A PROBLEMÁTICA DA INTELIGÊNCIA ARTIFICIAL E DOS VIESES ALGORÍTMICOS: CASO COMPAS

Leonardo Marques Vieira
Universidade Presbiteriana Mackenzie
GPDIT
Campinas, São Paulo
leo.mvieiraa@gmail.com

***Resumo*—Este artigo tem o intuito de analisar o risco das novas tecnologias, principalmente o programa COMPAS, o qual é usado para se avaliar o risco de reincidência de um réu em um processo criminal e quais são as possíveis reparações dos vieses existentes, haja vista que o preconceito poderá surgir em diversos estágios afetando a vida das pessoas no sistema jurídico criminal.**

Palavra-chave— viés; algoritmo; inteligência artificial.

I. INTRODUCTION

Não é novidade que as informações que circulam na internet não são inseridas somente por pessoas, mas também por algoritmos e plataformas que trocam dados entre si. Vivemos em um mundo digital, onde o homem dialoga com máquinas fazendo com que os algoritmos passem a tomar decisões e determinar avaliações e inclusive ações que outrora eram feitas por humanos. Como a relação entre homens e as novas tecnologias disruptivas é algo recente, essa nova cultura provoca reflexões éticas relevantes, tendo em vista as consequências que podem ser causadas pela inteligência artificial.

De forma geral, os algoritmos são modelos matemáticos (softwares) ordenados para uma determinada finalidade, buscando padrões de números. Contudo, sabe-se que os algoritmos são falíveis e limitados, pois são opiniões embutidas em um código, por meio do qual o homem ensina a máquina, ou seja, a máquina poderá tomar decisões enviesadas com base nos dados fornecidos. Logo, torna-se necessário nutrir os dados de forma precisa para que a inteligência artificial não cometa erros e não seja discriminatória.

Os algoritmos estão presente em todos os lugares podendo influenciar a vida diária de cada indivíduo, estão presentes no mercado financeiro, na área jurídica e diversos outros lugares, inclusive definindo os conteúdos que recebem nas redes sociais. À vista disso, surgem novas preocupações em relação a transparência algorítmica, privacidade pessoal e principalmente, aos vieses da inteligência artificial nos casos que envolvem raça, cor e gênero. Porém, como saber se um modelo de aprendizado de máquina é realmente justo? E o que significa justiça nesse modelo?

A inteligência artificial pode até ser inteligente, porém não possui a expertise de um humano, não é sábia. Tudo o que as máquinas sabem foi por que o homem o ensinou e em consequência, também ensinou os preconceitos ou aprenderam por meio do *machine learning*, tendo a capacidade de reproduzir os vieses humanos. Entretanto, para que não seja tarde demais, as pessoas devem tomar uma iniciativa para corrigir os vieses existentes em algoritmos. Nesse contexto de um mundo algorítmico a governança e proteção de dados são essenciais.

II. O VIÉS ALGORÍTMICO DO COMPAS

O algoritmo COMPAS (Perfil de Gerenciamento Corretivo de Infratores para Sanções Alternativas), foi elaborado pela empresa Northpointe (hoje com o nome *Equivant*), com o intuito de realizar avaliações de riscos sobre pessoas que voltam a praticar crimes, auxiliar nas informações de decisões e mitigar riscos futuros promovendo auxílio e orientação para os juízes nos tribunais dos Estados Unidos.

Esse algoritmo vinha sendo utilizado para determinar a probabilidade de reincidência de prisioneiros. Contudo, um estudo feito pela ProPublica (jornal de cunho investigativo) colocou em dúvida o seu uso, sendo constatado que o algoritmo era racialmente enviesado. O jornal conseguiu dados das pontuações de risco analisando mais de 7 mil pessoas presas no condado de Broward, Flórida nos anos de 2013 e 2014[1].

O *score* de avaliação de risco da empresa apontava as pessoas negras como de alto risco e as pessoas brancas como de baixo risco. Após as análises da ProPublica, detectaram que os negros que possuem alto risco não eram acusados de novos crimes e os brancos que eram caracterizados como de baixo risco vinham a cometer novos crimes, isto é, os negros tinham mais chances do que os bancos de serem taxados como alto risco. Observa-se que os dados eram viciados com

[1] Disponível em: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> Acesso em: 10 mai. 2019

informações anteriores, as quais influenciaram negativamente as decisões.

A empresa *Equivant*, mediante uma carta pública, refutou completamente da análise feita pela ProPublica dizendo que o software não possuía um viés racial e que seu algoritmo era preciso, além disso, concluíram que as análises feitas possuíam erros estatísticos e técnicos, tanto é que a empresa criou um relatório denominado "COMPAS *Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*" fornecendo uma resposta ao artigo publicado pela ProPublica[2].

O *software* avalia diversos critérios dentre eles: história criminal, criminalidade da família, colegas, abuso de substâncias, residência/estabilidade, ambiente social, educação, trabalho, lazer/recreação, isolamento social, personalidade criminosa, raiva e atitudes criminosas. Para isso, eram feitas 137 perguntas a serem respondidas pelos réus em um questionário contendo também os antecedentes criminais dos envolvidos[3]. A cor não era uma variável evidente introduzida nesse algoritmo, mas raça e gênero são integrados em diversas outras variáveis, como por exemplo, onde moramos, nossas redes sociais e nossa educação. Era perguntado inclusive se o indivíduo participava de gangues, se possuía pais separados, se tinha amigos presos, se no bairro em que vive era necessário portar arma e etc. Ao final desses questionamentos o algoritmo confirmava o risco dos indivíduos reincidir, classificando assim como de baixo, médio ou alto risco. O uso do COMPAS para um policiamento preditivo se tornou uma grande preocupação.

Fato é que, quando o software ajudava os juízes nos tribunais dos Estados Unidos para formarem conclusões sobre o futuro dos réus a fim de condená-los, a análise era feita com base em informações de outras pessoas e supostamente prevendo a futura reincidência, o que é completamente contrário aos princípios da garantia do estado de inocência e o devido processo legal do direito penal, tirando a personalidade da condenação e da pena. Conforme o algoritmo se baseava no histórico de condenações anteriores ocorreu esse enviesamento. Sendo assim, os processos seriam considerados inquisitórios, pois o juiz não estaria sendo imparcial.

Outro ponto importante a ser debatido é a revelação dos dados. Não se sabe ao certo como é o funcionamento do algoritmo do COMPAS, pois a empresa se recusou a divulgá-lo. Isso faz com que o réu não consiga questionar o resultado, pois não se sabe como foi calculado o seu risco. *Christopher Slobogin*, diretor do programa de justiça criminal, diz que tais avaliações de risco deveriam ser banidas, salvo se ambas as partes obtivessem acesso completo a todos os dados do algoritmo, consequentemente, possibilitando contraditório e ampla defesa[4]. Os únicos com acesso total aos algoritmos são seus próprios programadores, mas a pergunta que fica é: os programadores são mais bem preparados que os juízes para realizar justiça? A questão é se agride a separação de poderes o cidadão comum decidir no lugar do juiz?

Sabemos que as sentenças nos EUA variam de acordo com cada estado, porém todas as sentenças devem respeitar a Constituição, obedecendo os princípios e valores supremos. Por certo, o juiz que deve proferir uma sentença (obedecendo as normas da constituição), a dificuldade está quando o sistema é preconceituoso. O algoritmo pode ser falho e o juiz poderá abusar de tal prerrogativa sentenciando por meio desses *scores*, afetando a ética e a moral (como no caso a raça).

Defensores afirmam que as máquinas são menos tendenciosas do que os humanos, por mais bem intencionadas que as pessoas sejam[5]. Porém, substituir o pensamento crítico do juiz pelo das máquinas também é um problema, pois sabemos que o ser humano é falho, podendo transferir seus preconceitos para o algoritmo podendo resultar em um impacto negativo na vida de uma pessoa. Os *scores* não indicam se uma pessoa é perigosa ou se deve ir para a cadeia, mas sim para definir quais os programas de tratamento ou de condicional o réu terá. Os juízes não devem tomar essa avaliação de risco como alicerce e proferir a sentença em cima disso.

A ProPublica afirmou que o *software* da empresa prevê acertadamente 61% dos casos, porém os negros são duas vezes mais inclinados a reincidir e serem qualificados como risco maior. Inclusive, há casos em que juízes citaram as avaliações de risco e pontuações nas decisões, por exemplo, *Zilly* roubou objetos para manter seu vício em metanfetamina, porém estava se esforçando para se recuperar. *Zilly* foi apontado como tendo alto risco de reincidência e foi mandado para a prisão. Houve recurso e *Brennan* (criador do *software*) foi depor declarando que o software não tinha esse intuito de ser utilizado em sentenças, o objetivo era somente reduzir os crimes. A questão é que o COMPAS não deveria ser utilizado como meio de prova. A partir dessa declaração o juiz reduziu drasticamente a pena de *Zilly*[6].

Julia Dressel do *Dartmouth College* afirma que antes de chegar a justiça, torna-se necessário que haja uma certeza

---

[2] Disponível em: <http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf> Acesso em: 10 mai. 2019

[3] Disponível em: <https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html> Acesso em: 10 mai. 2019.

[4] Disponível em <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> Acesso em: 10 mai. 2019

[5] Disponível em: <https://www.ge.com/reports/will-smart-machines-be-less-biased-than-humans/> Acesso em: 10 mai. 2019.

[6] Disponívem em: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> Acesso em: 10 mai. 2019.

que tais ferramentas utilizadas para auxílio na justiça sejam precisas, caso contrário não será justo para ninguém[7].

De acordo com *John Giannandrea* (chefe de I.A. do *Google*) diz que a grande questão de segurança, é que se for atribuído um dado tendencioso as máquinas serão tendenciosas. *Giannandrea* está receoso quanto aos vieses da inteligência artificial aprendidas com as pessoas que tomam decisões a todo instante. Ainda disse que deve haver uma transparência sobre os dados de treinamento que são utilizados e ainda uma procura por vieses ocultos, senão o sistema resultante será enviesado[8].

III. O QUE VEM SENDO FEITO PELA EMPRESA?

Analisando o site acessível a todo público, a *Equivant* utiliza de seu *software* para auxílio de juízes, administradores e tomadores de decisão na condenação, gestão e outras necessidades vindas do regime carcerário, não deixando explícita nenhuma evidência questionada anteriormente pela ProPublica sobre o viés racial do algoritmo. Não há garantias de que a empresa tenha melhorado ou atualizado seu algoritmo após fortes críticas analisadas pela ProPublica.

IV. POSSÍVEIS REPARAÇÕES DOS VIESES EXISTENTES?

Há dados comprovando que os EUA possui cerca de 2,2 milhões de pessoas presas, tanto no âmbito federal quanto estadual, sendo que o número de presos aumentou se comparado aos anos 1980[9], o que pode estar ligado a existência de um algoritmo discriminatório. Como o *software* foi desenvolvido para encontrar padrões de reincidência, isso trouxe impactos significativos no modo em que o COMPAS estava sendo utilizado, afetando vida dos indivíduos que não possuíam uma probabilidade de reincidência.

Entretanto, durante as fases iniciais de testes de algoritmos que, como o COMPAS, utilizam a *clusterização* de dados (análise de dados estatísticos referente a um conjunto de grupos de pessoas, produtos etc, que tenham características similares)[10], é muito difícil detectar enviesamento, devido à própria natureza de código. Isso significa que, as soluções mais simples e de maior impacto, que em geral giram em torno dos dados colhidos pelo *software*, não podem ser implementadas de forma efetiva, uma vez que são inúteis depois que a máquina já adquiriu característica enviesada.

Corrigir os vieses existentes nas máquinas que utilizam machine learning é tarefa complexa e, se não for feito, pessoas podem ser afetadas, como no caso do COMPAS. Para que sejam corrigidos é necessário entender de onde vem e de onde surgem esses vieses. De acordo com o MIT, os preconceitos podem aparecer antes mesmo dos dados serem coletados, ou em outras fases do procedimento de Machine Learning. Assim sendo, elencam três pontos principais de como o viés pode ocorrer. O primeiro deles é o enquadramento do problema, ou seja, nesse ponto o programador define o que o programa fará especificamente, pois se fugir de sua finalidade poderá ocorrer, como no caso do COMPAS, uma discriminação. O segundo ponto é no momento da coleta de dados, no qual o preconceito poderá surgir, onde o algoritmo poderá ser treinado com dados tendenciosos. E o terceiro e último ponto elencado refere-se na preparação dos dados, isto é, o viés pode ser incorporado durante a elaboração dos dados. Se no momento da elaboração forem utilizados dados sensíveis, como raça, cor, gênero ou outros fatores implícitos poderá haver esse enviesamento[11].

Posto isto, as soluções passíveis de implementação requerem interferência mais invasiva ao código do *software*, e podem comprometer o aprendizado de máquina, tornando-a possivelmente menos precisa do que em versões onde o código se desenvolve sem tais alterações. Possíveis soluções a serem implantadas em estágios avançados do programa incluem a anulação de diretrizes enviesadas; implementação manual de novas diretrizes ou implementação de fatores de correção a diretrizes pré-existentes.

Em caso de *software* em desenvolvimento ou em fases iniciais, medidas mais eficientes e menos invasivas podem ser tomadas, tais como: filtrar os dados de entrada, ou seja, obter um novo questionário para a classificação do réu onde não leve em conta dados pessoais como: local de moradia, renda, fotos, entre outros, pois é justamente com tais opções que levam a condenar um maior número de negros, consequência de seu contexto histórico ainda enraizado na sociedade atualmente - podendo resultar em uma grande mudança na previsão de reincidência; Utilização de bancos de dados heterogêneos para não ocorrer algum tipo de enviesamento ou "alimentar" mais os dados algorítmicos.

V. CONCLUSÃO

O padrão humano de avaliação está longe de ser o ideal, sendo propício a influências. Os algoritmos estão buscando aperfeiçoar essas avaliações, mas certamente estão longe de

---

7 Disponível em: <https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/> Acesso em: 11 mai. 2019.

8 Disponível em: <https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/?fbclid=IwAR2d1_fJPi430OALq4HriyPZoK8v4vRSJ_Q6Ths8-n772T77BiYdcwyGr4s> Acesso em: 15 mai. 2019.

9 Disponível em: <https://www.bbc.com/portuguese/internacional-37195944> Acesso em: 14 mai. 2019.

10 Disponível em: <https://lamfo-unb.github.io/2017/10/05/Introducao_basica_a_clusterizacao/> Acesso em: 14 mai. 2019.

11 Disponível em: <https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/> Acesso em: 14 mai. 2019

serem perfeitos também. Sabe-se que em avaliações feitas por *softwares* existem falhas humanas integradas. Fato é que essas informações podem ser utilizadas para aumentar a sentença e punir alguém pelo crime que ainda não cometeu, se os indivíduos ainda não cometeram crimes, não há no que se falar em um julgamento quanto à possibilidade de violar a lei novamente.

Vale ressaltar que, para que haja uma análise detalhada do *software*, é necessário a revelação completa do código, visto que, sem essas informações não há como realizar um estudo elaborado para constatar se houve ou não uma discriminação intencional. Uma vez utilizado o COMPAS na justiça criminal dos Estados Unidos da América, os códigos adotados deveriam ser divulgados para que assim exista uma garantia de direitos, transparência e avaliação das decisões automatizadas, pois, como visto, estes *softwares* são capazes de causar impactos significativos na vida das pessoas.

Como já sabido, os algoritmos que utilizam *Machine Learning* realizam análises dos dados fornecidos inicialmente e conforme entram novos dados, procuram determinados padrões semelhantes. Caso as máquinas aprenderem códigos errados, haverá dados falhos e informações possivelmente influenciadas por diversos fatores. Geralmente nas delegacias os dados rastreados pelas polícias retratam os preconceitos preexistentes nas próprias instituições como no caso de renda, residência, raça e gênero, de modo que o agrupamento de dados aprendido pelas máquinas é enviesado desde o início.

De fato, para corrigir os vieses dos algoritmos é necessário compreender que eles não são imparciais, são objetivos se comparados com as pessoas, porém isso não o torna equitativo, assim poderá haver decisões injustas e discriminatórias. A inteligência artificial veio para auxiliar o ser humano, é um momento de disrupção e há grandes esforços por parte de pesquisadores e programadores em tentar corrigir esses vieses existentes nas máquinas.

Utilizar um algoritmo preditivo é um grande desafio, principalmente nas delegacias e tribunais que lidam com dados para tentar reduzir a criminalidade. Como dito, deve haver um sistema transparente que garanta uma segurança jurídica e previsibilidade, pois qualquer decisão tomada afetará a vida de pessoas.

REFERENCES

Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks. Disponível em: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> Acesso em: 10 mai. 2019.

How We Analyzed the COMPAS Recidivism Algorithm. Disponível em: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> Acesso em: 10 mai. 2019.

The accuracy, fairness, and limits of predicting recidivism. Disponível em: <https://advances.sciencemag.org/content/4/1/eaao5580.full> Acesso em: 10 mai. 2019.

The era of blind faith in big data must end - Cathy O'Neil. Disponível em: <https://www.youtube.com/watch?v=_2u_eHHzRto> Acesso em: 06 mai. 2019.

Preconceito das máquinas: como algoritmos podem ser racistas e machistas. Disponível em: <https://noticias.uol.com.br/tecnologia/noticias/redacao/2018/04/24/preconceito-das-maquinas-como-algoritmos-tomam-decisoes-discriminatorias.htm> Acesso em: 10 mai. 2019.

Rise of the racist robots – how AI is learning all our worst impulses. Disponível em: <https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses> Acesso em: 10 mai. 2019.

Sample Risk Assessment COMPAS Core. Disponívem em: <https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html> Acesso em: 10 mai. 2019.

When an Algorithm Helps Send You to Prison. <https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html> Acesso em: 14 mai. 2019.

EQUIVANT - Compas Classification. Disponível em: <http://www.equivant.com/wp-content/uploads/Classification.pdf>. Acesso em: 11 maio 2019.

Racial Bias and Gender Bias Examples in AI systems. Disponível em: <https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1> Acesso em: 14 mai. 2019.

Forget Killer Robots - Bias Is the Real AI Danger. Disponível em: <https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/?fbclid=IwAR2d1_fJPi430OALq4HriyPZoK8v4vRSJ_Q6Ths8-n772T77BiYdcwyGr4s> Acesso em: 15 mai. 2019.

A Popular Algorithm Is No Better at Predicting Crimes Than Random People. Disponível em: <https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/> Acesso em: 11 mai. 2019.

Will Smart Machines Be Less Biased Than Humans? Disponível em: <https://www.ge.com/reports/will-smart-machines-be-less-biased-than-humans/> Acesso em: 10 mai. 2019.

Big Risks, Big Opportunities: the Intersection of Big Data and Civil Rights. Disponivel em:<https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights> Acesso em: 19 mai. 2019.

This is how AI bias really happens—and why it's so hard to fix. Disponível em: <https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/> Acesso em: 14 mai. 2019.