paulo e diretor do Center for Artificial Intelligence (C4AI).

DORA KAUFMAN é professora do Programa de Tecnologias Inteligentes e Design Digital (TIDD) da Faculdade de Ciências Exatas e Tecnologias da PUC/SP.

tem início com uma pergunta-chave, que remete ao artigo seminal de Alan Turing (1950), que definiu o que hoje é conhecido como "Máquina de Turing": "Você acha que haverá uma máquina que pode pensar como os humanos e ser mais inteligente do que os humanos?". Para Feynman, as futuras máquinas não pensarão como os seres humanos, da mesma forma que um avião não voa como os pássaros. Dentre outras diferenciações, os aviões não batem asas; são processos, dispositivos e materiais distintos. Quanto à questão das máquinas superarem a inteligência humana, na visão do físico o ponto de partida está na própria definição de "inteligência".

De fato, é difícil definir o que entendemos por "inteligência". Segundo Stuart Russell, pesquisador referência no campo da IA, uma entidade é inteligente na medida em que o que faz é capaz de alcançar o que deseja, ou seja, seus objetivos. Escreve Russell (2019, p. 9): "Todas essas outras características da inteligência – perceber, pensar, aprender, inventar e assim por diante – podem ser compreendidas por meio de suas contribuições para nossa capacidade de agir com sucesso". Russell lembra que o conceito de inteligência, desde os primórdios da filosofia grega antiga, está associado a capacidades humanas (perceber, raciocinar e agir), o que não seria o caso da IA, "meros" modelos de otimização com objetivos definidos pelos humanos e não dotados desses atributos. Outros autores não consideram a "inteligência" uma prerrogativa humana, como o próprio Marvin Minsky (1985), um dos fundadores do campo da IA, ao argumentar que os sistemas de IA têm habilidades, apesar de limitadas, de aprendizagem e raciocínio. Complicando ainda mais esse debate, as técnicas atuais de IA lidam com percepção, análise de texto, processamento de linguagem natural (PNL), raciocínio lógico, sistemas de apoio à decisão, análise de dados e análise preditiva (*Stanford Encyclopedia*, 2020, apud Stone et al., 2016).

Outro tema abordado por Feynman em sua palestra de 1985 foi o reconhecimento de padrões em grandes conjuntos de dados, à época um desafio ainda não totalmente viabilizado por técnicas empíricas de IA. A programação computacional, pondera o físico, não contemplaria as nuances da realidade, como, por exemplo, luminosidades, distâncias e ângulos de inclinação da cabeça num conjunto de fotos – os seres humanos são capazes de reconhecer uma pessoa pelo movimento do corpo ao andar, pela maneira como mexe no cabelo e outros pequenos e sutis detalhes.

Resolver tarefas executadas pelos humanos intuitivamente, e com relativo grau de subjetividade, era um desafio dos primórdios do campo da IA. Várias tentativas envolvendo linguagens formais, apoiadas em regras de inferência lógica, tiveram êxito limitado, sugerindo a necessidade de os sistemas gerarem seu próprio conhecimento extraindo padrões de dados, ou seja, "aprender" com os dados sem receber instruções explícitas. Esse processo é usualmente denominado "aprendizado de máquina" (*machine learning*), subcampo da IA criado em 1959 e hoje certamente o maior subcampo da IA em número de praticantes (Domingos, 2015; Goodfellow, Bengio & Courville, 2016; Alpaydin, 2016).

Um algoritmo de aprendizado de máquina é um algoritmo capaz de aprender com experiências; como definido por Tom Mitchell (1997): "Diz-se que um programa

de computador aprende com a experiência E em relação à classe de tarefas T e à medida de desempenho P, se seu desempenho nas tarefas medidas por P melhora com a experiência E". O processo de aprendizagem desses sistemas é influenciado por múltiplos fatores observáveis ou não observáveis no mundo físico, sujeitos a efeitos de fontes externas: por exemplo, os pixels em uma imagem de um carro vermelho podem estar muito próximos do preto à noite, e a forma da silhueta de um carro varia com o ângulo de visão (Goodfellow, Bengio & Courville, 2016).

A técnica de aprendizado de máquina que hoje melhor resolve esses desafios é o aprendizado profundo (*deep learning*), que introduz representações complexas, frequentemente referidas como "redes neurais profundas", expressas em termos de outras representações mais simples organizadas em diversas camadas. As entradas (*inputs*) são apresentadas a uma camada visível, assim chamada porque contém as variáveis observáveis, seguida de uma série de camadas ocultas contendo variáveis não observáveis e internas ao próprio modelo (origem do problema da não explicabilidade). Essa estrutura codifica uma função matemática que mapeia conjuntos de valores de entrada (*inputs*) para valores de saída (*output*); redes com maior profundidade (mais camadas) têm apresentado resultados positivos em várias áreas, particularmente em visão computacional, reconhecimento de voz e imagem (Goodfellow, Bengio & Courville, 2016).

Em redes neurais profundas, os parâmetros aprendidos a partir de dados são chamados de *weights* (pesos); após a fase de treinamento (ou aprendizado), esses pesos compõem o algoritmo e passam a ser fixos. No caso de uma imagem, em que os pixels são os dados de entrada, a saída do sistema reflete a soma das multiplicações de pesos pelos pixels de entrada. Cada camada processa o que supõe-se serem conceitos mais abstratos do que da camada anterior, gerando o nível de abstração requerido pela saída. Por exemplo, a saída pode ser *dog vs cat*, e a entrada pode ser a imagem (conjunto de pixels); cada camada mais "profunda" (mais próximo da saída) tem valores representando conceitos mais abstratos que ajudam, eventualmente, a concluir se é gato ou cachorro. A questão da interpretabilidade (ou opacidade, ou não explicabilidade) decorre do desconhecimento do que as camadas realmente representam.

Essa relativamente nova técnica de aprendizado de máquina, baseada fortemente em redes neurais de aprendizado profundo (*deep learning neural networks* - DLNN), tem sua inspiração no funcionamento do cérebro biológico. As DLNN são capazes de lidar com dados de alta dimensionalidade, por exemplo, milhões de pixels num processo de reconhecimento de imagem. Adicionalmente, DLNN estabelecem correlações não perceptíveis aos desenvolvedores humanos, cuja tendência é considerar apenas as correlações "mais fortes", embora as correlações "mais fracas", quando agrupadas, possam impactar sensivelmente a acurácia dos modelos.

Para avaliar o desempenho das técnicas de aprendizado de máquina mede-se sua precisão, ou seja, a proporção de exemplos para os quais o modelo produz a saída correta (ou, inversamente, a taxa de erro, ou seja, a proporção de exemplos para os quais o modelo produz uma saída incorreta). Em 2012, uma rede neural convolu-

cional (CNN) chamada AlexNet, uma das arquiteturas das DLNN, venceu por uma ampla margem o Desafio ImageNet 2012[2], reduzindo a taxa de erro de reconhecimento de imagem de 26,1% para 15,3%. Desde então, a competição é consistentemente vencida por essas redes, com a taxa de erro declinando para 3,6% (equiparável ao erro humano). Esses resultados positivos são função da disponibilidade de grandes conjuntos de dados e da maior capacidade computacional (Kaufman, 2019). Assim, as DLNN tornaram-se fator estratégico de processos decisórios pela capacidade de gerar *insights* preditivos com taxas relativamente altas de acurácia, permeando a maior parte das aplicações atuais de IA.

Entretanto, as DLNN ainda possuem limitações, já que requerem abundância de dados, pois a qualidade dos resultados é função da quantidade de dados utilizados no desenvolvimento, treinamento e aperfeiçoamento dos modelos. De fato, a arquitetura complexa dos modelos demanda *hardware* com grande capacidade de processamento. Ademais, dentre as externalidades negativas, destacam-se a opacidade dos modelos, ou falta de explicabilidade (ou seja, como os algoritmos chegaram à saída com base nos dados de entrada), e o viés contido nos resultados dos modelos, tema central deste artigo.

Em geral, atribui-se vieses integralmente às bases de dados tendenciosas. Porém, vieses podem emergir antes da coleta de dados em função das decisões tomadas pelos desenvolvedores (os atributos e variáveis contemplados no modelo, inclusive, determinam a seleção dos dados). No caso de viés associado aos dados, existem duas principais origens: os dados coletados não representam a composição proporcional do universo objeto em questão, ou os dados refletem os preconceitos existentes na sociedade. O primeiro caso pode ocorrer, por exemplo, se uma base de dados de treinamento contiver mais observações de uma categoria que de fato é minoritária; o segundo caso, por exemplo, é ilustrado pelo sistema de triagem de recrutamento automatizado da Amazon implantado em 2014 (descontinuado em início de 2017): em 2015, a empresa identificou que seu sistema não estava sendo neutro em termos de gênero, privilegiando candidatos homens. Os algoritmos de IA do sistema foram treinados, ou seja, "aprenderam" a identificar padrões na base de dados de currículos enviados à Amazon ao longo de um período de dez anos, refletindo o domínio masculino na indústria de tecnologia, ou seja, a realidade enviesada (viés histórico)[3].

Resultados tendenciosos podem decorrer, igualmente, de erros na rotulagem da base de dados que antecede o aprendizado supervisionado e na própria geração de dados, por exemplo, a não desagregação por gênero. A constatação do viés em um modelo, em geral, ocorre tardiamente, dificultando identificar retroativamente sua origem e,

2 ImageNet, base de dados para treinar algoritmos de IA, apresentada publicamente em 2009, na Conferência sobre Visão Computacional e Reconhecimento de Padrões (CVPR).

3 Fonte: Reuters, 10/10/2018. Disponível em: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Acesso em: 10/8/2021.

consequentemente, as formas de eliminá-lo. Especialistas em IA estão empenhados em identificar formas de eliminar ou, ao menos, mitigar os vieses dos modelos, a partir de variadas abordagens (Hao, 2019).

O recente avanço das tecnologias de IA associadas, por exemplo, à detecção e análise facial, se, por um lado, produz resultados mais assertivos com benefícios sociais, por vezes tipifica o racismo sistêmico além de agregar novas formas de discriminação derivadas de amostragens desequilibradas de dados, prática de coleta e rotulagem (Leslie, 2020). O propósito deste artigo é descrever e refletir sobre as principais origens do viés contido na aplicação de IA com o uso da técnica de redes neurais profundas (DLNN) e alguns dos caminhos, técnicos e sociais, para sua mitigação.

VIÉS NOS MODELOS BASEADOS EM DLNN

O uso das tecnologias de análise facial não tem sido neutro no que diz respeito à distribuição das externalidades positivas e negativas, apresentando desempenhos distintos para distintos grupos demográficos. Batya Friedman, da Universidade de Washington, e Helen Nissenbaum, da Universidade de Cornell, escreveram um dos primeiros artigos sobre sistemas de computação tendenciosos, alertando sobre o potencial impacto na sociedade dado o custo relativamente baixo de disseminação desses sistemas (Christian, 2020).

Buolamwini e Gebru (2018) auditaram sistemas de classificação de gênero produzidos pela Microsoft e IBM e apuraram que a taxa de classificação incorreta para as mulheres de pele escura era, em média, 35% mais alta do que para os homens brancos. O modelo FaceDetect da Microsoft, por exemplo, demonstrou uma taxa de erro geral de 6,3% em suas tarefas de classificação de gênero, contudo, quando seu desempenho foi analisado em termos da interseção de gênero e raça, os resultados mostraram que, enquanto a aplicação teve uma taxa de erro de 0% para homens de pele clara, teve uma taxa de erro de 20,8% para mulheres de pele escura. Para as autoras, esses resultados enviesados evidenciam que as bases de dados usadas no treinamento e validação desses sistemas sub-representaram pessoas de cor e mulheres.

Os sistemas de reconhecimento facial automatizados (*facial detection and recognition technologies* - FDRT) podem ser agrupados em três categorias de detecção: a) se há uma pessoa na imagem (*face detection*), sem atribuir identidade e atributos específicos; b) que tipo de pessoa aparece na imagem (sexo, idade, etnia, estado emocional/expressão facial); e c) quem é a pessoa na imagem, estabelecendo e/ou verificando identidades pessoais. As FDRT são usadas em instituições financeiras (adicionam segurança às operações bancárias), produtos de consumo (laptops, celulares), eventos de entretenimento, habitação (sistemas de câmaras), polícia (apoiar investigação, pesquisa em banco de dados, identificação), escolas (verificar frequência e avaliar atenção do aluno), varejo (sistemas de pagamento), transporte (aeroportos, transporte público), locais de trabalho (acesso) (Learned-Miller et al., 2020a).

A extensão do uso das FDRT está permeada pelo dilema entre os ganhos em segurança, proteção e eficiência, e as ameaças às liberdades civis e aos direitos humanos

fundamentais. O grau de acurácia depende de decisões dos desenvolvedores dos sistemas e da base de dados usada no treinamento e validação do sistema. Cada uma dessas etapas está sujeita a efeitos tendenciosos (viés dos modelos).

Karen Hao (2019) adverte que para detectar é imprescindível compreender como o viés surge na base de dados, sendo comum se atribuir características tendenciosas aos dados de treinamento, quando isso pode surgir nas várias etapas do processo, particularmente: a) no enquadramento do problema, quando o desenvolvedor traduz o objetivo a ser alcançado em linguagem computável; b) na coleta dos dados, no caso da base não ser representativa da realidade ou refletir os preconceitos existentes na sociedade; e c) na preparação dos dados, quando cabe ao desenvolvedor selecionar os atributos a serem considerados pelo algoritmo, que difere de (a) porque os mesmos atributos podem ser usados para objetivos diferentes. Hao alerta que, mesmo quando detectado, é difícil corrigir o viés nesses sistemas, inclusive porque a detecção pode ocorrer quando o sistema já está plenamente em uso, o que explica os casos em que seus proprietários optaram por descontinuar (por exemplo, sistema de seleção de candidatos para vagas de tecnologia da Amazon e o *chatbot* para interagir com adolescentes Ty da Microsoft, dois dos mais citados casos de discriminação algorítmica). Vejamos as principais origens do viés.

Viés na geração dos dados

A discriminação na produção de dados está presente tanto na predominância de usuários dos países desenvolvidos com mais acesso a tecnologias e às redes sociais, o que engendra uma base de dados enviesada pelo biotipo racial de pele clara, quanto na não desagregação dos dados por gênero e/ou o tratamento dado ao homem como "humano padrão". Ademais, a internet não é um espaço público inteiramente democrático; as informações não circulam com a mesma velocidade nem com o mesmo alcance; as redes que têm mais conexões ampliam as oportunidades de gerar mais conexões. Pontos com número maior de conexões são denominados *hubs* (concentrador e/ou conector): quanto mais links um *hub* capta, maior sua visibilidade na rede. Plataformas tecnológicas como Google, Amazon, Facebook, ao concentrar parte relevante do fluxo de informações na internet, constituem-se em poderosos *hubs* (Barabási, 2009), alavancados significativamente pelos algoritmos de IA.

Caroline Criado Perez (2021), por meio de um extenso levantamento histórico da "invisibilidade" feminina, constata que, como a técnica de IA que permeia a maior parte das aplicações atuais é baseada em dados, a sociedade está tomando decisões enviesadas por gênero em número maior do que o percebido. Na Inglaterra, por exemplo, as mulheres têm 50% mais chances de serem diagnosticadas erroneamente após um ataque cardíaco, em função da predominância de homens nos estudos científicos sobre insuficiência cardíaca (Perez-Criado, 2021). A prática de não coletar dados desagregados por gênero, tratando os homens como neutros e/ou "padrão humano", e a partir dessas bases de dados tendenciosas identificar padrões de comportamento humano, distorce a suposta objetividade e a acurácia dos resultados dos modelos estatísticos baseados em IA.

Viés nas escolhas dos desenvolvedores

No desenvolvimento de um modelo de DLNN, a tarefa inicial dos cientistas da computação é identificar o problema a ser resolvido pelo sistema, em que situação e com qual objetivo o sistema será utilizado. O segundo passo é traduzir esse problema a ser resolvido em variáveis que possam ser observadas e manipuladas (*feature engineering process*). São eles que definem, por exemplo, quais termos de pesquisa serão usados para coletar os dados, o número de camadas ocultas e o número de nós em cada camada. Identificar a influência da subjetividade humana no projeto e na configuração do algoritmo de IA não é trivial, além de não ser possível eliminá-la mesmo se identificada (Hao, 2019).

O Alan Turing Institute (Leslie, 2020) aponta como um dos problemas críticos que permitem que os vieses sistêmicos se infiltrem nos dados deriva da postura dos desenvolvedores e designers de algoritmos, que não priorizam as ações para identificar e corrigir desequilíbrios potencialmente discriminatórios na representação demográfica e fenotípica. O instituto atribui esses vieses à complacência dos produtores de tecnologia, em geral, parte do grupo dominante e, logo, isentos dos efeitos adversos de resultados discriminatórios. Equipes inter e multidisciplinares de desenvolvedores, potencialmente, podem atenuar esses efeitos discriminatórios, mas sua eficácia depende de construir "pontes" entre os pesquisadores de diferentes campos de conhecimento (Kaufman, 2021).

Viés na base de dados

O viés ocorre se os dados de referência forem menos diversificados demograficamente do que a população-alvo, ou seja, se a base de dados contiver poucos ou nenhum exemplo de uma determinada subpopulação por etnia e/ou gênero. A diferença entre ambientes controlados (laboratórios) e ambientes não controlados (mundo real), igualmente, tem o potencial de gerar resultados tendenciosos; nas ruas, por exemplo, as câmaras podem captar imagens em baixa resolução, o ângulo captado da face e a luminosidade podem dificultar a extração de características faciais ou mesmo distorcê-las provocando erro no reconhecimento facial (Learned-Miller et al., 2020b).

É recente a sensibilidade de pesquisadores, e da sociedade em geral, para o problema do viés nos dados, consequentemente, durante anos diversas bases de dados tendenciosas foram utilizadas para desenvolver e treinar os algoritmos de IA (e continuam sendo). O ImageNet, por exemplo, demorou uma década (2009 a 2019) para reconhecer o viés na rotulagem de suas imagens, mesmo assim por iniciativa do artista americano Trevor Paglen (ver "Viés no processo de rotulagem dos dados", abaixo). Outro exemplo de banco de dados enviesado de domínio público é o Labeled Faces in the Wild (LFW), organizado em 2007 com base em artigos de notícias *on-line* e rotulado por uma equipe da Umass Amherst. Em 2014, Hu Han e Anil Jain, da Michigan Sates, notaram que nesse banco de dados mais de 77% das imagens eram de homens e, nesse conjunto, mais de 83% de homens de pele clara; o ex-presidente dos EUA, George W.

Bush, tinha 530 imagens exclusivas, mais do que o dobro do conjunto de imagens de todas as mulheres de pele escura combinadas. Cinco anos depois, e 12 da data de constituição do LFW, seus gestores postaram um aviso de isenção de responsabilidade, alertando que muitos grupos não estão bem representados (Christian, 2020).

O United States Office of the Director of National Intelligence – supervisor da implementação do Programa de Inteligência Nacional, principal assessor do presidente, do Conselho de Segurança Nacional e do Conselho de Segurança Interna para assuntos de inteligência relacionados à segurança nacional –, em 2015, lançou um banco de dados de imagens faciais denominado IJB-A, supostamente contemplando a diversidade da população americana. Entretanto, estudo de Gebru e Buolamwini constatou que 75% eram imagens de homens e 80% de homens de pele clara, e apenas 4,4% do conjunto de dados eram de mulheres de pele escura (Christian, 2020).

O viés algorítmico, em geral ético/moral ou legal, é difícil de ser detectado por estar atrelado a sistemas proprietários (não auditáveis sem consentimento), mas também pela diversidade de composição dos sistemas de IA mais sofisticados (desenvolvidos em distintos locais e treinados em múltiplas base de dados).

Viés no processo de rotulagem dos dados

Criar uma base de dados de treinamento significa amostrar um mundo quase infinitamente complexo e variado, e fixá-lo em taxonomias compostas de classificações. Em 2006, cientistas da computação das universidades de Stanford e Princeton, liderados por Fei-Fei Li, deram início ao desenvolvimento do ImageNet, base de dados para treinar algoritmos de IA; o projeto foi apresentado publicamente em 2009, na *"Conference on Computer Vision and Pattern Recognition"* (CVPR) realizada na Flórida, EUA, constituindo-se numa base de dados padrão dos desenvolvedores de IA.

Kate Crawford (2021) investigou as falhas na rotulagem do ImageNet, com resultados surpreendentes. O banco de dados hoje contém aproximadamente 14 milhões de exemplos rotulados de mais de 20 mil classes/categorias, basicamente rotulados a mão por participantes da Amazon Mechanical Turk (*crowdsourcing* de força de trabalho distribuída e, relativamente, de baixa remuneração). Manter a uniformidade na classificação manual de grandes conjuntos de dados é um desafio, que se torna quase inviável quando envolve classificar imagens de pessoas; são inúmeras as categorias classificatórias, incluindo raça, idade, nacionalidade, profissão, *status* econômico, comportamento, caráter e até mesmo moralidade. Estruturar uma taxonomia para classificar imagens de pessoas com a lógica utilizada para objetos gera inúmeras distorções e, consequentemente, vieses. Por uma década, o ImageNet teve 2.832 subcategorias na categoria "pessoa": "avô", com 1.662 imagens; "pai", com 1.643 imagens; e "diretor executivo", com 1.614 imagens, a maioria homens.

No ImageNet, a categoria "corpo humano" é enquadrada como objeto natural – corpo – corpo humano; as subcategorias incluem "pessoa", "corpo masculino", "corpo juvenil", "corpo adulto" e "corpo feminino".

A suposição explícita é que apenas os corpos masculinos e femininos são reconhecidos como "naturais", seguindo uma classificação biológica, ou seja, binária, não reconhecendo as pessoas de gênero não binário, como os transexuais (Crawford, 2021).

Em 2019, o artista americano Trevor Paglen, dedicado ao tema da vigilância em massa e da coleta de dados, a pesquisadora em IA Kate Crawford e o especialista em tecnologias Leif Ryge desenvolveram o aplicativo ImageNet Roulette como parte de uma exposição de arte sobre os sistemas de reconhecimento de imagem no museu Fondazione Prada, em Milão, intitulada "Training Humans"[4]. Baseado num modelo de DLNN com código aberto Caffe, criado na UC Berkeley, o propósito do aplicativo era facilitar ao público a compreensão sobre os sistemas de aprendizado de máquina. Quando o usuário efetua o *upload* de sua foto, o aplicativo retorna a imagem com o rótulo atribuído pelo aplicativo. "'Training Humans' explora duas questões fundamentais: como os humanos são representados, interpretados e codificados por meio de conjuntos de dados de treinamento e como os sistemas tecnológicos coletam, rotulam e usam este material" (texto da curadoria da exposição).

Segundo relatório do Alan Turing Institute, desde 2019 nenhum dos dez maiores conjuntos de dados de imagens de rosto em grande escala foi rotulado ou anotado para tipo de pele, tornando as disparidades de desempenho entre diferentes grupos raciais praticamente invisíveis para aqueles que usaram esses conjuntos de dados para treinar seus modelos de IA (Leslie, 2020).

Viés nos dados de treinamento dos algoritmos

Considera-se que existe um enviesamento na base de dados quando o sistema exibe um erro sistemático no resultado ("enviesamento estatístico" ou "discriminação algorítmica"). Estritamente, qualquer conjunto de dados poderá ser imparcial para a execução de uma determinada tarefa, contudo, potencialmente existe o risco de que, se usado para uma tarefa distinta, seja tendencioso para essa segunda tarefa. Um sistema citado com frequência nos debates sobre discriminação algorítmica é o Compas.

O Correctional Offender Management Profiling for Alternative Sanctions (Compas) é um sistema desenvolvido por Tim Brennan, da Universidade do Colorado, em parceria com Dave Wells, na empresa fundada por ambos em 1998, Northpointe. Em 2001, o estado de Nova York iniciou um programa piloto usando o Compas na automatização de decisões sobre liberdade condicional; no final de 2017, todos os 57 condados fora da cidade de Nova York haviam adotado o sistema Compas em seus departamentos encarregados de "liberdade condicional". Os resultados, aparentemente, eram tão promissores que, em 2011, uma lei estadual estabeleceu que todas as decisões sobre liberdade condicional deveriam decorrer de sistemas automatizados de avaliação de riscos. Até 2015, o Compas recebeu cobertura favorável da mídia; a partir de junho de 2016, o tom

4 Exposição com curadoria compartilhada com Kate Crawford. Disponível em: http://digicult.it/slider/training-humans-an-exhibition-by-kate-crawford-and-trevor-paglen/. Acesso em: 15/9/2021.

de abordagem da mídia mudou, surgindo reportagens denunciando as decisões tendenciosas do sistema. A mudança de perspectiva decorreu de estudo e publicação da ProPublica, corporação sem fins lucrativos com sede em Nova York, dedicada ao jornalismo investigativo.

A equipe da ProPublica, liderada por Julia Angwin, empreendeu um longo percurso investigativo do Compas, na ocasião adotado não apenas em Nova York, mas na Califórnia, Wisconsin, Flórida e em mais cerca de 200 jurisdições americanas. Em abril de 2015, Angwin enviou uma *Freedom of Information Act* (Lei de Liberdade de Informação) requerendo do Broward County, Flórida, informações sobre as 18 mil pontuações Compas dos anos de 2013-14 (dados entregues cinco meses após a solicitação). Não convencidos da validade desses dados, Angwin e equipe, com a colaboração de funcionários do condado, unificaram essa base de dados com os dados de antecedentes criminais de todos os 18 mil condenados. A primeira constatação do grupo foi sobre a baixa qualidade dos registros, com inúmeros erros de digitação e ortográficos, o que por si só compromete a assertividade dos resultados. O artigo da ProPublica, "*Machine bias: there's software used across the country to predica future criminais. And it's biased against blacks*"[5], publicado em maio de 2016, sinalizava como resultados que os réus de pele escura tinham duas vezes mais probabilidade de serem classificados como de alto risco, e não reincidirem; e os réus de pele branca tinham duas vezes mais chances de serem classificados como de baixo risco, e reincidirem (Christian, 2020). A determinação exata do viés no Compas é dificultada por ser ele um sistema proprietário, mas, provavelmente, um dos fatores está na disparidade social: a composição racial e étnica das prisões dos EUA é substancialmente diferente da demografia do país. Em 2018, os negros americanos representavam 33% da população carcerária condenada, quase o triplo de sua participação de 12% na população adulta americana; os brancos representavam 30% dos presos, cerca de metade de sua participação de 63% na população adulta; e os hispânicos representavam 23% dos reclusos, em comparação com 16% da população adulta, ou seja, a população carcerácia é enviesada por raça e etnia[6]. A constatação do enviesamento da base de dados do Compas contribuiu para a visibilidade do problema.

CAMINHOS DE MITIGAÇÃO DO VIÉS

Dada a crescente visibilidade dos efeitos danosos do viés contido nas decisões automatizadas por IA, particularmente as aplicações em campos sensíveis como saúde e educação, especialistas acadêmicos e não acadêmicos, das ciências exatas e das ciências sociais, estão empenhados em encontrar abordagens para detectar e remover, ou ao menos mitigar, o viés dos sistemas de IA.

5 Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 16/9/2021.

6 Disponível em: https://www.pewresearch.org/fact-tank/2020/05/06/share-of-black-white-hispanic-americans-in-prison-2018-vs-2006/. Acesso em: 15/9/2021.

A maior parte das propostas de institutos e pesquisadores do campo das humanidades carece de viabilidade prática, conflita com a natureza e a prática de aprendizado de máquina (incluindo o caráter proprietário dos algoritmos e a complexidade dos sistemas como barreiras ao entendimento leigo). Algumas delas são: disponibilizar ao público informações de como foram desenvolvidas e implementadas as tecnologias de reconhecimento facial; criar estruturas de governança para garantir a proteção, segurança, confiabilidade e precisão dos sistemas; criar trilhas de auditoria por meio de protocolos robustos de registros de atividades, consolidados em documentação e transmitidos por relatórios públicos; esclarecer os fundamentos e resultados aos usuários afetados em linguagem não técnica (Leslie, 2020).

Sugestões para mitigar os danos dos sistemas de reconhecimento facial de Learned-Miller et al. (2020b), por exemplo, assemelham-se mais a princípios gerais e/ou "lista de desejos" do que a alternativas efetivas: constituição de bancos de dados diversificados por etnia, gênero e idade para desenvolvimento e treinamento dos algoritmos; elaboração de padrões e princípios éticos para o desenvolvimento; e legislação para proibir ou restringir certos usos desses sistemas.

Diversos documentos, inclusive a proposta de regulamentação da IA da Comissão Europeia (*Artificial Intelligence Act – AIA*, de 21/4/ 2021)[7], sugerem a constituição de órgãos reguladores responsáveis por auditar os sistemas de IA. O Ada Lovelace Institute (ALI), por exemplo, propõe a sistemática de "auditoria dos preconceitos", em que reguladores avaliam os sistemas quanto à conformidade aos regulamentos e às normas, centrada em duas etapas: "avaliação de risco algoritmo", avaliação dos potenciais danos antes do lançamento do sistema, e "avaliação de impacto algorítmico", avaliação dos efeitos pós-lançamento. No primeiro caso, os testes seriam executados pelos próprios pesquisadores por meio da metodologia de "contrafactuais" (variar um atributo mantendo os demais idênticos), mesmo reconhecendo as limitações por conta da opacidade desses sistemas ("problema de interpretabilidade" mencionado anteriormente). O ALI sugere outras abordagens de identificação de viés, como a criação de contas falsas para checar se o sistema responde. Na prática, essas propostas não se mostram factíveis.

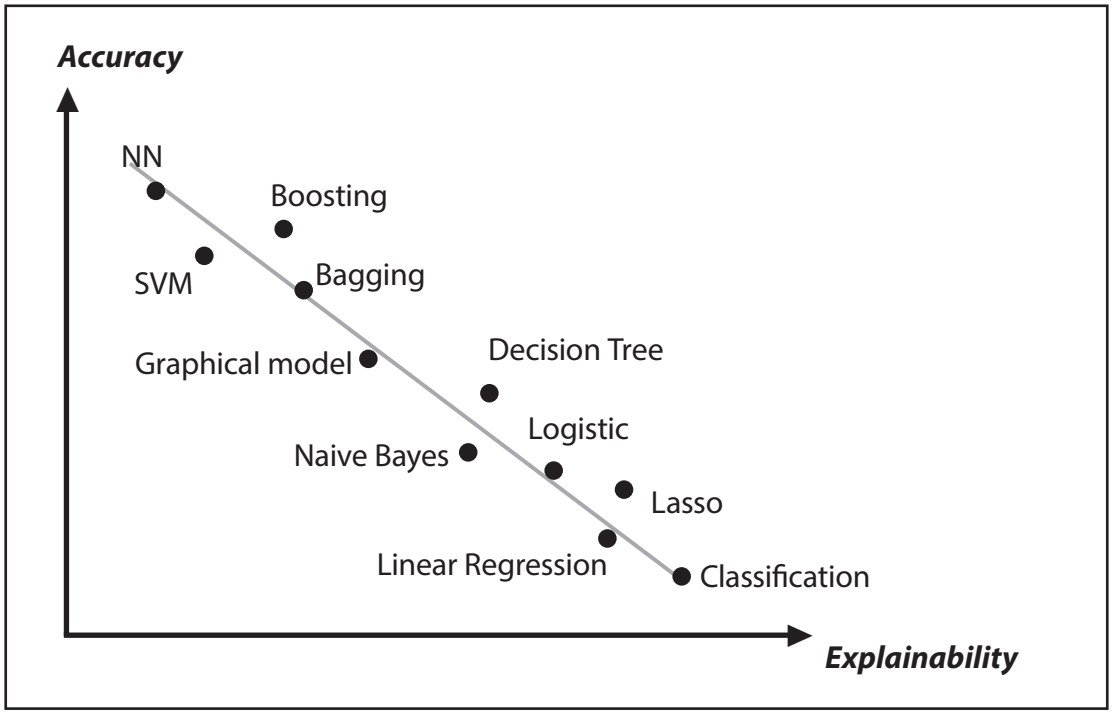
Auditoria dos sistemas de IA

O problema da auditoria tem duas abordagens: a) os métodos técnicos de identificar a origem do viés nos resultados gerados pelos modelos; e b) as barreiras operacionais. A capacidade de interpretar o modelo permite definir e incorporar os atributos/variáveis com menor potencial de gerar distorções nos resultados; adicionalmente, facilita a justificativa das decisões aos usuários diretamente afetados. Por exemplo, as DLNN utilizam grandes conjuntos de dados, ou seja, com nível de complexidade maior estabelecendo uma

7 Disponível em: https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682. Acesso em: 10/8/2021.

GRÁFICO 1

Interpretabilidade *vs* acurácia



Fonte: Explainable Artificial Intelligence (XAI); Duval (2019)

relação inversa entre os graus de interpretabilidade e acurácia (Gráfico 1).

Métodos técnicos

Existem técnicas para identificar a origem do viés nos modelos de aprendizado de máquina a partir da análise de variáveis iniciais e hiperparâmetros e, a partir da identificação, mitigar seus efeitos. Várias dessas técnicas são disponibilizadas por plataformas de tecnologia tais como a interface interativa do Google "What if tool" (Wexler et al., 2019), que gera gráficos correlacionando variáveis e viés, e a da IBM "AIF-360" (Bellamy et al., 2019), que identifica e mitiga o viés. Essas técnicas de interpretabilidade, mesmo não sendo inteiramente assertivas, ampliam o grau de confiança dos usuários e dos afetados pelos modelo ao promover um entendimento sobre o comportamento e a influência dos atributos.

Não existe, atualmente, *benchmark* consistente que permita comparar o grau de eficiência das técnicas de interpretabilidade e mitigação do viés. Uma técnica de interpretabilidadde externa ao modelo e bastante aceita como referência é a denominada SHAP (*SHapley Additive exPlanations*) (Lundberg & Lee, 2017). Baseada na teoria dos jogos cooperativos, a técnica SHAP calcula a contribuição de cada atributo no resultado preditivo gerado pelo modelo. SHAP interpreta a contribui-

ção de variáveis individualmente, ou seja, estima o efeito das interações de atributos separadamente, além de avaliar o modelo em sua totalidade. Na verdade, SHAP não é uma única técnica, mas um conjunto de técnicas em que cada uma delas tem distintos níveis de adequação a distintos modelos de IA. A visualização dos resultados de SHAP é relativamente fácil de interpretar pelos usuários, a técnica produz gráficos intuitivos (Cesaro, 2021).

Outra forma de avaliar o viés dos modelos é por meio da base de dados, como ilustrado no sistema Compas, em que a etnia do condenado é a variável sensível do modelo (ver "Viés nos dados de treinamento dos algoritmos", acima).

Barreiras operacionais

No caso de auditoria privada, Alfred Ng, jornalista especializado em privacidade e vigilância, indica a falibilidade dessa opção por meio de um caso real. A empresa HireVue, especializada em modelos de IA para auxiliar no processo de contratação, em face dos constantes escrutínios e denúncias de viés em seus sistemas, contratou a auditoria da empresa de Cathy O'Neil, autora do livro *Weapons of math destruction*: não foram evidenciados problemas nos sistemas nem lacuna entre a promessa contratada e o efetivamente entregue, o que resultou em legitimação dos modelos da HireVue. Ng pondera, contudo, que o próprio resultado da auditoria pode estar enviesado, pela ausência de padrões que definam o que seria uma auditoria de qualidade. Ademais, a falta de transparência da auditoria (O'Neil se recusou a dar detalhes do processo) tem o potencial de transformá-la numa mera "lavagem ética"[8].

Mesmo defendendo a ideia de auditoria, tarefa que poderia caber a um órgão governamental, a um contratado terceirizado ou a uma função especialmente designada em organizações multilaterais, Mokander e Floridi (2021) apontam restrições conceituais, técnicas, econômicas, sociais, organizacionais e institucionais (Quadro 1).

Complementando as restrições e/ou desafios indicados por Mokander e Floridi (2021), ressalta-se: a) a agregação de novos dados nos sistemas baseados em aprendizado de máquina, como mencionado anteriormente, implica retreinamento dos algoritmos, gerando a necessidade de auditoria contínua; b) a velocidade e a descentralização no desenvolvimento de novos modelos/algoritmos de IA dificultam replicar o arcabouço regulatório, por exemplo, da indústria farmacêutica (concentrada em poucos produtores, mais fácil de monitorar/fiscalizar); c) os algoritmos de IA são, em geral, proprietários, ou seja, são protegidos por sigilo comercial; e d) as tecnologias de IA são sofisticadas, demandando conhecimento especializado que, em geral, escapa aos reguladores/legisladores.

CONCLUSÃO

Os resultados positivos da aplicação de aprendizado de máquina na execução de variadas tarefas, em variados setores, esti-

8 Disponível em: https://themarkup.org/ask-the-markup/2021/02/23/can-auditing-eliminate-bias-from-algorithms. Acesso em: 10/8/2021.

QUADRO 1

Restrições à auditoria como mecanismo para garantir IA confiável

Tipo	Restrições
Conceitual	Há uma falta de consenso em torno dos princípios éticos gerais. Valores normativos entram em conflito e exigem compensações. Difícil quantificar as externalidades de sistemas complexos de IA. Infalibilidade da informação perdida em meio a explicações reducionistas.
Técnica	Sistemas de IA podem ser opacos e difíceis de interpretar. Integridade e privacidade dos dados são expostas a riscos durante auditorias. Mecanismos de conformidade linear são incompatíveis com o desenvolvimento ágil de software. Testes podem não ser indicativos do comportamento dos sistemas de IA no ambiente do mundo real.
Econômica e social	Auditorias podem prejudicar ou sobrecarregar desproporcionalmente setores/grupos específicos. Garantir o alinhamento ético deve ser equilibrado com incentivos à inovação. Auditoria baseada na ética é vulnerável ao comportamento adversário. Efeitos transformadores da IA apresentam desafios sobre como acionar auditorias. Auditoria baseada na ética pode refletir e reforçar as estruturas de poder existentes.
Organizacional e institucional	Falta clareza institucional sobre quem audita quem. Auditores podem não ter acesso às informações necessárias para avaliar os sistemas de IA. Natureza global dos sistemas de IA desafia as jurisdições nacionais.

Fonte: Floridi et al. (2021)

mula sua implementação em larga escala. São mandatórias, contudo, a consciência e a vigilância da sociedade em relação às externalidades negativas desses modelos, parte explicitada no quarto volume do *The State of AI Ethics Report* do Montreal AI Ethics Institute (Maiei). O relatório sinaliza a variedade de impactos sociais dos sistemas de IA em temas como discriminação, condições de trabalho, direitos humanos, desinformação e democracia (Maiei, 2021).

Outro aspecto a ser considerado é que o desempenho dos sistemas varia do ambiente em que foram treinados e testados (base de dados de treinamento e validação) ao comportamento quando os sistemas interagem com os dados do mundo real. Como ponderam Daniel Ho et al. (2020), grande parte do treinamento desses sistemas ocorre com dados "higienizados", por exemplo, imagens bem iluminadas e frontais, enquanto no mundo real as condições de iluminação não são ideais e os fotografados não necessariamente estão olhando direto para a câmara; essas diferenças impactam fortemente o desempenho dos sistemas.

Categorizar por nível de risco, como consta da proposta de regulamentação da Comissão Europeia, é um caminho apropriado; a atenção das autoridades reguladoras deve se concentrar nos usos de IA que representam mais riscos para os indivíduos e a sociedade,

contemplando o *trade-offs* entre os riscos e benefícios a serem observados pelos desenvolvedores, usuários e reguladores.

Talvez o maior equívoco no uso dos sistemas de IA seja a "promessa de objetividade", isto é, supor que os algoritmos garantem objetividade e/ou neutralidade por serem processados por máquinas e protegidos dos erros humanos. O recomendado é considerar os sistemas de IA como parceiros dos especialistas humanos, e não soberanos. A ética da IA é sobre mitigar riscos e a abordagem não é global, como também não é possível controlar todos os desenvolvimentos e usos. É crítico eleger as prioridades e focar os maiores riscos.

REFERÊNCIAS

FALPAYDIN, E. *Machine learning*. Cambridge, MIT Press, 2016.

ANGWIN, J. et al. "Machine bias". *ProPublica*, 2016.

BARABÁSI, A.-L. *Linked: a nova ciência dos networks*. São Paulo, Leopardo Editora, 2009.

BELLAMY, R. et tal. "AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias". *IBM Journal of Research and Development*, 2019, pp. 1-15.

BUOLAMWINI, J.; GEBRU, T. "Gender shades: intersectional accuracy disparities in commercial gender classification". *Proceedings of machine learning research*, 2018.

CESARO, J. *Avaliação de discriminação em aprendizagem de máquina usando técnicas de interpretabilidade*. Dissertação de mestrado. São Paulo, Escola Politécnica da Universidade de São Paulo, 2021.

CHRISTIAN, B. *The alignment problem: machine learning and human values*. New York, W.W. Norton & Company, 2020.

CRAWFORD, K. *Atlas of AI*. New Haven/London, Yale University Press, 2021.

DOMINGOS, P. *The master algorithm: how the quest for the ultimate learning machine will remake our world*. New York, Basic Books, 2015.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. Cambridge, MIT Press, 2016.

HAO, K. "Intelligent machines: this is how AI bias really happens – and why it's so hard to fix". *MIT Technology Review*, 2019.

HO, D. et al. "How regulators can get facial recognition technology right". *Tech Stream*, 2022.

KAUFMAN, D. "Equipes interdisciplinares: não basta 'juntar campos', tem que construir pontes", *Época Negócios*, 2021.

LEARNED-MILLER, E. et al. "Facial recognition technologies in the wild: a call for a federal office". *White Paper*, 2020a.

LEARNED-MILLER, E. et al. "Facial recognition technologies". *A Primer*, 2020b.

LESLIE, D. "Understanding bias in facial recognition technologies". *Alan Turing Institute*, 2020.

LUNDBERG, S. M.; LEE, S. I. "A unified approach to interpreting model predictions". *Advances in neural information processing systems*, 2017, pp. 4.765-74.

LIAO, S. M. *Ethics of artificial intelligence*. New York, Oxford University Press, 2020.

MINSKY, M. *Communication with alien intelligence*. Cambridge, Cambridge University Press, 1985.

MITCHELL, T. M. *Machine learning*. New York, McGraw-Hill, 1997.

MOKANDER, J.; FLORIDI, L. "Ethics-based auditing to develop trustworthy ai. Minds and machines". *Springler*, 2021.

PEREZ-CRIADO, C. *Invisible women: data bias in a world designed for men*. New York, Abrams Press, 2021.

RUSSELL, S. *Human compatible: artificial intelligence and the problem of control*. New York, Viking, 2019.

WEXLER, J. et al. "The what-if tool: interactive probing of machine learning models". *IEEE Transactions on visualization and computer graphics*, 2019, pp. 56-65.