Inteligência artificial, Direito e equidade algorítmica

Discriminações sociais em modelos de *machine learning* para a tomada de decisão

RICARDO SILVEIRA RIBEIRO

Resumo: Este artigo expõe como a inteligência artificial pode apresentar vantagens quando comparada com a opção tradicional de decidir com base em julgamentos exclusivamente humanos, ainda que reproduza discriminações sociais em decisões jurídicas relevantes. Por meio da revisão dos principais problemas decorrentes do uso de modelos de classificação de risco de reincidência em processos criminais, o artigo demonstra como pesquisas empíricas recentes sugerem a superioridade de decisões baseadas em algoritmos em aplicações socialmente relevantes. Em conformidade com a literatura sobre equidade algorítmica, o artigo indica que não se deve proibir por lei o uso da tecnologia, mas simplesmente regulamentá-lo de acordo com critérios de avaliação de equidade estabelecidos previamente no âmbito do debate público nacional em cada domínio de aplicação. A complexidade dos problemas suscitados pela inteligência artificial sugere a adoção de estratégias especializadas de regulamentação bastante diferentes da utilizada no art. 20 da Lei Geral de Proteção a Dados Pessoais.

Palavras-chave: inteligência artificial; aprendizado de máquina; equidade algorítmica; Direito; discriminação social.

Artificial intelligence, Law and algorithmic fairness: social discrimination in machine learning models for decision making

Abstract: This paper presents how Artificial Intelligence, despite reproducing social discrimination in relevant legal decisions, can represent advantages when compared to the traditional option of deciding based on exclusively human judgments. From the review of the main problems arising from the use of risk classification models for recidivism in criminal

Recebido em 29/3/21
Aprovado em 10/6/22

proceedings, the paper shows how recent empirical research suggests the superiority of decisions based on algorithms in socially relevant applications. Consistent with the literature on algorithmic equity, the paper indicates that the use of technology should not be prohibited by Law, but simply regulated according to equity assessment criteria previously established within the scope of the national public debate in each domain of application. The complexity of the problems raised by Artificial Intelligence suggests the adoption of specialized regulatory strategies quite different from the one used in art. 20 of the General Personal Data Protection Act.

Keywords: artificial intelligence; machine learning; algorithmic fairness; Law; social discrimination.

1 Introdução

Em 2012, matéria de *The Wall Street Journal* revelou indícios de que *sites* de venda na internet estavam utilizando informações do código postal dos usuários para discriminar consumidores por faixa de renda: enquanto compradores de áreas mais pobres receberam ofertas de preço menos atrativas, residentes de áreas mais ricas tiveram acesso a descontos significativos para compras realizadas no mesmo *site* (VALENTINO-DEVRIES; SINGER-VINE; SOLTANI, 2012). Poucos anos depois, a imprensa internacional noticiou que um aplicativo de reconhecimento de imagens classificou, automaticamente, pessoas negras em uma foto como "gorilas" (GOOGLE…, 2015). Pesquisadores do MIT Media Lab e da Microsoft Research também detectaram que sistemas comerciais de reconhecimento facial erravam em até 34,7% quando usados na identificação de imagens de *mulheres de pele mais escura*, ao passo que o mesmo erro de classificação não passava da insignificante proporção de 0,8% no reconhecimento de imagens de *homens de pele mais clara* (BUOLAMWINI; GEBRU, 2018).

O *fio condutor* desses três casos não são apenas práticas discriminatórias de renda, raça e gênero produzidas por sistemas computacionais: a novidade é que esses *softwares* não foram projetados para fomentar desigualdades, preconceitos e estereótipos. Simplesmente seus algoritmos,[1]

[1] *Algoritmo* é um termo técnico para qualquer conjunto de instruções que se dá a um computador por meio de programação. Intuitivamente, é equivalente a uma "receita de bolo" que indica à máquina todas as tarefas a serem executadas em determinada ordem.

baseados em inteligência artificial, "aprenderam" com o meio social a reproduzir essas práticas discriminatórias de forma até então não antevista pelos responsáveis por projetos de criação e manutenção do *software*. Para compreender a razão pela qual isso pode ocorrer, precisamos entender como a Inteligência Artificial – como área da Ciência da Computação – se transformou desde seu surgimento formal no *workshop* The Dartmouth Artificial Intelligence Summer Research Project, realizado nos EUA em 1956.[2] Na origem, os cientistas esperavam que programas complexos pudessem ensinar computadores a simular tarefas tipicamente humanas (RUSSELL; NORVIG, 2016, p. 17). Os sistemas computacionais não teriam então a habilidade de aprender, por si mesmos, a mimetizar o comportamento humano; teriam que ser detalhadamente programados para isso.

Na década de 1980, contudo, fortaleceu-se uma segunda tradição de pesquisa, denominada *machine learning* (aprendizado de máquina). Nessa abordagem, há o desenvolvimento de algoritmos capazes de ensinar o computador a aprender com os dados a ele informados como *inputs* (imagens, vídeos, planilhas, textos ou sons) (GOODFELLOW; BENGIO; COURVILLE, 2016, p. 2; SEJNOWSKI, 2018, p. 40). Assim como aprendemos desde crianças a distinguir, intuitivamente e por experiência, imagens e sons de diferentes animais e pessoas, os algoritmos de *machine learning* permitem "alimentar" um sistema computacional com um número suficientemente grande de "exemplos" (*inputs*), para que ele seja "ensinado" a distinguir ou revelar padrões nos dados. Com isso, informações preciosas podem ser analisadas e sintetizadas automaticamente com o objetivo de fornecerem parâmetros objetivos para a tomada de decisão.

A despeito dessa vantagem, há um risco latente nessa abordagem: como a sociedade é permeada por desigualdades, modelos[3] de *machine learning* são também capazes de "aprender" a reproduzir políticas discriminatórias presentes no ambiente socioeconômico e, sem medidas corretivas, podem até mesmo gerar um círculo vicioso de retroalimentação, no qual o sistema de inteligência artificial contribui para o agravamento de problemas sociais.

Por exemplo, se uma grande empresa no passado adotou políticas discriminatórias na contratação de mulheres, usar dados de contratações anteriores num sistema de inteligência artificial pode fazer com que o algoritmo de *machine learning* simplesmente recomende a contratação de uma proporção maior de homens sem que os engenheiros de *software* tenham desenhado o sistema para operar de forma discriminatória. Ao contratar

[2] O primeiro trabalho sobre inteligência artificial foi o de McCulloch e Pitts (1943), mas a área de estudo somente foi fundada formalmente em 1956.

[3] Define-se *modelo* como qualquer representação *matemática* da realidade por meio de fórmulas, funções e estruturas (arquiteturas), esquemas e desenhos gráficos.

mais homens em razão do uso da inteligência artificial, os dados a serem processados pelo sistema são alimentados com novas informações sobre contratação de uma proporção maior de homens; e, futuramente, quando forem usados em novas decisões de contratação, podem levar o sistema a reforçar ainda mais o problema da discriminação de gênero. O sistema pode então amplificar o viés discriminatório de gênero da empresa sem haver sido projetado para isso. É como se a "criatura" ganhasse "vida própria" diante de seus "criadores" porque simplesmente foi capaz de "aprender" a discriminar e, nesse círculo vicioso, pouco importa o que os responsáveis pela concepção e *design* do *software* pensavam.

Será então que o risco desse novo tipo de discriminação, a *discriminação algorítmica*, indica a superioridade da decisão exclusivamente humana em questões jurídicas importantes como prender ou não prender cautelarmente um réu, conceder ou não conceder um empréstimo, contratar ou não contratar um empregado? Será que esse risco de discriminação algorítmica sugere que decisões exclusivamente humanas são tecnicamente superiores às baseadas no uso de *machine learning*, devendo a legislação simplesmente proibir o uso de inteligência artificial em processos decisórios jurídicos com grande impacto para os seres humanos?

A resposta a esses problemas está longe de ser trivial. Dúvidas como essas levaram os cientistas da computação à criação de uma nova área de pesquisa transdisciplinar, denominada *machine learning and fairness* (aprendizado de máquina e equidade)[4] ou simplesmente *algorithmic fairness* (equidade algorítmica). Nela são sugeridas técnicas e critérios para a avaliação crítica dos algoritmos e dados utilizados por *softwares* com o objetivo de propor soluções que minimizem os riscos concretos de discriminação algorítmica nas diversas aplicações dessa tecnologia. Com amparo na revisão dessa literatura, este artigo objetiva apresentar ao público da área jurídica e das Ciências Sociais em geral as principais discussões sobre equidade em sistemas computacionais. Mais especificamente, demonstrará como operam modelos de classificação em *machine learning*, indicará quais são os principais problemas decorrentes da adoção de critérios para a avaliação de equidade algorítmica e, com fundamento na apresentação de resultados de pesquisas empíricas desenvolvidas na Ciência da Computação, Psicologia e Economia, discutirá as possíveis vantagens de sistemas automatizados de decisão baseados em *machine learning*, quando comparados com o *status quo* de deixar todo o processo decisório nas mãos da percepção subjetiva humana. Esperamos, ao final, confirmar a hipótese do trabalho de que, a despeito dos problemas reportados pela literatura, o uso de *machine learning* não só representa uma vantagem em termos de acurácia, como pode até mesmo minimizar a ocorrência de discriminações sociais decorrentes de julgamentos exclusivamente humanos em aplicações relevantes para o Direito e para a sociedade em geral.

Para sustentar essa hipótese, o trabalho foi dividido em três seções. Na primeira, introduziremos conceitos básicos de *machine learning* para o público da área jurídica e das Ciências Sociais. O objetivo é fazer o leitor compreender, com alguma discussão teórica e exemplos, como funcionam modelos de classificação em inteligência artificial. A segunda seção do artigo apresentará aplicações de modelos de classificação no Direito. Utilizaremos como mote para discussão o debate norte-americano sobre equidade e avaliação do risco de reincidência

[4] Traduzimos *fairness* como "equidade", embora os cientistas da computação usem o termo de modo mais genérico, como sinônimo também de "justiça", "não discriminação" ou "correção" sob o ponto de vista ético.

de réus em ações criminais. Esse é o momento em que serão problematizados os critérios de equidade comumente utilizados pela literatura em equidade algorítmica, especialmente o *balanceamento das taxas de erro* e a *calibragem*. A terceira seção, por fim, discutirá se, diante dos problemas salientados, há alguma vantagem no uso de *machine learning* para a tomada de decisão em aplicações jurídicas. Metodologicamente, será o momento de revisar os resultados de pesquisas empíricas recentes sobre as vantagens de decisões automatizadas, ao menos quando comparadas com o julgamento exclusivamente humano.

2 Modelos de classificação em *machine learning*: treinamento, testagem, inferência e outros conceitos básicos da inteligência artificial

A compreensão da natureza dos problemas em torno do uso de técnicas de *machine learning* exige uma ideia precisa de como esses modelos funcionam. Para isso, começaremos por um exemplo didático de classificação de imagens com o uso de um algoritmo de *machine learning* baseado em "redes neurais convolucionais" (*convolutional neural networks*), um tipo de estrutura (arquitetura) de "neurônios artificiais" apropriado ao processamento de imagens apresentadas a seguir.

Suponha-se, por exemplo, que desejemos reconhecer, por meio de um computador, se a imagem de determinado animal é de um gato.[5] Em *machine learning*, isso pode ser feito se ensinarmos o sistema a identificar gatos em fotos. Para isso, temos que montar um grande arquivo, em nosso computador pessoal, preferencialmente com milhares de imagens de animais. Subdividimos então esse arquivo em duas outras pastas: a primeira, o *training data* (dados para treinamento), será usada para ensinar o sistema a distinguir gatos de outros animais durante uma *fase de treinamento*; a segunda, o *test data* (dados para testagem), servirá para testar se efetivamente o modelo matemático estimado aprendeu a fazer essa distinção numa espécie de *fase de avaliação*.[6]

Há uma forte analogia aqui com o que ocorre em nosso cotidiano escolar: primeiramente, aprendemos algum assunto por meio do estudo e, em momento posterior, fazemos testes que avaliam nossa capacidade de acertar. Nos modelos de *machine learning*, o sistema funciona inspirado nessa ideia e, durante o aprendizado, os erros são usados pelo sistema para sua própria correção. Ou seja: o sistema não só treina como também é capaz de aprender com seus erros, minimizando-os.

Para a tarefa de classificação de fotos, usaremos um banco público de imagens, o Oxford-IIIT Pet Dataset (PARKHI; VEDALDI; ZISSERMAN; JAWAHAR, [20--]).[7] Nesse banco, há 4.978 fotos digitais de cachorros e 2.371 de gatos, classificadas por raças, que serão usadas para treinarmos e testarmos a capacidade efetiva de o sistema aprender a classificar uma nova imagem como um gato. Após o treinamento e a testagem do aprendizado com o uso de imagens desse banco, obtivemos um modelo matemático

[5] Essa tarefa foi desenvolvida com a linguagem de programação Python e as bibliotecas FastAi e PyTorch, de acordo com as orientações técnicas de Howard e Gugger (2020, p. 67).

[6] Na prática, não é necessário dividir em duas pastas. Programas podem fazer esse trabalho. Poderíamos, por exemplo, reservar 90% das fotos como *training data* e as 10% restantes para *test data*. No exemplo a ser dado, usamos 30% das imagens para a testagem do aprendizado do modelo.

[7] De acordo com o *site*, o banco de imagens pode ser usado para fins comerciais ou não sob a licença da Creative Commons (Attribution-ShareAlike 4.0 International – CC BY-SA 4.0).

que permite ao sistema reconhecer se uma nova foto é de um gato ou não. Por exemplo, apresentamos a Figura 1 ao modelo matemático de *machine learning* e este classificou corretamente a imagem como a de um gato com 100% de probabilidade.

Figura 1

Gato. O modelo de *machine learning* classifica corretamente essa imagem como a de um gato (com 100% de probabilidade)

Fonte: elaborada pelo autor.

A imagem da Figura 1, contudo, é trivial. Explicitamente, estamos diante da foto de um gato. Tentamos então complicar a tarefa de

reconhecimento de imagens apresentando a nosso modelo a imagem da Figura 2. Como a face do animal não é revelada, a tarefa de classificação é um pouco mais difícil que a anterior.

Figura 2

Gato com face ocultada. O modelo de *machine learning* classifica corretamente essa imagem como a de um gato (com 99,84% de probabilidade)

Fonte: elaborada pelo autor.

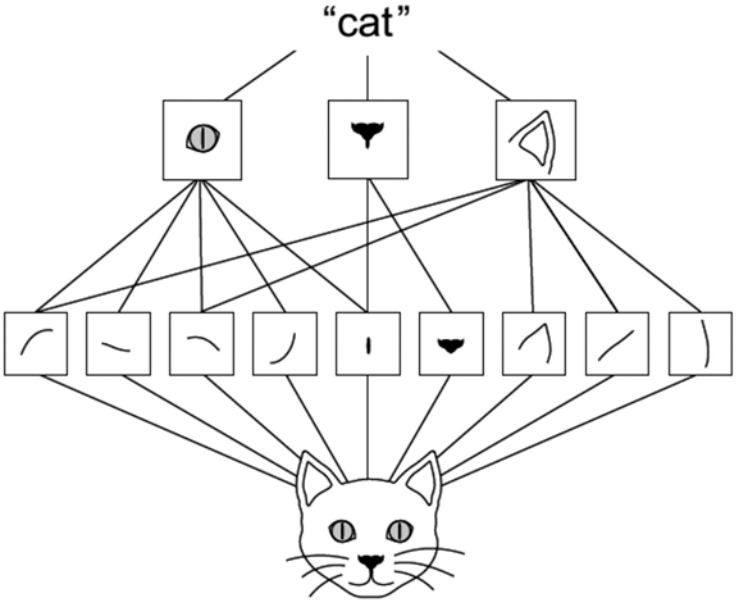
Mais uma vez, o modelo matemático classificou a imagem corretamente como a de um gato; mas, como a tarefa foi um pouco mais difícil que a anterior, a probabilidade de ser um gato foi estimada em 99,84%. A probabilidade de classificar uma nova imagem como a de um gato, portanto, caiu um pouco; ainda assim, a classificação foi correta.

Essa capacidade de aprender a classificar a imagem corretamente ocorreu graças à estrutura de *deep learning* utilizada por este autor para treinar e testar o modelo. Nessa estrutura, cada imagem dentre as milhares fotos de gato do Oxford-IIIT Pet Dataset é analisada em seus mínimos detalhes, os quais são armazenados em níveis (*layers*) com sucessivos filtros organizados em forma análoga a uma rede. Pormenores técnicos desse processo não são objeto deste artigo, mas a Figura 3, reproduzida por Chollet (2018, p. 122), dá uma ideia de como

características (*features* ou atributos) dos gatos são armazenadas nessas espécies de "neurônios" artificiais.

Figura 3

Rede neural com *deep learning*



Fonte: Chollet (2018, p. 123).

Esse tipo de estrutura possibilita a identificação de padrões nas imagens, tais quais linhas, curvas, focinhos, vibrissas, orelhas, patas, suas possíveis cores e gradientes. Em *deep learning*, esses padrões são reconhecidos automaticamente pelo modelo, permitindo que o sistema represente desde características mais concretas (ex.: um pequeno pedaço da garra, uma vibrissa) até conceitos mais abstratos (ex.: expressões faciais de medo ou de raiva). Os níveis (*layers*) que armazenam informações sobre características de determinado objeto são construídos automaticamente pelo próprio modelo e este pode ainda identificar estruturas, relações ou padrões jamais observados por seres humanos, assim como ignorar informações irrelevantes para a tarefa de classificação (LECUN; BENGIO; HINTON, 2015).

A despeito da complexidade dessa tecnologia, o modelo pode falhar por diferentes razões. Para demonstrarmos tal fato, dificultamos ainda mais a tarefa do modelo e apresentamos um desenho infantil de gato na Figura 4.

Figura 4

Desenho infantil de um gato. O modelo de *machine learning*, incorretamente, não reconhece essa imagem como a de um gato. Estima ser muito pequena a probabilidade de ela representar um gato: apenas 4,4%

Fonte: elaborada pelo autor.

Desta vez, o modelo de *machine learning* não conseguiu identificar que se tratava de um gato. Simplesmente classificou como de apenas 4,4% a probabilidade de essa foto ser de um gato e, com isso, concluiu incorretamente que a imagem não seria de um gato. A pergunta é: por quê?

Como vimos, nosso modelo aprendeu a distinguir gatos de outros animais por meio de *inputs* de um banco de dados com imagens reais de gatos e cachorros. Como o modelo não foi treinado e testado em desenhos infantis, não pôde aprender a identificar adequadamente gatos em desenhos elaborados por crianças. Assim, se treinarmos modelos de *machine learning* com *inputs* de outro tipo ou de má qualidade, a capacidade de classificar novas informações, denominada tecnicamente de *inferência*, será muito ruim ou ao menos limitada ao que foi aprendido. Existe ainda outra razão de ordem técnica para o erro: em nosso modelo, a decisão por classificar a imagem como de um gato foi diretamente influenciada pela probabilidade de a imagem corresponder ao que foi aprendido a partir do treinamento e testagem em imagens reais de cachorros e gatos. Assim, se o modelo estimar como *alta* a probabilidade de a nova imagem corresponder a um gato, concluirá que a imagem é de um gato; se, por outro lado, a probabilidade de ser de um gato for estimada como muito *baixa*, o modelo concluirá que a nova imagem não é de um gato.

Esse tipo de problema em que se procura obter uma classificação conclusiva (ex.: é um gato ou não?) com base em um número (ex.: 95% de probabilidade) é um dos mais comuns em *machine learning*, e os modelos que procuram resolvê-lo recebem o nome intuitivo de *modelos de classificação* (MURPHY, 2012, p. 3). Num modelo de classificação, verifica-se, por meio de um escore (um número qualquer), se o objeto a ser classificado (ex.: foto) pertence ou não a determinada *categoria* (*label* ou *target*) (ex.: ser um gato). Se o número indicado sugerir a maior probabilidade de o objeto pertencer à categoria, isso significa que o modelo deve classificar esse objeto dentro dessa categoria (ex.: a foto é de um gato com 80% de probabilidade de acerto).

Possível exemplo de modelos de classificação é o uso de inteligência artificial para decidir sobre concessão de empréstimos a pessoas físicas. Imagine-se que um banco deseja avaliar se deve ou não emprestar dinheiro a um cliente e, para isso, usa suas informações pessoais como *histórico de pagamento de contas*, *idade*, *renda mensal* e *padrões de consumo* observados em compras de cartão de crédito para, por meio de um modelo de *machine learning*, decidir se empresta ou não alta quantia ao cliente. Nesse caso, o sistema estimará o risco de inadimplência do cliente para que o gerente tome a decisão de emprestar ou não com base na probabilidade estimada de o cliente pagar ou não pagar o empréstimo no futuro. Se a probabilidade de inadimplência (escore) for calculada como alta, o sistema classificará o cliente como "alto risco" de inadimplência; se for baixa, classificará o cliente como "baixo risco". Com base nesse número estimado, o gerente decidirá se emprestará ou não o dinheiro. Outro exemplo seria o uso de informações pessoais de candidatos a vagas de emprego para estimar se serão bons empregados ou não no futuro. Se o sistema estimar que o empregado terá "alta produtividade", a seção de recursos humanos da empresa poderá usar esse escore para selecionar o candidato para uma entrevista de emprego; do contrário, jamais o contratará.

Em suma, nesses dois exemplos, modelos de classificação são usados para orientar seres humanos a tomarem decisões com base num escore que avalia o risco de inadimplência de um cliente ou a probabilidade de um candidato a emprego ser produtivo futuramente no trabalho. Aparentemente, como esses cálculos são baseados em informações objetivas dos próprios clientes ou dos candidatos a emprego, evitariam que o banco ou a empresa de vendas discriminassem pessoas. A lógica subjacente ao mundo real, contudo, não é tão simples, pois as informações (*inputs*) utilizadas para alimentar sistemas de inteligência artificial podem estar associadas a estereótipos, preconceitos e políticas discriminatórias latentes na sociedade, que podem ser refletidas pelo modelo de *machine learning* numa espécie de "espelhamento" dos problemas sociais.

3 Inteligência artificial como espelho da sociedade: sistemas de classificação do risco de reincidência em processos criminais e critérios de equidade

Nesta seção, discutiremos como modelos de classificação baseados em *machine learning* podem refletir vieses sociais não antevistos pelos programadores e analistas de dados. Para ilustrarmos a discussão, tomaremos, como ponto de partida, o debate norte-americano sobre a adoção de sistemas de classificação de risco de acusados em processos criminais.

Há muitos anos, o Poder Judiciário em vários estados dos EUA tem usado um *software*

comercial da empresa Northpointe para classificar o risco de réus reincidirem, ou seja, cometerem novos crimes.[8] Com um modelo de *machine learning*, o sistema denominado Compas[9] avalia o risco de reincidência de determinado réu, sugerindo um escore que, após determinado limiar (*threshold*), classifica se o acusado tem alto risco de cometer um novo delito nos próximos dois anos. O juiz envolvido no caso, por sua vez, com amparo na classificação de risco, decide se o réu merece ser preso ou se deveria ser fixada uma pena alternativa em lugar da simples prisão. A lógica de tal decisão é simples: se o réu for classificado como de "alto risco" de reincidência, o juiz tende a mantê-lo preso; se o risco for baixo, pode soltá-lo e fixar uma pena alternativa pecuniária.

Em março de 2015, a empresa responsável pelo desenvolvimento e manutenção do Compas informava que a avaliação de risco de reincidência baseava-se em variáveis que mensuravam características como idade do réu, idade quando da primeira prisão, *número de passagens pela prisão* e violação de liberdade condicional anterior. O *software* também permitia a produção de várias outras avaliações de risco baseadas em diferentes teorias criminológicas (NORTHPOINTE, 2015, p. 28, 32). Em maio de 2016, uma organização não governamental dedicada ao jornalismo investigativo, a ProPublica, divulgou uma análise crítica dos dados de avaliação de risco produzidos pelo Compas e chegou à conclusão de que réus negros tinham *duas vezes mais chances* de serem incorretamente classificados como de alto risco quando comparados com *réus brancos* em circunstâncias análogas (ANGWIN; LARSON; MATTU; KIRCHNER, 2016; LARSON; MATTU; KIRCHNER; ANGWIN, 2016). Isso aumentaria, segundo a ProPublica, a probabilidade de os juízes, ao usarem essa informação de avaliação de risco, determinarem a prisão de negros em maior proporção que de brancos, mesmo que ambos os grupos possuam perfis criminais semelhantes. Como consequência, o modelo do Compas poderia fazer com que a discriminação se perpetuasse e até se ampliasse, pois o uso dos escores como elemento de informação para a tomada de decisão tenderia a fomentar novas prisões indevidas de negros no futuro em um círculo vicioso que a literatura denomina *feedback loop*. Se esse fosse o caso, "quanto mais o modelo é usado, mais enviesados os dados se tornam, fazendo o modelo mais enviesado ainda e assim por diante" (HOWARD; GUGGER, 2020, p. 83, tradução nossa).[10]

---

[8] O Compas hoje está sob a responsabilidade da empresa Equivant (NORTHPOINTE…, *c*2022).

[9] *Compas* significa *Correctional Offender Management Profiling for Alternative Sanctions* (em português, Gestão de Perfis de Infratores para Fixação de Penas Alternativas).

[10] No original: "the more the model is used, the more biased the data becomes, making the model even more biased, and so forth".

Para diversos críticos, esses problemas e riscos latentes sugerem que a inteligência artificial não cumprira a promessa de ser objetiva no trato de questões sociais relevantes (DONEDA; MENDES; SOUZA; ANDRADE, 2018) e poderia ainda ser utilizada a serviço de narrativas e práticas de opressão contra grupos vulneráveis, minorias étnicas e pessoas mais pobres, o que reforçaria a necessidade urgente de sua contenção e regulação (BENJAMIN, 2019; CRAWFORD, 2021; O'NEIL, 2016). Não é simples, contudo, apontar se há ou não discriminação deliberada num sistema concreto como o Compas, dado que ele não foi desenhado para ter aversão a negros. Um exemplo mais próximo da realidade brasileira examinará esse ponto.

Imagine-se que, numa ação criminal, o Ministério Público requeira ao juiz a prisão cautelar de um réu sob o argumento de que existe risco concreto de reiteração de práticas delituosas violentas pelo acusado, como um novo homicídio ou latrocínio. Com amparo nos arts. 312 e 313 do Código de Processo Penal (CPP), o juiz então terá que decidir se decreta a prisão cautelar ou permite que o réu responda em liberdade. Normalmente, tais decisões são tomadas com base numa análise judicial idiossincrática das circunstâncias do caso. Suponha-se então que o Poder Judiciário use um *software* de avaliação do "perigo gerado pelo estado de liberdade do imputado" (art. 312, *caput*, do CPP (BRASIL, [2022a])) para analisar se o acusado é de alto risco ou não.

De acordo com esse sistema, os acusados podem receber uma espécie de "nota", o escore, que avalia de 0% a 100% sua probabilidade de voltar a praticar delitos. Acusados que recebam uma "nota" de 0% a 40% seriam considerados como de *baixo risco* de praticar novos delitos; acusados que recebam escores superiores ao limiar (*threshold*) de 80%, por outro lado, representam *alto risco* de nova delinquência. Avaliações intermediárias, com "notas" superiores a 40% e inferiores a 80%, significam que os acusados teriam um *risco moderado* de voltar a delinquir. Se o acusado do nosso exemplo recebeu um escore superior ao limiar, como 90%, isso significa dizer que foi considerado como de "alto risco" pelo sistema de avaliação e, nesse caso, se o juiz levar em consideração esse escore, decretará a prisão do acusado em atenção ao elevado risco sugerido pelo sistema. A pergunta é: se o acusado for negro, o sistema discriminou o acusado nesse caso concreto? Ao analisar a situação concreta de réu, o algoritmo foi "racista"?

A resposta a essa pergunta não é simples e depende do critério de equidade utilizado para avaliar a decisão. No mundo ideal, o algoritmo de *machine learning* seria capaz de sempre acertar suas análises de risco, classificando como de "alto risco" os acusados extremamente perigosos e como de "moderado" ou "baixo risco" os demais acusados. No mundo real, contudo, assim como humanos, algoritmos podem produzir dois tipos de erro: (1) classificar como de *alto risco* (escore acima do limiar) um acusado *não perigoso*; e/ou (2) classificar como de *moderado ou baixo risco* (escore menor ou igual ao limiar) um acusado *perigoso*. O primeiro tipo de erro é denominado *falso positivo* e o segundo é conhecido pelo nome *falso negativo* (KRZANOWSKI; HAND, 2009, p. 8-9).

No debate ProPublica, a organização não governamental denunciou que o sistema seria propenso a classificar negros como réus de maior risco quando comparados com os brancos. Por outras palavras, a discriminação ocorreria porque a probabilidade de um negro ser classificado incorretamente como de maior risco seria duas vezes superior à probabilidade de um branco ser considerado de maior risco em condições análogas. O sistema, portanto, teria uma proporção de falsos positivos superior para negros, errando mais ao classificá-los.

Seguindo essa lógica no nosso exemplo da determinação de uma prisão cautelar no Brasil, a discriminação racial ocorreria se o sistema classificasse incorretamente uma proporção maior de negros quando comparados com a de brancos, pois teria sido violado um critério de equidade denominado *balanceamento das taxas de erro* (*error rate balance*). Nesse balanceamento, exige-se que a proporção de erros de classificação dos grupos dominante e vulnerável seja a mesma (VERMA; RUBIN, 2018). Se o sistema erra mais na classificação de risco de negros, quando comparada com a classificação de risco dos brancos, viola-se esse critério de equidade de acordo com este quadro geral de falhas:[11]

> (1) *maior taxa de falsos positivos dos negros em relação aos brancos*, com o efeito indesejável de uma proporção maior de prisões cautelares de acusados negros classificados *incorretamente* como perigosos, quando *de fato* não seriam de alto risco; e
>
> (2) *menor taxa de falsos negativos dos negros em relação aos brancos*, ou seja, uma proporção maior de brancos livra-se da prisão por ter sido classificada *incorretamente* como de baixa periculosidade, quando, em verdade, seria de alto risco.

A despeito da clareza dessas consequências díspares (*disparate impact*) para ambos os grupos, é extremamente difícil operar a utilização do critério de balanceamento das taxas de erro, pois para isso o sistema ou os seres humanos teriam que saber, no momento da tomada de decisão, se o acusado avaliado cometeria ou não um crime no futuro. Isso não é possível, pois se o acusado for preso, *não haverá como testar seu perigo real para a sociedade*, sendo impossível avaliar se cometeria ou não um delito no futuro, isto é, se seria um falso positivo; e, se o acusado for solto, poderá cometer novos crimes sem que o aparato repressivo do Estado consiga identificar a autoria do delito nem se o crime existiu. Também é muito difícil calcular a taxa de falsos negativos. Nem o sistema nem o responsável pela tomada da decisão, o juiz, poderão antever como o réu se comportaria no futuro se não houvesse sido preso ou como efetivamente se comportou se foi solto, isto é, se efetivamente não cometeu crimes. Nesse caso, as taxas de falsos positivos e falsos negativos somente podem ser estimadas muito indiretamente por meio de técnicas estatísticas complexas executadas muito tempo após o momento da tomada de decisão e, ainda assim, esses cálculos servem apenas para uma análise crítica de como o fluxo de justiça vem tratando os acusados historicamente; mas

[11] Com o objetivo de ser mais didática para o público da área jurídica, a definição de *balanceamento das taxas de erro* foi apresentada de forma intuitiva nesta seção. No apêndice, apresenta-se a definição formal para públicos das demais áreas, sobretudo da Ciência da Computação, da Economia e da Ciência Política.

não é viável usá-los cotidianamente no momento da tomada de decisão, quando realmente poderia ser de extrema relevância.

Em resposta às críticas da ProPublica, a empresa responsável pelo *software* lançou um relatório refutando a alegação de viés racial do sistema Compas e, de modo geral, argumentou que as diferenças entre brancos e negros nas probabilidades de serem erroneamente classificados como de alto risco ocorreram porque a proporção de infrações é maior na população negra, o que sugere que as diferenças de avaliação de risco são explicáveis por *diferenças estatísticas reais entre os dois grupos* (DIETERICH; MENDOZA; BRENNAN, 2016, p. 8). Essas diferenças entre "brancos" e "negros", portanto, não foram "fabricadas" artificialmente pelo sistema de *machine learning*; seriam uma simples decorrência de ser prevalente o encarceramento de negros nos EUA. Cor ou raça, portanto, não foram explicitamente levadas em consideração na predição do risco de reincidência; mas, como negros tendem a ter uma incidência de prisão mais alta que brancos no universo de sua população, o sistema "espelhou" tais diferenças sociais em seus escores e capturou a discriminação racial pré-existente na sociedade norte-americana.[12]

Análises estatísticas posteriores também detectaram que não haveria viés contra negros no sistema Compas, ao menos quando analisado à luz de critérios de equidade tradicionalmente reportados pela literatura. Flores, Bechtel e Lowenkamp (2016), por meio de várias técnicas estatísticas, não encontraram viés contra negros nas estimativas de probabilidade de reincidência do sistema, pois, para qualquer valor dos escores de avaliação de risco, *demonstraram que tanto brancos como negros tiveram uma probabilidade semelhante de reincidência*. Isso significa dizer que o sistema de *machine learning* obedecia ao critério de equidade denominado *calibragem* (*calibration*).[13]

Não é difícil perceber que, a depender do critério de equidade utilizado, é possível obter conclusões bem diferentes sobre eventuais práticas discriminatórias do sistema. Assim, não se registram vencedores e vencidos no debate iniciado pela ProPublica em 2016, pois as grandes divergências

---

[12] Pode-se defender que, para solucionar esse problema da maior proporção de prisões sobre a população negra, seria necessário um limiar diferente para brancos e negros no sistema Compas. Usando nosso exemplo da prisão cautelar, poder-se-ia definir que os negros fossem considerados de alto risco somente se tivessem uma "nota" maior que 90; em contrapartida, para os brancos, bastaria que tivessem um escore de 80 para os classificar como de alto risco. Não é difícil perceber um dilema jurídico forte nessa solução, pois, para corrigir as distorções, seria necessário ser mais duro com os brancos, baseando-se de forma explícita no critério racial. Perceba-se que aqui não é a sociedade que cria a discriminação, mas sim o próprio algoritmo. Usando dados do Compas, Corbett-Davies, Pierson, Feller, Goel e Huq (2017) chegaram à conclusão de que diferentes limiares por raça levam a uma maior prisão de brancos e maior soltura de negros que em tese teriam o mesmo risco de reincidência, ou seja, teriam o mesmo perfil criminológico.

[13] A definição formal de *calibragem* está no apêndice deste artigo.

nas conclusões dos autores dos diferentes estudos dependem do *critério de equidade* utilizado como parâmetro de avaliação crítica. Como existem muitos, mais de vinte na literatura (ŽLIOBAITĖ, 2017; VERMA; RUBIN, 2018; HUTCHINSON; MITCHELL, 2019; CATON; HAAS, 2020), são esperadas conclusões extremamente contraditórias a depender do modo como se mensura a discriminação entre grupos dominantes e vulneráveis. Não existe, portanto, um só critério de aferição de equidade que possa ser utilizado de forma consensual em qualquer domínio social ou econômico e, ainda que o sistema "passe" em alguns testes de equidade, jamais se deve interpretar que foi expedido um "certificado de que o sistema é justo", dadas as diferentes conclusões a que se pode chegar a depender do critério de avaliação adotado (BAROCAS; HARDT; NARAYANAN, 2019, p. 121, tradução nossa).[14]

Como complicador, Kleinberg, Mullainathan e Raghavan (2016) e Chouldechova (2017) demonstraram matematicamente que, ressalvadas situações muito improváveis, os critérios de equidade são inconciliáveis, isto é, apresentam *trade-offs*. Em linguagem simples, isso significa que se o sistema satisfaz aos critérios de calibragem, *é impossível* cumprir o requisito do balanceamento das taxas de erro; se o sistema gera um balanceamento das taxas de erro, será impossível cumprir o critério da calibragem. Em suma, os critérios não são apenas diferentes: são incompatíveis. Seria, pois, impossível criar um sistema que satisfizesse concomitantemente a todos os critérios de equidade e, por essa razão, quando se avalia se um sistema computacional cumpre requisitos de equidade, há sempre a necessidade de explicitarmos previamente na metodologia o critério utilizado para auditar criticamente a equidade algorítmica, pois o uso de outros critérios de avaliação poderá levar a conclusões bem diferentes. Consequentemente, a rigor não faz sentido afirmar que um sistema é ontologicamente predestinado a violar critérios de equidade em geral.

Por outro lado, não parece haver dissenso sobre o quanto os dados informados aos sistemas computacionais podem explícita ou implicitamente afetar, de modo negativo, modelos de *machine learning*. Sem dúvida, os dados sobre as proporções de encarceramento nos EUA indicam maior repressão sobre a população negra e, embora as causas possam ser objeto de profunda controvérsia, devemos esperar que o sistema "espelhe", por meio de uma análise algorítmica, essa estrutura discriminatória subjacente. Como o Departamento de Justiça dos EUA admitiu formalmente (NATIONAL INSTITUTE OF CORRECTIONS, 2017, p. 18), as avaliações de risco não têm como oferecer uma solução para problemas como maior proporção no encarceramento de minorias. Elas apenas mapeiam o risco real de novas prisões por reincidência, mas não têm como coibir uma grande incidência de repressão policial sobre comunidades negras. É possível ainda que brancos cometam uma proporção superior de crimes de outro tipo não identificados pelo aparato repressivo do Estado justamente porque a polícia tem um viés contra negros em áreas geográficas segregadas.

Acrescente-se a esses problemas a impossibilidade de saber se o sistema foi discriminatório com um acusado específico – o indivíduo. Normalmente os critérios de equidade definidos pela literatura preocupam-se com a discriminação decorrente do pertencimento a um grupo (ex.: latinos, LGBTQ+, mulheres). Uma avaliação de risco como a do Compas ou do nosso exemplo da prisão cautelar, entretanto, classifica determinado acusado ao avaliar o comportamento estatístico passado de outros réus e de grupos em centenas ou, preferencialmente, milhares de

[14] No original: "if a system passes a fairness test, not interpret it as a certificate that the system is fair".

casos criminais. Ao reportar uma avaliação de risco, o sistema produz uma representação estatística do réu, ou seja, uma representação de como a média de acusados com as características pessoais do indivíduo avaliado (ex.: idade, passagem anterior pela polícia, tipo de crime supostamente cometido) se comportou no meio social, isto é, se voltou ou não a delinquir. A informação não é uma representação exata de uma pessoa física concreta, mas um modelo estatístico idealizado do ser humano avaliado.

Diante de todos esses problemas, o uso de modelos de *machine learning* teria alguma vantagem? Deveríamos insistir no uso da inteligência artificial ou simplesmente rejeitar a tecnologia em benefício da opção *default*, o *status quo*, de somente nos valermos do julgamento humano em processos decisórios importantes para o Direito e para a sociedade em geral?

4 Julgamento humano vs. *machine learning*: razões para apostar na futura automatização do processo decisório em aplicações jurídicas

As limitações apontadas anteriormente podem levar críticos a refutar o uso de *machine learning* em favor exclusivo do julgamento humano em qualquer circunstância. Poder-se-ia concluir que o caminho a seguir, para combater os problemas salientados na seção anterior, seria não usar processos automatizados em decisões jurídicas relevantes em razão dos riscos latentes de "espelhamento" dos preconceitos, vieses e discriminações nos resultados produzidos pelos sistemas computacionais. Os riscos de erro e o medo podem sugerir que se deveria evitar o uso dessa tecnologia na tomada de decisões jurídicas com consequências relevantes para os seres humanos.[15] A alternativa ao uso de algoritmos, entretanto, é simplesmente manter o *status quo* de continuar a conviver com dois outros tipos de risco com os quais lidamos há milhares de anos em nosso cotidiano – o risco da *decisão pouco refletida* (baseada no "Sistema 1"[16]) e o risco da *decisão inconsistente*.

Sobre o primeiro, é comum acharmos que a discriminação é apenas consequência da vontade deliberada de tratar desigualmente determinado grupo. Nesse tipo de abordagem, quem discrimina teria uma espécie de

[15] Ver, por exemplo, a abordagem do recente documentário *Coded Bias*, de 2020, disponível na plataforma Netflix (CODED…, 2020).

[16] Na linguagem da Psicologia Experimental e da Economia Comportamental, a *decisão pouco refletida* é movida pelo "Sistema 1" de nosso modo de pensar (KAHNEMAN, 2012, p. 29). Esse tipo de sistema mental é responsável por um raciocínio automatizado, impulsivo ou com baixo nível reflexivo. Em contraposição, como seres humanos temos também um "Sistema 2" em nosso cérebro, responsável por uma forma de pensar baseada em decisões analíticas bem refletidas e ponderadas.

"gosto por discriminação" (BECKER, 1971, p. 16) que o levaria a discriminar para assegurar preferências individuais ou de grupo. Pesquisa de Adida, Laitin e Valfort (2016) sobre a discriminação contra imigrantes muçulmanos na França exemplifica esse tipo de discriminação deliberada: em interações sociais entre franceses e imigrantes muçulmanos nas contratações de mão de obra, equipes de recursos humanos das empresas preferem contratar trabalhadores cristãos para não terem que lidar com dificuldades e custos decorrentes da prática religiosa cotidiana dos muçulmanos no ambiente da empresa (ex.: jejum durante o Ramadã, uso de véu por mulheres, orações frequentes e rituais durante o expediente de trabalho).

Entretanto, nessa mesma pesquisa, experimentos sociais, entrevistas etnográficas e análises estatísticas revelaram que, em situações corriqueiras, muçulmanos são vistos como ameaça de forma gratuita pelo cidadão francês, havendo vários componentes "irracionais" de discriminação que não são claramente percebidos pelos atores envolvidos nesses processos sociais (ADIDA; LAITIN; VALFORT, 2016, cap. 6-7). Isso significa que um fator como "raça" pode influenciar a sentença condenatória em processos criminais, fazendo com que negros recebam penas mais duras quando comparados com brancos em formas mais sutis de racismo decorrentes da presença de estereótipos não explicitados (SWEENEY; HANEY, 1992) ou sequer racionalizáveis. O juiz pode, então, decidir de modo discriminatório sem que tenha propriamente consciência de que as diferenças raciais influenciaram seu julgamento subjetivo.

Quanto ao segundo risco, o da inconsistência decisória, é comum a crença de que uma mesma pessoa, autoridade ou órgão deveria tomar decisões aproximadamente idênticas ao julgar casos análogos sucessivamente ao longo do tempo. A despeito desse ideal normativo, é possível que o mesmo responsável pela decisão trate situações análogas de modo completamente diferentes, como o faria um professor que pontuasse alunos com notas bem distintas em provas com respostas análogas ou o juiz que, para situações semelhantes, decidisse de modo diferente.

Em pesquisa de ampla repercussão na Ciência Política, na Psicologia Experimental e na Economia Comportamental, Danziger, Levav e Avnaim-Pesso (2011) constataram que a probabilidade de decisões favoráveis aos réus presos caía abruptamente após um longo período de trabalho e em momento imediatamente anterior à refeição. De modo surpreendente, após a refeição a probabilidade de decisões favoráveis voltava a subir radicalmente para seu patamar esperado e, de acordo com os achados da pesquisa, esses fatores fisiológicos influenciaram decisões judiciais extremamente relevantes sobre a liberdade dos réus num processo que com muita dificuldade se pode considerar "racional".

Num primeiro olhar, talvez o leitor não perceba a relevância desse tipo de pesquisa, mas poderíamos explicitar os resultados de outra forma menos sutil: sem usar a terminologia jurídica, os autores indicam que fome, cansaço e humor figuram como uma espécie de fonte "não formal" do Direito, pois afetam a produção da norma jurídica de decisão num caso concreto, mesmo em tema socialmente relevante como a prisão de um acusado.

Diferentemente dos julgamentos humanos, algoritmos de *machine learning* não têm "gosto por discriminação", não precisam ser influenciados por estereótipos e seus resultados não são afetados por limitações fisiológicas ou psíquicas. Têm ainda a vantagem de poderem ser desenhados para explicitar com precisão o processo decisório subjacente (KEARNS; ROTH, 2020, p. 191), que jamais é de todo revelado por decisões humanas.

Ademais, a literatura registra que modelos automatizados têm maior capacidade de corretamente classificar um objeto como *verdadeiro positivo* e/ou *verdadeiro negativo* em ambientes permeados por incerteza, imprevisibilidade e complexidade. Kleinberg, Lakkaraju, Leskovec, Ludwig e Mullainathan (2017), ao analisarem decisões judiciais de 758.027 acusados presos na cidade de Nova Iorque entre 2008 e 2013, chegaram à conclusão de que um modelo de *machine learning* diminuiria significativamente erros de classificação de risco dos acusados pelos juízes. Os autores estimaram que, se tivesse sido utilizado um modelo de inteligência artificial adequado à classificação de risco, a taxa de encarceramento poderia ter caído 42%, ou seja, poderia haver um encarceramento muito menor de seres humanos em razão da capacidade de os modelos identificarem corretamente acusados de maior risco.

Estudos em áreas tão distintas como diagnósticos de doenças, esportes, previsão de comportamentos violentos e preços futuros de vinhos também confirmam que modelos desse tipo são dotados de acurácia igual ou frequentemente superior ao julgamento humano tradicional (MEEHL, 1954; GROVE; ZALD; LEBOW; SNITZ; NELSON, 2000; KAHNEMAN, 2012, p. 278; TETLOCK; GARDNER, 2015, p. 27), com a vantagem de podermos aperfeiçoar constantemente algoritmos e análise de dados para desempenharem melhor seus papéis.

Embora não possam ser estendidos automaticamente a todos os domínios da vida social ainda não testados, esses resultados servem como alerta de que, para determinadas atividades e decisões, modelos computacionais podem ter desempenho superior ao humano e, como a discriminação social contém nuances difíceis de aferir, algoritmos podem ter um papel relevante nesse processo, pois sujeitam-se a processos contínuos de correção, desenvolvimento e auditoria para não refletirem vieses sociais.

Exemplos: um modelo de reconhecimento facial de criminosos que erra mais ao identificar peles negras pode ser treinado em tonalidades de cor mais escura com o objetivo de diminuir erros de classificação; bancos de dados de contratação de empregados que reflitam vieses de gênero podem ser devidamente tratados para suprimir preconceitos contra a diversidade de gênero. Algoritmos mais complexos podem ser até mesmo desenvolvidos especificamente com essa finalidade. Zemel, Wu, Swersky, Pitassi e Dwork (2013) conseguiram criar um mecanismo de otimização capaz de ofuscar discriminações contra grupos vulneráveis em algoritmos de *machine learning*. Nesse caso, o algoritmo cria uma representação dos dados que remove a influência de variáveis sensíveis, como raça, cor, gênero e classe social. Estritamente, é como se, na tomada de decisão ou classificação, o modelo de *machine learning* aprendesse a corrigir discriminações sociais presentes na própria sociedade.

Podemos ainda usar, subsidiária ou complementarmente, o julgamento humano em conjunto com essas análises em *sistemas computacionais híbridos*. Seria possível, por exemplo, adotar uma estimativa de risco do acusado como a produzida pelo sistema Compas em conjunto com uma estimativa de risco baseada no julgamento humano de um perito judicial ou mesmo de um servidor treinado nesse tipo de avaliação. Ao decidir, o juiz poderia analisar as duas avaliações com o objetivo de fundamentar melhor sua decisão final, e divergências muito grandes de percepção de risco entre o sistema de inteligência artificial e o perito judicial poderiam ser uma sinalização, para o magistrado, de que ele estaria diante de um caso atípico, que mereceria maior atenção na avaliação de risco para a ordem pública. Avaliações similares, por outro lado, poderiam reforçar a convicção do magistrado de que uma decisão compatível com as sugestões do sistema e do perito estariam no rumo certo. Em modelos de inteligência artificial, poder-se-ia ainda adotar um banco de dados com critérios estatísticos e de avaliação humana subjetiva para aperfeiçoar o algoritmo ou permanentemente alimentar o sistema com informações sobre erros de classificação, para corrigi-lo continuamente.

As possibilidades de correção do fluxo do uso de algoritmos de *machine learning*, portanto, são diversas e não têm limites *ex ante*. Se pesquisas indicarem a superioridade de um algoritmo na tomada de decisão, quando comparado com o simples julgamento humano, a rejeição a uma alternativa potencialmente melhor é equivalente a uma simples "aversão algorítmica" (DIETVORST; SIMMONS; MASSEY, 2015). Por outras palavras, se existir um algoritmo disponível com desempenho igual ou superior ao exclusivo julgamento humano, não há razão *a priori* para não o utilizar, isolada ou conjuntamente, na busca do aperfeiçoamento de processos decisórios relevantes para os seres humanos. Como adverte Kahneman (2012, p. 285) ao enfatizar as ideias de Meehl (1954), pode ser até considerado "antiético apoiar-se em julgamentos intuitivos para decisões importantes se um algoritmo que cometerá poucos erros está disponível".

5 Conclusão

Este artigo expôs como modelos de classificação em *machine learning* podem ser afetados por discriminações sociais em algumas aplicações de alto impacto jurídico e socioeconômico. Na primeira seção, procuramos explicar ao leitor conceitos básicos de modelos de classificação em inteligência artificial. Em seguida, foram analisados os principais problemas de equidade algorítmica em sistemas de avaliação de

risco de reincidência em processos criminais e a última seção sustentou teoricamente que, diante das limitações do julgamento humano (risco de decisão pouco refletida e/ou inconsistente), algoritmos apresentam inúmeras vantagens que motivam a necessidade de não se descartarem *a priori* suas aplicações em decisões jurídicas relevantes.

Em resposta ao problema suscitado por este artigo, com base na análise dos resultados de pesquisas empíricas demonstramos que decisões automatizadas tendem a ter uma acurácia igual ou superior à do julgamento humano. O que ratifica a hipótese do trabalho de que, mesmo com os riscos de discriminação algorítmica, modelos de *machine learning* não devem ser descartados *aprioristicamente* pela legislação, pois o desempenho do julgamento humano mensurado pelo critério *acurácia* é equivalente ou inferior ao realizado por dispositivos automatizados, sobretudo em temas complexos e permeados de incerteza.

Vimos também que, se algoritmos podem ser continuamente aperfeiçoados, técnicas de *machine learning* podem até representar uma vantagem diante do processo decisório exclusivamente humano, pois são dotadas de aptidão para corrigir distorções que normalmente influenciam a percepção subjetiva da realidade sem que sequer as percebam os atores responsáveis pela tomada de decisão. Por isso, e por ironia, os esforços da literatura sobre equidade algorítmica são justamente dedicados a livrar os sistemas computacionais dos equívocos resultantes da contaminação dos dados por estereótipos, preconceitos, vieses e inconsistências do julgamento exclusivamente humano.

No âmbito da produção legislativa, tais resultados reforçam a tese de que o caminho a ser trilhado não deve ser o da simples rejeição a decisões baseadas em algoritmos, com a proibição de seu uso em complementação ou substituição a julgamentos humanos. Os potenciais riscos da inteligência artificial (FERRARI; BECKER; WOLKART, 2018), a relevância internacional da tecnologia (LEE, 2019) e a impossibilidade de, em democracias constitucionais, haver a simples proibição do uso de modelos de *machine learning* em verdade indicam a necessidade de sua adequada regulação jurídica.

Para isso, não basta apenas adotar o caminho do art. 20 da Lei Geral de Proteção de Dados Pessoais (LGPD) (BRASIL, [2022b]), que prevê genericamente a possibilidade de revisão de decisões automatizadas a pedido do interessado titular dos dados ou, em última análise, a auditoria do sistema.[17] A multiplicidade dos tipos de sistema de *machine learning*

[17] Na Ciência da Computação, a área de pesquisa *privacidade algorítmica* (*algorithmic privacy*), mais afeita à LGPD, não se confunde com a *equidade algorítmica* (*algorithmic fairness*), de maior abrangência e com algoritmos e critérios de avaliação de equidade próprios, embora ambas apresentem vários pontos teóricos de interseção. Ver Kearns e

exigirá que, em cada domínio de aplicação (ex.: saúde, criminal, bancário, contratações trabalhistas), haja uma regulação setorial descentralizada baseada no prévio amadurecimento da discussão teórica e no debate público em torno dos critérios de equidade a adotar para a avaliação crítica da tecnologia, pois os parâmetros adotados na auditoria não são consensuais e exigem uma definição *ex ante* do critério de equidade a ser utilizado – se *balanceamento das taxas de erro*, *calibragem*, *paridade estatística* ou qualquer outro dentre os vários existentes na literatura especializada de cada setor.

Fica como sugestão ao leitor, para trabalhos futuros, revisar tais critérios e sugerir a sua utilidade em dado segmento da vida social, política ou econômica, pois é viável concebermos diretrizes para a fixação de um marco normativo de testagem, vigilância social e responsabilização jurídica pela utilização de sistemas de *machine learning*. Como vimos, esses sistemas podem ser aperfeiçoados para que "espelhem" o que há de melhor em nós mesmos, como seres humanos, em lugar de refletirem as discriminações sociais do passado.

Sobre o autor

Ricardo Silveira Ribeiro é doutor em Direito pela Universidade Federal de Pernambuco, Recife, PE, Brasil; procurador federal da Advocacia-Geral da União, Aracaju, SE, Brasil.
E-mail: ricardosribeiro1976@gmail.com

Como citar este artigo

(ABNT)

RIBEIRO, Ricardo Silveira. Inteligência artificial, Direito e equidade algorítmica: discriminações sociais em modelos de *machine learning* para a tomada de decisão. *Revista de Informação Legislativa*: RIL, Brasília, DF, v. 59, n. 236, p. 29-53, out./dez. 2022. Disponível em: https://www12.senado.leg.br/ril/edicoes/59/236/ril_v59_n236_p29

(APA)

Ribeiro, R. S. (2022). Inteligência artificial, Direito e equidade algorítmica: discriminações sociais em modelos de *machine learning* para a tomada de decisão. *Revista de Informação Legislativa: RIL*, *59*(236), 29-53. Recuperado de https://www12.senado.leg.br/ril/edicoes/59/236/ril_v59_n236_p29

Roth (2020), em especial o tratamento diferenciado dessas duas áreas de pesquisa nos capítulos 1 e 2 do livro.

Referências

ADIDA, Claire L.; LAITIN, David D.; VALFORT, Marie-Anne. *Why Muslim integration fails in Christian-heritage societies*. Cambridge, MA: Harvard University Press, 2016.

ANGWIN, Julia; LARSON, Jeff; MATTU, Surya; KIRCHNER, Lauren. Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, [*s. l.*], May 23, 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 11 jul. 2022.

BAROCAS, Solon; HARDT, Moritz; NARAYANAN, Arvind. *Fairness and machine learning*: limitations and opportunities. [*S. l.*]: fairmlbook.org, 2019. Disponível em: https://fairmlbook.org/. Acesso em: 11 jul. 2022.

BECKER, Gary S. *The economics of discrimination*. 2nd ed. Chicago: University of Chicago Press, 1971. (Economics Research Studies of the Economics Research Center of the University of Chicago).

BENJAMIN, Ruha. *Race after technology*: abolitionist tools for the new Jim Code. Medford, MA: Polity, 2019.

BRASIL. *Decreto-lei nº 3.689, de 3 de outubro de 1941*. Código de Processo Penal. Brasília, DF: Presidência da República, [2022a]. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto-lei/del3689compilado.htm. Acesso em: 11 jul. 2022.

_____. *Lei nº 13.709, de 14 de agosto de 2018*. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, [2022b]. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 11 jul. 2022.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender shades: intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, [*s. l.*], v. 81, p. 1-15, 2018. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Disponível em: https://proceedings.mlr.press/v81/buolamwini18a.html. Acesso em: 11 jul. 2022.

CATON, Simon; HAAS, Christian. Fairness in machine learning: a survey. *In*: ARXIV. [*S. l.*]: Cornell University, 4 Oct. 2020. DOI: https://doi.org/10.48550/arXiv.2010.04053. Disponível em: https://arxiv.org/abs/2010.04053. Acesso em: 11 jul. 2022.

CHOLLET, François. *Deep learning with Python*. Shelter Island, NY: Manning Publications Co., 2018.

CHOULDECHOVA, Alexandra. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*, [*s. l.*], v. 5, n. 2, p. 153-163, June 2017. DOI: http://doi.org/10.1089/big.2016.0047.

CODED Bias. Direção: Shalini Kantayya. [*S. l.*]: Netflix, 2020. 1 vídeo (85 min). Disponível em: https://www.netflix.com/br-en/title/81328723. Acesso em: 11 jul. 2022.

CORBETT-DAVIES, Sam; PIERSON, Emma; FELLER, Avi; GOEL, Sharad; HUQ, Aziz. Algorithmic decision making and the cost of fairness. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 23., Aug. 13-17, 2017, Halifax, CA. *Proceedings* […]. New York: Association for Computing Machinery, 2017. p. 797-806. DOI: https://doi.org/10.1145/3097983.3098095.

CRAWFORD, Kate. *Atlas of AI*: power, politics, and the planetary costs of artificial intelligence. New Haven: Yale University Press, 2021.

DANZIGER, Shai; LEVAV, Jonathan; AVNAIM-PESSO, Liora. Extraneous factors in judicial decisions. *PNAS*, [*s. l.*], v. 108, n. 17, p. 6.889-6.892, Apr. 26, 2011. DOI: https://doi.org/10.1073/pnas.1018033108. Disponível em: https://www.pnas.org/doi/10.1073/pnas.1018033108. Acesso em: 11 jul. 2022.

DIETERICH, William; MENDOZA, Christina; BRENNAN, Tim. *COMPAS risk scales*: demonstrating accuracy equity and predictive parity. [*S. l.*]: Northpointe Inc. Research

Department, 2016. Disponível em: https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf. Acesso em: 11 jul. 2022.

DIETVORST, Berkeley J.; SIMMONS, Joseph P.; MASSEY, Cade. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology*: general, [*s. l.*], v. 144, n. 1, p. 114-126, Feb. 2015. DOI: https://doi.org/10.1037/xge0000033.

DONEDA, Danilo Cesar Maganhoto; MENDES, Laura Schertel; SOUZA, Carlos Affonso Pereira de; ANDRADE, Norberto Nuno Gomes de. Considerações iniciais sobre inteligência artificial, ética e autonomia pessoal. *Pensar*: Revista de Ciências Jurídicas, Fortaleza, v. 23, n. 4, p. 1-17, out./dez. 2018. DOI: https://doi.org/10.5020/2317-2150.2018.8257. Disponível em: https://periodicos.unifor.br/rpen/article/view/8257. Acesso em: 11 jul. 2022.

FERRARI, Isabela; BECKER, Daniel; WOLKART, Erik Navarro. *Arbitrium ex machina*: panorama, riscos e a necessidade de regulação das decisões informadas por algoritmos. *Revista dos Tribunais*, São Paulo, v. 107, n. 995, p. 635-655, set. 2018.

FLORES, Anthony W.; BECHTEL, Kristin; LOWENKAMP, Christopher T. False positives, false negatives, and false analyses: a rejoinder to "Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks". *Federal Probation*, Washington, DC, v. 80, n. 2, p. 38-46, Sept. 2016. Disponível em: https://www.uscourts.gov/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-rejoinder. Acesso em: 11 jul. 2022.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep learning*. Cambridge, MA: MIT Press, 2016. (Adaptive Computation and Machine Learning).

GOOGLE apologises for photos app's racist blunder. *BBC News*, [*s. l.*], 1 July 2015. Disponível em: https://www.bbc.com/news/technology-33347866. Acesso em: 11 jul. 2022.

GROVE, William M.; ZALD, David H.; LEBOW, Boyd S.; SNITZ, Beth E.; NELSON, Chad. Clinical versus mechanical prediction: a meta-analysis. *Psychological Assessment*, [*s. l.*], v. 12, n. 1, p. 19-30, 2000. DOI: https://doi.org/10.1037/1040-3590.12.1.19.

HOWARD, Jeremy; GUGGER, Sylvain. *Deep learning for coders with fastai and PyTorch*: AI applications without a PhD. Sebastopol, CA: O'Reilly Media, 2020.

HUTCHINSON, Ben; MITCHELL, Margaret. 50 years of test (un)fairness: lessons for machine learning. *In*: CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, Jan. 29-31, 2019, Atlanta, GA. *Proceedings* […]. New York: Association for Computing Machinery, 2019. p. 49-58. DOI: https://doi.org/10.1145/3287560.3287600.

KAHNEMAN, Daniel. *Rápido e devagar*: duas formas de pensar. Tradução de Cássio de Arantes Leite. Rio de Janeiro: Objetiva, 2012.

KEARNS, Michael; ROTH, Aaron. *The ethical algorithm*: the science of socially aware algorithm design. New York: Oxford University Press, 2020.

KLEINBERG, Jon; LAKKARAJU, Himabindu; LESKOVEC, Jure; LUDWIG, Jens; MULLAINATHAN, Sendhil. Human decisions and machine predictions. *NBER Working Paper*, Cambridge, MA, n. 23.180, p. 1-76, Feb. 2017. DOI: https://doi.org/10.3386/w23180. Disponível em: https://www.nber.org/papers/w23180. Acesso em: 11 jul. 2022.

KLEINBERG, Jon; MULLAINATHAN, Sendhil; RAGHAVAN, Manish. Inherent trade-offs in the fair determination of risk scores. *In*: ARXIV. [*S. l.*]: Cornell University, 17 Nov. 2016. DOI: https://doi.org/10.48550/arXiv.1609.05807. Disponível em: https://arxiv.org/abs/1609.05807v2. Acesso em: 11 jul. 2022.

KRZANOWSKI, Wojtek J.; HAND, David J. *ROC curves for continuous data*. Boca Raton: CRC Press, 2009. (Monographs on Statistics and Applied Probability, 111).

LARSON, Jeff; MATTU, Surya; KIRCHNER, Lauren; ANGWIN, Julia. How we analyzed the COMPAS recidivism algorithm. *ProPublica*, [*s. l.*], May 23, 2016. Disponível em: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Acesso em: 11 jul. 2022.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. *Nature*, [*s. l.*], v. 521, n. 7.553, p. 436-444, May 28, 2015. DOI: https://doi.org/10.1038/nature14539.

LEE, Kai-Fu. *Inteligência artificial*: como os robôs estão mudando o mundo, a forma como amamos, nos relacionamos, trabalhamos e vivemos. Tradução de Marcelo Barbão. Rio de Janeiro: Globo Livros, 2019.

MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, [*s. l.*], v. 5, n. 4, p. 115-133, 1943. DOI: https://doi.org/10.1007/BF02478259.

MEEHL, Paul E. *Clinical versus statistical prediction*: a theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press, 1954.

MURPHY, Kevin P. *Machine learning*: a probabilistic perspective. Cambridge, MA: MIT Press, 2012. (Adaptive Computation and Machine Learning Series).

NATIONAL INSTITUTE OF CORRECTIONS. *Myths & facts*: using risk and need assessments to enhance outcomes and reduce disparities in the Criminal Justice System. Written by Dr. Cara Thompson. Washington, DC: NIC, 2017. (Myths & Facts). Disponível em: https://nicic.gov/myths-facts-using-risk-and-need-assessments-enhance-outcomes-and-reduce-disparities-criminal-justice. Acesso em: 11 jul. 2022.

NORTHPOINTE. *Practitioner's guide to COMPAS core*. [*S. l.*]: Northpointe Inc., 2015.

NORTHPOINTE suite risk needs assessments. *In*: OUR PRODUCTS. [*S. l.*]: Equivant, c2022. Disponível em: https://www.equivant.com/northpointe-risk-need-assessments. Acesso em: 11 jul. 2022.

O'NEIL, Cathy. *Weapons of math destruction*: how big data increases inequality and threatens democracy. New York: Crown, 2016.

PARKHI, Omkar M.; VEDALDI, Andrea; ZISSERMAN, Andrew; JAWAHAR, C. V. *The Oxford-IIIT Pet Dataset*. Oxford, UK: University of Oxford, Department of Engineering Science, [20--]. Disponível em: https://www.robots.ox.ac.uk/~vgg/data/pets/. Acesso em: 11 jul. 2022.

RUSSELL, Stuart J.; NORVIG, Peter. *Artificial intelligence*: a modern approach. 3rd ed. Harlow, UK: Pearson, 2016. (Pearson Series in Artificial Intelligence).

SEJNOWSKI, Terrence J. *The deep learning revolution*: artificial intelligence meets human intelligence. Cambridge, MA: MIT Press, 2018.

SWEENEY, Laura T.; HANEY, Craig. The influence of race on sentencing: a meta-analytic review of experimental studies. *Behavioral Sciences and the Law*, [*s. l.*], v. 10, n. 2, p. 179-195, 1992. DOI: https://doi.org/10.1002/bsl.2370100204.

TETLOCK, Philip E.; GARDNER, Dan. *Superprevisões*: a arte e a ciência de antecipar o futuro. Tradução de Cássio de Arantes Leite. Rio de Janeiro: Objetiva, 2015.

VALENTINO-DEVRIES, Jennifer; SINGER-VINE, Jeremy; SOLTANI, Ashkan. Websites vary prices, deals based on users' information. *The Wall Street Journal*, [*s. l.*], Dec. 24, 2012. Disponível em: https://www.wsj.com/articles/SB10001424127887323777204578189391813881534. Acesso em: 11 jul. 2022.

VERMA, Sahil; RUBIN, Julia. Fairness definitions explained. *In*: INTERNATIONAL WORKSHOP ON SOFTWARE FAIRNESS, 40., May 29, 2018, Gothenburg, SE. *Proceedings* […]. New York: Association for Computing Machinery, 2018. p. 1-7. DOI: https://doi.org/10.1145/3194770.3194776.

ZEMEL, Richard; WU, Yu; SWERSKY, Kevin; PITASSI, Toniann; DWORK, Cynthia. Learning fair representations. *Proceedings of Machine Learning Research*, [*s. l.*], v. 28, n. 3, p. 325-333, 2013. Proceedings of the 30th International Conference on Machine Learning. Disponível em: https://proceedings.mlr.press/v28/zemel13.html. Acesso em: 12 jul. 2022.

ŽLIOBAITĖ, Indrė. Measuring discrimination in algorithmic decision making. *Data Mining Knowledge Discovery*, [*s. l.*], v. 31, n. 2, p. 1.060-1.089, July 2017. DOI: https://doi.org/10.1007/s10618-017-0506-1.

Apêndice técnico

Definições formais dos critérios *balanceamento das taxas de erro* e *calibragem*

Suponha-se que duas variáveis binárias denominadas *e_positivo* e *e_grupo_vulnerável* foram categorizadas de acordo com o critério abaixo:

e_positivo = 1 se o risco é real (positivo) e 0 se não é real (negativo);

e_grupo_vulnerável = 1 se pertence a algum grupo vulnerável no contexto examinado (ex.: negros, mulheres, pobres) e 0 se pertence ao grupo dominante (ex.: brancos, homens, ricos).

Para um escore, s, calculado pela função s = f(X), e um limiar, t, fixado pelo tomador de decisão, tem-se que:

f(.) é a função usada para classificação (*score function*);

X é o conjunto de características (atributos) do objeto a ser classificado;

s > t significa instância classificada como positiva;

s ≤ t significa instância classificada como negativa;

p(.) é a proporção ou probabilidade.

Definição de *balanceamento das taxas de erro*: um escore s satisfaz a balanceamento das taxas de erro se, em determinado limiar, t, a proporção, p, de falsos positivos é igual à proporção, p, de falsos negativos no grupo dominante e no grupo vulnerável, ou seja:

$$p(s>t|e\_positivo=0, e\_grupo\_vulnerável=1)=p(s>t|e\_positivo=0, e\_grupo\_vulnerável=0)$$

$$p(s\leq t|e\_positivo=1, e\_grupo\_vulnerável=1)=p(s\leq t|e\_positivo=1, e\_grupo\_vulnerável=0)$$

Definição de calibragem: um escore s é calibrado se a probabilidade, p, de o escore, s, classificar instâncias como positivas for independente da variável e_vulnerável, ou seja, essa probabilidade é igual para ambos os grupos (dominante e vulnerável):

$$p(s>t|e\_grupo\_vulnerável=1)=p(s>t|e\_grupo\_vulnerável=0)$$