Auditoria para detecção de vieses discriminatórios em métodos de *machine learning*

André Kazuo Takahata [1]

Universidade Federal Do ABC (Brasil)

andre.t@ufabc.edu.br

Selma Carloto [2]

Universidade Federal Do ABC (Brasil)

selmacarloto@hotmail.com

Recibido el: 05.07.2022 *Aceptado el: 25.07.2022*

Resumo

Em uma sociedade digital e algorítmica a proteção de dados assume um papel fundamental frente às novas tecnologias cada vez mais invasivas e surge a

[1] André Kazuo Takahata é docente dos cursos de graduação em Engenharia de Informação e Pós-graduação lato e stricto sensu em Engenharia da Informação (PPG-INF) na Fundação Universidade Federal do ABC (UFABC). Possui graduação, mestrado e doutorado em Engenharia Elétrica pela Universidade Estadual de Campinas. Atualmente ministra disciplinas ligadas à eletrônica, programação e processamento de linguagem natural (PLN) na graduação e a respeito de métodos de engenharia para aprendizado de máquina e aprendizado profundo. Possui trabalhos nas áreas de otimização e processamento digital de sinais nas áreas de processamento de sinais sísmicos, interface cérebro-computador e caracterização de sinais de potenciais local de campo para doença de Parkinson. Além disso, tem trabalhado com PLN com linguagem biomédica, em especial em textos científicos a respeito da COVID-19 em língua portuguesa e em voicebots para a área da saúde no Brasil. Também está orientando trabalho de doutorado a respeito de combate à discriminação em processos de recrutamento e seleção com uso de algoritmos de aprendizado de máquina.

[2] Autora das obras: Compliance Trabalhista; Compliance Trabalhista e a efetividade dos Direitos Humanos; Lei Geral de Proteção de Dados enfoque nas relações de Trabalho; Manual Prático de Adequação à Lei Geral de Proteção de Dados; Lei Geral de Proteção de Dados e Segurança da Informação:perguntas e respostas; coordenadora/autora da obra LGPD comentada, enfoque nas relações de trabalho. Pesquisadora da Usp, departamento de Direito Civil, de Lei Geral de Proteção de Dados. Exin Data Protection Officer certificada. Professora convidada da Fundação Getúlio Vargas da FGV Direito Rio e dos MBAs de Gestão de Pessoas, Gestão Empresarial e Gestão Comercial da área de Direito. Professora premiada como destaque da área de Direito pela rede FGV Management, e pelo IDE, nos anos 2011, 2012, 2013 e 2014 consecutivamente, dos cursos de pós-graduação. Prêmio de destaque no MBA de Direito do Trabalho da Fgv Management Rio de 2016. Condecorada pela FGV Direito Rio com o prêmio de desempenho como docente nos cursos de pós-graduação da FGV em 2011, 2013 e 2015. Membro do Instituto Nacional de Proteção de Dados. Mestre pela Universidade de São Paulo (USP). Doutoranda em engenharia da informação na Universidade Federal do ABC em Inteligência Artificial. Doutorado em Direito do Trabalho na Universidade Federal de Buenos Aires (UBA). Especialista pela FADISP. Autora dos livros publicados na Argentina, Editorial Quorom, Manual de Derecho Laboral e Interesses Metaindividuais e ações coletivas. e-mail: selmacarloto@hotmail.com inst.selmacarloto@hotmail.com

preocupação com os estereótipos e vieses discriminatórios, já que estes poderão ser reforçados e constatados em atividades de tratamento por meio de inteligência artificial, em relações em geral, mas principalmente em processos seletivos com a consequente lesão a direitos de personalidade.

Palavras chave: algoritmos; vieses discriminatórios; aprendizado de máquina; detecção e Lei Geral de Proteção de Dados

Resumen

En una sociedad digital y algorítmica, la protección de datos asume un papel fundamental frente a las nuevas tecnologías cada vez más invasivas y surgen preocupaciones con los estereotipos y sesgos discriminatorios, ya que estos pueden reforzarse y verificarse en las actividades de tratamiento a través de la inteligencia artificial, en las relaciones en general, pero principalmente en procesos selectivos con la consiguiente lesión de los derechos de la personalidad.

Palabras clave: algoritmos; sesgos discriminatorios; aprendizaje automático; detección y Ley General de Protección de Datos

1. INTRODUÇÃO

O que parecia um sonho do futuro inalcançável, transformou-se em realidade, em busca de maior produtividade e efetividade, na realização de tarefas, as empresas passaram a utilizar robôs e algoritmos de Inteligência Artificial, mas estes nem sempre são, como aqueles que sempre imaginamos e vimos em filmes de ficção científica, em formato humanoide.

Em um mundo cada vez mais tecnológico e digital permeado por mentes inteligentes e computacionais e do outro lado máquinas que imitam a inteligência humana, ambas interagindo, passam a ser criados algoritmos cada vez mais avançados e novos métodos, ou softwares de inteligência artificial. Um algoritmo é um conjunto de instruções e pode ser ou não de inteligência artificial. Um programa é um conjunto de instruções, ou operações, que processam dados iniciais, o qual poderá ter como sinônimo o algoritmo, incluindo a transformação de dados naquilo que é almejado.

Um dos dilemas éticos mais complexos e que gera atualmente intensa preocupação é a proteção à privacidade e a outros direitos fundamentais dos trabalhadores, diante da utilização massiva de novas e cada vez mais invasivas e avançadas tecnologias digitais pelas empresas, que surgem no atual cenário de big data, principalmente durante a pandemia, o que aumentou a necessidade do contato digital entre pessoas e nas empresas, com isolamento social. Já existem empresas que utilizam microchips sob a pele dos trabalhadores, o que pode ser considerado uma nova e verdadeira "escravidão eletrônica", marcada pelos avanços tecnológicos.

O direito à privacidade renasce num cenário de novos tempos da tecnologia digital, como resposta aos ataques e às consequências sofridas pelos trabalhadores, principalmente com a utilização de algoritmos de inteligência artificial e outras novas tecnologias, que permitem a invasão da privacidade e práticas discriminatórias em prejuízos dos trabalhadores.

O aumento do uso de novas tecnologias resulta na crise do direito do trabalhado, direito fundamental, que se encontra ameaçado por tecnologias mais baratas que o custo do trabalhador para as empresas e que os controlam e as substituem, dominam esse novo cenário de busca incessante do lucro por parte do poder econômico das empresas.

A inteligência artificial cerca nosso dia-a-dia e os algoritmos de aprendizado de máquina aprendem e reconhecem padrões em dados existentes e os aplicam a novas instâncias para replicar rapidamente o que lhes foi ensinado. Por um lado temos a redução do fator de erro humano e a aceleração de processos, levando menos de um segundo para que uma decisão que um um trabalhador humano levaria várias horas.

Ao montarmos um **algoritmo**, dividimos o problema apresentado em três fases fundamentais representadas na figura 1.

Entrada: corresponde aos **dados de entrada** do **algoritmo**.

Processamento: são os procedimentos utilizados para se alcançar um resultado final.

Saída: que são os **dados** já processados.

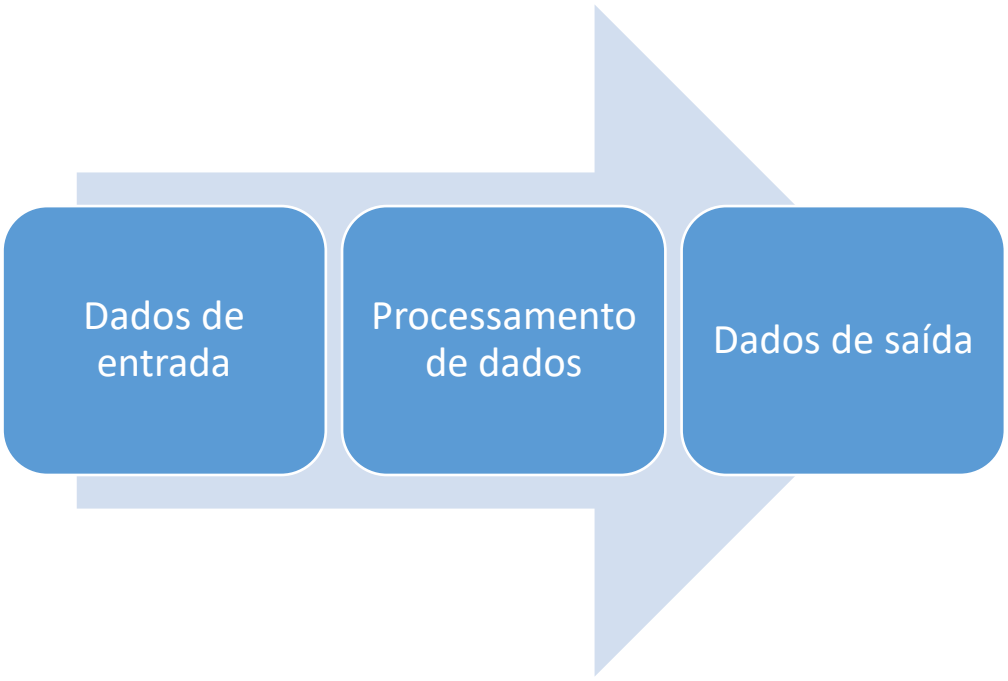

Figura 1:algoritmo- fases

A informática nasceu inicialmente para auxiliar os homens em tarefas rotineiras e repetitivas e a Inteligência Artificial (IA), que inclui estatística, cálculo e programação também é muito bem-vinda na eliminação destas tarefas, assim como aquelas repetitivas e mais pesadas para os seres humanos.

Ocorre que podemos nos deparar com vieses discriminatórios oriundos de esteriótipos. Um algoritmo que concede um salário mais alto a um empregado homem do que a uma mulher terá viés discriminatório sexista, assim como aquele que priorize ou elimine homens ou mulheres, mais jovens, ou idosos, terá viés por idade.

Este fato pode ocorrer, pois quando treinados, com dados históricos anteriores, de seleções para determinados ofícios, ou em uma leitura apenas estatística, estes algoritmos poderão privilegiar os homens às mulheres, apenas em decorrência destes serem prevalentes em determinadas tarefas ou trabalhos, mas muitas vezes o treinamento da máquina pode ser fruto de preconceitos inerentes aos homens e à sociedade e que são levados às máquinas que estão aprendendo com aqueles. Se quem a treina discrimina a mesma irá discriminar de forma mais intensa. Desta forma, urge a existência de algoritmos que aproximem o diálogo da tecnologia com o Direito.

Em recente estudo de 2021, que resultou no artigo de (BANDY, 2021) *Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits*, Northwestern University, USA é ressaltada a importância de um software para auditar métodos de IA e sua importância pública, sendo destacado que existem ainda poucos estudos a respeito, não obstante a importância atual do tema com a utilização e criação de novas tecnologias de forma cada vez mais intensa. Nesse trabalho, os estudos são sintetizados e organizados principalmente por comportamento, incluindo discriminação, distorção, exploração e erro de julgamento, com códigos também fornecidos para o domínio e organização, como, por exemplo, em grandes empresas como o Google, Facebook, Amazon, entre outras e os métodos de auditoria, por exemplo, fantoche de meia, scrape direto, crowdsourcing, etc. A revisão mostra como estudos de auditoria anteriores expuseram o contato com o público e algoritmos que exibem comportamento problemático, como algoritmos de pesquisa culpados de discriminação.

Os estudos revisados também sugerem alguns comportamentos, incluindo-se discriminação com base em identidades intersetoriais, domínios como algoritmos de publicidade, métodos de auditoria de código em organizações como Twitter, TikTok e LinkedIn, os quais exigem atenção de uma auditoria futura. O trabalho é concluido oferecendo os ingredientes comuns de auditorias bem-sucedidas e discutindo a auditoria de algoritmo no contexto de pesquisa mais ampla trabalhando em direção à justiça algorítmica. (BANDY, 2021)

Nesse contexto, cabe-se ressaltar que um comportamento é dito problemático quando causa dano ou dano potencial. Destaca-se no contexto brasileiro que estes aspectos são abordados, nos termos da recente Lei Geral de Proteção de Dados (LGPD), que entrou em vigor final de 2020, seguida das sanções da Autoridade Nacional de Proteção de Dados (ANPD), a partir de agosto passado de 2021, a qual poderá auditar estes aspectos discriminatórios e impor altas sanções, em decorrência desta desconformidade, já que existe lesão aos direitos de personalidade, sendo que

a legislação, assim como a Constituição Federal, se preocupam e protegem o dano não apenas físico, mas o moral ou extrapatrimonial, que poderá trazer consequências muito graves ao cidadão, à pessoa natural, na linguagem dos direitos humanos e ao homem, (que inclui homens e mulheres), na linguagem geral.

O tema é de alta relevância, já que é ainda muito pouco explorado, com pesquisas ainda escassas e hoje o mundo é movido por tecnologias de inteligência artificial, cada vez mais utilizadas, além de ter sido previsto exatamente pela recente legislação de proteção de dados do Brasil que as máquinas poderão ser auditadas por máquinas e ainda, caso remanescentes aspectos discriminatórios que a Autoridade Nacional de Proteção de Dados poderá auditar as novas tecnologias e aplicar elevadas sanções a quem estiver em desacordo.

Este tema é interdisciplinar, sendo a Inteligência Artificial uma das áreas da ciência mais interdisciplinares e como a ciência jurídica não é matemática e estamos falando de valores humanos, princípios, a tecnologia terá de ir além de um modelo preditivo de prognósticos ou regressão. Devemos assegurar que os direitos fundamentais e humanos das pessoas naturais não sejam lesados e afastados por meio de novas tecnologias e passa a ser urgente, neste novo cenário, a criação de uma nova tecnologia para auditar estes aspectos, que têm trazido cada vez mais lesão, principalmente estes dois últimos anos e que agora poderão e deverão ser auditados por determinação legal.

O aprendizado de máquina, a partir de um conjunto de dados, poderá reforçar estereótipos, induzindo um viés, discriminando e repetindo preconceitos anteriores, mas muitas vezes uma região poderá ter mais homens, mulheres, brancos, negros, o que poderia influenciar no aprendizado da máquina que poderá discriminar em razão de idade, gênero, origem racial, entre outros.

O controlador deverá fornecer, nos termos da Lei Geral de Proteção de Dados, sempre que solicitadas, informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada, ou oriunda de métodos de inteligência artificial e em caso de não oferecimento destas informações por aquela, observando-se o segredo comercial e industrial, a Autoridade Nacional de Proteção de Dados poderá realizar auditoria para verificação de aspectos discriminatórios em tratamento automatizado de dados pessoais.

No aspecto jurídico, uma das problemáticas da *machine learning* é a possibilidade de tratamento diferenciado, de uma discriminação negativa do titular de

dados, que, automaticamente, ofenderia diretamente a Lei Geral de Proteção de Dados. O artigo 6, que traz os princípios da Lei Geral de Proteção de Dados, inciso IX, da LGPD, reza: “IX - não discriminação: impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos.” (BRASIL. Lei 13.709, 2018)

Ao tratarmos dados sensíveis, adotando uma das bases legais do artigo 11 da Lei 13.709/2018 (LGPD), estamos diante de uma prática lícita, mas desde que o tratamento não inclua vieses discriminatórios, uma vez, que, devemos atender sempre o princípio da boa-fé e os demais princípios da Lei Geral de Proteção de Dados, previstos no artigo 6º e que incluem não apenas a finalidade, adequação, necessidade, que equivale à minimização no Regulamento Geral de Proteção de Dados (RGPD) da União Europeia, como a transparência, prevenção, segurança, qualidade de dados, livre acesso, responsabilidade e prestação de contas, assim como o princípio da não discriminação. Estes 10 princípios deverão estar sempre concomitantemente presentes em todos os tratamentos de dados ou teremos uma desconformidade à Lei Geral de Proteção de Dados.

A discriminação positiva, ou real, para corrigir desigualdades, medida de correção de assimetrias e forma de justiça distributiva, protegendo minorias, não é vetada pela LGPD, mas a discriminação negativa ofende um dos princípios basilares da LGPD, o da não discriminação para fins ilícitos ou abusivos.

Sempre que estivermos diante de vieses preconceituosos em processo seletivo, ou em outras atividades de tratamento, passamos a enfrentar práticas discriminatórias de dados sensíveis por algoritmos e a empresa será responsável por esta prática, já que o risco é potencializado em tratamentos automatizados e o risco da atividade econômica é do empregador que assume o risco de discriminar.

Outra questão a analisar ainda, o robô poderá será equiparado a um empregado, ou preposto do empregador, ainda que uma máquina apenas dotada de Inteligência Artificial? Será que o robô apenas reforçou o esteriótipo já existente, que se potencializou eliminando um grupo com histórico discriminatório? Ou foi o empregado que inseriu os dados de treinamento, ou que realizou esta atividade de tratamento, com dados que foram inseridos no treinamento de máquina, ou machine learning, que sempre discriminou e ensinou a máquina a discriminar?

Nos termos do artigo 932 do Código Civil: “são também responsáveis pela reparação civil: (…) “III - o empregador ou comitente, por seus empregados, serviçais e prepostos, no exercício do trabalho que lhes competir, ou em razão dele.” (BRASIL. Lei 10.406, 2002)

2. Lei Geral de Proteção de Dados

Preocupado com os aspectos discriminatórios, que poderão ser reforçados e constatados em atividades de tratamento por meio de inteligência artificial, o legislador incluiu o artigo 20 da Lei geral de Proteção de Dados, o qual dispõe sobre os aspectos discriminatórios decorrentes de tratamentos em métodos de Inteligência Artificial em relações em geral, trabalhistas, corporativas, de consumo e crédito, sempre que envolvida lesão a direitos de personalidade:

> Art. 20. **O titular dos dados tem direito a solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais** que afetem seus interesses, incluídas as decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade. (Redação dada pela Lei nº 13.853, de 2019) Vigência
>
> § 1º O controlador deverá fornecer, sempre que solicitadas, informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada, observados os segredos comercial e industrial.
>
> § 2º Em caso de não oferecimento de informações de que trata o § 1º deste artigo baseado na observância de segredo comercial e industrial, **a autoridade nacional poderá realizar auditoria para verificação de aspectos discriminatórios em tratamento automatizado de dados pessoais**." (BRASIL. Lei 13.709, 2018) (destaques nossos)

Com base neste dispositivo garante-se a possibilidade de revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais e que afetem os seus interesses. Na União Europeia, o artigo 22 Regulamento Geral de Proteção de Dados traz a necessidade de revisão destas atividades de tratamento de dados por métodos de Inteligência Artificial, por pessoa natural, mas o mesmo não foi repetido pela legislação nacional, podendo máquinas revisar aspectos discriminatórios de máquinas.

3. Autoridade Nacional de Proteção de Dados

Em consequência surgem em grande massa situações de titulares de dados não satisfeitos e sentindo-se lesados poderão não apenas ajuizar ações individuais, mas denunciar estas práticas ilícitas à ANPD, órgão vinculado ao governo que tem por função regular alguns institutos da LGPD, no Brasil, assim como auditar e aplicar sanções que poderão alcançar o patamar de 50 (cinquenta milhões de reais) por infração:

> “Art. 52. Os agentes de tratamento de dados, em razão das infrações cometidas às normas previstas nesta Lei, ficam sujeitos às seguintes sanções administrativas aplicáveis pela autoridade nacional: (Vigência)
>
> I - advertência, com indicação de prazo para adoção de medidas corretivas;
>
> II - multa simples, de até 2% (dois por cento) do faturamento da pessoa jurídica de direito privado, grupo ou conglomerado no Brasil no seu último exercício, excluídos os tributos, limitada, no total, a R$ 50.000.000,00 (cinquenta milhões de reais) por infração;
>
> III - multa diária, observado o limite total a que se refere o inciso II;
>
> IV - publicização da infração após devidamente apurada e confirmada a sua ocorrência;
>
> V - bloqueio dos dados pessoais a que se refere a infração até a sua regularização;
>
> VI - eliminação dos dados pessoais a que se refere a infração” (BRASIL. Lei 13.709, 2018)

Não podemos nos esquecer da possibilidade de ações civis públicas por parte de entidades legitimadas, nos termos da Lei de Ação Civil Pública e mesmo de ações coletivas nos termos do Código de Defesa ao Consumidor.

Nos termos do artigo 20 da Lei Geral de Proteção de Dadso, poderá a ANPD realizar uma auditoria para verificação de aspectos discriminatórios em tratamento automatizado e por meio de outro método de Inteligência Artificial.

Wilson, 2021, destaca que é complexo para o desenvolvedor se envolver concomitantemente com as facetas sociais e jurídicas de “justiça” e a dificuldade de desenvolver um software que concretize estes valores e passar por uma auditoria de algoritmo independente para garantir a correção técnica e a responsabilidade social de seus algoritmos.

> “Academics, activists, and regulators are increasingly urging companies to develop and deploy sociotechnical systems that are fair and unbiased. Achieving this goal, however, is complex: the developer must (1) deeply engage with social and legal facets of “fairness” in a given context,
>
> (2) develop software that concretizes these values, and
>
> (3) undergo an independent algorithm audit to ensure technical correctness and social accountability of their algorithms. To date, there are few examples of companies that have transparently undertaken all three steps.” (WILSON *et al*. 2021)

4.Vieses algorítmicos em métodos de machine learning

As preocupações vêm se acentuando cada vez mais com a crescente utilização e adoção de métodos de machine learning. A adoção de técnicas de machine learning na contratação é extremamente controversa e impulsionada por preconceitos humanos, um problema antigo, mas com consequências atuais com a criação das novas tecnologias, chegando à necessidade de ser positivado pelo legislador, tendo o dispositivo legal entrado em vigor em setembro de 2020 e a autoridade fiscalizadora recém criada poderá aplicar sanções, as quais poderão ser desde advertências até altas sanções, como vimos anteriormente, a partir de agosto de 2021, caso a empresa utilize métodos de inteligência artificial que deflorem aspectos discriminatórios.

A problemática maior vem sendo no uso dos métodos de *machine learning* (ML) em recrutamento e seleção, o que precisará ser auditado por outro software, como determinado pela LGPD e em última instância não sendo eliminados estes aspectos a ANPD, vinculada ao governo, deverá possuir uma nova tecnologia, que aparentemente ainda não existe, para auditar as empresas que desrespeitem as normas de proteção de dados.

Não há razão de assumir-se, a priori, que os sistemas de ML sem domínio de contratação serão automaticamente "objetivos", "neutros" ou "livres de preconceitos". (WILSON *et al*. 2021)

4. Estudo de casos

4.1 Pymetrics

Existem algumas empresas que têm realizado as etapas anterirmente abordadas, envolvendo as questões sociais e jurídicas, com a concretização de valores, passando por uma auditoria de algoritmo independente para garantir a correção técnica e a responsabilidade social de seus algoritmos. Como exemplo trazemos a Pymetrics, uma startup que oferece um serviço de triagem de candidatos, na avaliação pré-emprego, para os empregadores, com base nos dados e aprendizado de máquina. A pymetrics proativamente elimina os aspectos discriminatórios da *machine learning* antes da implantação para cumprir com as normas e diretrizes sobre Procedimentos de Seleção de Funcionários. A auditoria pode ser interna ou externa, a primeira conduzida por funcionários e a segunda por especialistas externos. Neste caso, esta se trata de uma auditoria externa, mas com a facilidade de acesso ao código fonte, o qual contem os algoritmos de tratamento. (WILSON *et al*. 2021)

Com relação ao processo, apresentaram a auditoria cooperativa como uma estrutura para auditores de algoritmos externos para auditar os sistemas de empresas privadas dispostas. (WILSON *et al*. 2021)

As auditorias externas sempre são muito importantes para as empresas eliminarem aspectos discriminatórios e não correrem risco de uma auditoria embasada na própria norma de proteção de dados do Brasil pelo governo, podendo evitar as altas sanções anteriormente mencionadas da Autoridade Nacional de Proteção de Dados.

4.2 Redes neurais profundas

A aprendizagem profunda, em inglês *Deep Learning,* consiste em um ramo de aprendizado de máquina que é baseado em algoritmos que tentam modelar abstrações de alto nível de dados e os quais se utilizam de um grafo profundo, o qual possui diversas camadas de processamento, sendo as mesmas compostas por várias transformações lineares e não lineares.

No artigo *Mitigating Algorithmic Bias: Evolving an Augmentation Policy that is Non-Biasing* (SMITH, Philip; RICANEK, 2020) os autores destacam que o aprendizado profundo tem se mostrado eficaz para uma ampla variedade de circunstâncias anteriormente apenas superadas por humanos. Lembre-se que IA é uma área da Ciência da Computação que objetiva replicar habilidades próprias da inteligência humana em computadores.

As redes neurais profundas (DNNs) foram aplicadas a uma infinidade de tecnologias emergentes, como veículos autônomos, moderação de conteúdo automatizado, sistemas de deteção de intrusão e muitas aplicações de DNNs se concentram em ser capaz para entender os seres humanos. Os estudos sobre a análise de sentimento, reconhecimento de fala e o processamento de linguagem natural demonstram avanços promissores em direção a esse objetivo.

Não obstante alguns bons resultados, deparamo-nos no problema de que se o conjunto de dados que costumavam treinar uma rede neural é tendencioso, então o modelo resultante também será tendencioso. Ao considerar a idade e estimativa de gênero, o viés algorítmico pode surgir devido a grupos não representados na população.

Alguns subgrupos podem ser mais difíceis de identificar, o que exigiria mais dados, ou um modelo mais robusto para se obter resultados imparciais.

Restou provado que o aumento de dados escolhidos, como estratégia, diminui o erro, mas também pode aumentá-lo, não sendo ainda a solução. Logo, a aplicação

de políticas de aumento de dados foi mostrada para reduzir o erro em uma quantidade significativa, mas se por um lado poderá diminuir vieses para conjunto de dados maior, poderá também piorar para um conjunto de dados menor.

Os resultados deste trabalho mostram que vieses em um método de IA poderão ser reduzidos sem sacrificar o desempenho do modelo. Na verdade, o desempenho aumentou para quase todos subgrupos isolados. Os resultados foram obtidos por gênero e idade no conjunto de dados MORPH-II, e resultados para o conjuntos de dados IMDB e Wiki são relatados pela primeira vez com a finalidade de dar uma ideia de desempenho geral. Executar o algoritmo avaliando centenas de modelos é um processo caro, podendo ser solução, segundo o artigo, mas o que ainda não é conclusivo.

4.3 Auditar um algoritmo como método de pesquisa

No artigo *Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits* (BANDY, 2021) estudamos que auditar um algoritmo tornou-se um método de pesquisa importante para diagnosticar comportamentos problemáticos em sistemas algorítmicos e se questiona se os algoritmos de publicidade direcionada facilitam a discriminação? Se algoritmos de recomendação do YouTube elevam os vídeos extremistas? Se algoritmos de reconhecimento facial têm pior desempenho em mulheres de pele mais escura? E as questões continuam a se multiplicar à medida que os algoritmos ganham mais força e são mais utilizados na sociedade.

Destaca-se no artigo que apesar da crescente importância pública das auditorias de algoritmo, as quais abordam esses problemas, principalmente de vieses discriminatórios, relativamente pouco trabalho foi feito para esclarecer a trajetória passada e a agenda futura de auditoria de algoritmo. Revisões sistemáticas da literatura têm sido uma parte importante da pesquisa de computação e que ajudam na identificação de lacunas existentes. O artigo conduziu um SLR de escopo de auditorias de algoritmo, rastreando mais de 500 artigos de uma variedade de jornais e conferências. Foram identificados comportamentos de máquina problemáticos que ajudam a organizar este trabalho, bem como a criar estratégias para trabalhos futuros.

Foram identificados quatro tipos principais de comportamento problemático em sistemas algorítmicos: a discriminação, distorção, exploração e erro de julgamento, sendo a maioria dos estudos revisados focados na discriminação (N = 21) ou distorção (N = 29). Os estudos de auditoria também deram mais atenção a algoritmos de pesquisa (N = 25), algoritmos de publicidade (N = 12) e algoritmos de recomendação

(N = 8), o que ajudou a diagnosticar uma série de comportamentos problemáticos nesses sistemas.

A revisão demonstrou que as auditorias de algoritmo diagnosticaram uma série de máquinas problemáticas com discriminação algorítmica acentuada, concentrada em algoritmos de publicidade e distorção em algoritmos de pesquisa. Estes estudos fornecem evidências empíricas de que os danos públicos dos sistemas algorítmicos atuam em sistemas públicos do mundo real afetando milhões de pessoas. Esta revisão também sugeriu que algumas áreas já estão prontas para futuras auditorias de algoritmos, incluindo o tratamento discriminatório em termos de identidades intersetoriais, explorando ainda mais algoritmos de publicidade que são a espinha dorsal econômica de grandes empresas de tecnologia, e empregando métodos como auditoria de código. Sugere-se que algumas empresas merecem maior atencão e que as futuras auditorias continuem a examinar algoritmos voltados para o público, concentrando-se em comportamento problemático, incluindo vieses discriminatórios.

4.4 Algoritmos pré-emprego

No estudo do artigo *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices Manish* (RAGHAVAN; BAROCAS; KLEINBERG; LEVY, 2021) pode-se constatar o interesse crescente na utilização de algoritmos de recrutamento. Neste artigo foram documentadas e analisadas as reivindicações e práticas de empresas que oferecem algoritmos para avaliação de empregos.

Foram identificados fornecedores de algoritmos de pré-emprego para selecionar candidatos. Foram documentados seus procedimentos de desenvolvimento e validação, e avaliadas suas práticas, com foco particular nos esforços para detectar e mitigar vieses. Esta análise considerou os aspectos técnicos e perspectivas legais. Tecnicamente, foram considerados esforços para detectar e mitigar preconceitos. Foram consideradas as perspectivas legais e técnicas, em relação à coleta de dados e metas de previsão, e explorados os riscos e compensações que essas escolhas representam. Também foi discutido como as técnicas de desenviesamento algorítmico se relacionam e criam desafios para a lei.

Como as avaliações algorítmicas são construídas, validadas e examinadas para tendência? Neste trabalho, documentamos e analisamos as reivindicações e práticas de empresas que oferecem algoritmos para avaliação de empregos.

Em particular, identificamos fornecedores de algoritmos de pré-emprego avaliações (ou seja, algoritmos para selecionar candidatos), documentar o que eles divulgaram seus procedimentos de desenvolvimento e validação, e avaliaram suas

práticas, com foco particular nos esforços para detectar e mitigar vieses. Foram considerados aspectos técnicos e as perspectivas legais. Tecnicamente, foram consideradas várias opções de fornecedores, o que que fazem em relação à coleta de dados e as de previsão, de exploração dos riscos e compensações que essas escolhas representam. Também se discutiram técnicas de desviesamento algorítmico que fazem interface com, e criam desafios para a lei antidiscriminação.

Em nossa análise, foram encontramos 18 fornecedores que fornecem serviços baseados em algoritmos de avaliações pré-emprego. Os tipos de avaliações oferecidos variam por fornecedor. Os tipos de avaliação mais populares foram perguntas (11 fornecedores), análise de entrevista de vídeo (6 fornecedores) e jogabilidade (por exemplo,quebra-cabeças ou jogos de vídeo) (6 fornecedores).. A maioria dos fornecedores (15) oferece avaliações personalizadas ou personalizáveis, adaptando a avaliação ao dados específicos do cliente ou requisitos de trabalho.

Os fornecedores em geral deixam para os clientes determinar quais resultados eles desejam prever, incluindo, por exemplo, análise de desempenho, número de vendas e tempo de retenção. Estes determinam qual conjunto pré-determinado de competências são mais relevantes para o trabalho específico.

Neste trabalho foi feita uma análise aprofundada das práticas relacionadas a vieses de fornecedores de avaliação algorítmica e as descobertas tiveram amplas implicações para a descoberta de sistemas algorítmicos e socioténicos e que dada a natureza sensível dos modelos construídos é inviável uma realização de uma auditoria tradicional e que em termos gerais os modelos resultam da aplicação das práticas de um fornecedor. Aprendendo sobre esta prática pode-se tirar conclusões e levantar questões sobre o modelo resultante.

4.5 Dependência algorítmica e preconceitos

No artigo *Bias in algorithmic decision-making* (ROVATSOS; MITTELSTADT; KOENE, 2020) verificamos que à medida que nossas sociedades se tornam cada vez mais dependentes de algoritmos, refletindo a volta de preconceitos em forma digital. Mas o os sistemas algorítmicos que usamos também têm o potencial de amplificar, acentuar e sistematizar nossos vieses em uma escala sem precedentes, ao mesmo tempo que apresentam a aparência de objetivos, neutros árbitros. Este resumo da paisagem reúne a literatura e os debates em torno do viés algorítmico,os métodos e estratégias que podem ajudar a mitigar seu impacto, e explora quatro setores em que este fenômeno já está começando a ter consequências no mundo real - serviços financeiros, governo local, crime, justiça e recrutamento. Identificamos um estudo de

caso para cada setor que pode ter consequências significativas para indivíduos e grupos no Reino Unido: empréstimo algorítmico redlining, bem-estar infantil, avaliações de risco de infratores e seleção de currículos.

O artigo foi estudado anteriormente, no caso atinente ao processo de recrutamento, mais uma vez, demonstrando as indagações e dificuldades mais uma vez desta temática. No recrutamento online, os algoritmos agora são frequentemente usados para filtrar automaticamente os formulários de emprego, com base em critérios definidos, que criam uma lista restrita para recrutadores humanos, para depois filtrar manualmente. A Unilever é um exemplo proeminente de uma empresa que está usando algoritmos para rastrear seus candidatos. Todos os anos, esta empresa processa mais de 1,8 milhões de pedidos de emprego e recruta mais de 30.000 trabalhadores. Eles contrataram a empresa Pymetrics, a qual é especialista em recrutamento de IA, para criar uma plataforma online, que realiza avaliações de triagem inicial. O chefe de RH da empresa disse que com este tratamento economizou cerca de 70.000 horas-pessoa de trabalhadores para entrevistar e avaliar candidatos.

Um dos problemas potenciais com este tipo de abordagem é que os dados que eles estão usando para treinar seus algoritmos às vezes refletem e perpetuam estereótipos arraigados e suposições sobre gênero e raça que continuam existindo até hoje. Por exemplo, um estudo descobriu que as ferramentas da PNL podem aprender a associar nomes afro-americanos a nomes negativos e nomes femininos com trabalho doméstico, em vez de profissionais ou outras ocupações. Outro estudo descobriu de forma semelhante que os sistemas treinados aprenderam a associar mulheres à família e às artes e humanidades, enquanto os homens foram associados a carreiras, matemática e ciências.

4.6 Amazon

Embora esses pesquisadores observem que essas associações são precisas na medida em que refletem tendências e preconceitos do mundo real, apresentam problemas quando as empresas procuram romper com os padrões históricos de emprego, diversificando sua força de trabalho. Por exemplo, relatórios em 2018 afirmavam que a Amazon havia interrompido o desenvolvimento de uma máquina "sexista", uma ferramenta baseada em aprendizagem desenvolvida em seu escritório de Edimburgo para auxiliar o recrutamento interno devido a preocupações com os preconceitos de gênero que poderia incorporar. O desenvolvimento desta ferramenta envolveu treinamento de até 500 modelos diferentes para reconhecer até 50.000 termos relevantes nos currículos dos candidatos, mas a empresa descobriu que pegou

mais os termos que os candidatos do sexo masculino usavam em seus currículos frequentemente quando vem com recomendações. Não está claro se a Amazon abandonou este projeto principalmente devido a essas preocupações, pois parece que geralmente não estava produzindo resultados.

Embora o uso de IA tenha o potencial de melhorar a precisão e ser um método econômico de filtrar funcionários em potencial durante o processo de recrutamento, há uma necessidade de pesquisar como os algoritmos usados desta forma podem levar a resultados diferentes para redes sociais e grupos étnicos.

Mais uma vez, constatamos a preocupação com os aspectos discriminatórios em recrutamento e seleção e a necessidade de superar este problema e lembremos que a Autoridade Nacional de Proteção de Dados poderá auditar aspectos discriminatórios em métodos de IA.

CONCLUSÃO

No presente artigo, pesquisamos os vieses discriminatórios que vêm sendo também constatados não apenas no Brasil, como em outros países, incluindo métodos de *machine learning* e *deep learning,* com o surgimento de um arcabouço envolvendo algoritmos, ferramentas de análise e boas práticas que permitam auditar os aspectos discriminatórios das empresas privadas. Destacamos que este software também poderá ser utilizado pelo governo, para auditoria, nos termos da LGPD.

Demonstrou-se a relevância e dificuldade do aprendizado supervisionado de máquina e o porquê há geralmente reforço dos estereótipos tecnicamente acarretando vieses discriminatórios.

A justificativa de coleta de dados em um processo seletivo deverá ser mérito para uma vaga específica e a necessidade destes dados para cada finalidade, devendo ser assegurada a privacidade por padrão, *privacy by default,* sub príncipio do *privacy by design*, em segurança da informação e devendo ser sempre garantida a proteção de dados dos titulares e consequentemente a sua privacidade, que lhe é umbilcalmente inerente.

Destacamos aqui, por fim, a importância da criação de um arcabouço técnico-jurídico para auditar os aspectos discriminatórios dos programas de recrutamento e seleção, além do papel fundamental da função fiscalizatória e de auditoria da Autoridade Nacional de Proteção de Dados, permitida pela Lei Geral de Proteção de Dados.

Referências

Bandy(2021) *Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits*, Northwestern University,

Lei Geral de Proteção de Dados (LGPD), 2020-Brasil.

Raghavan; Barocas; Kleinberg; Levy(2021) *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices Manish*

Smith, P.; Ricanek (2020) *Mitigating Algorithmic Bias: Evolving an Augmentation Policy that is Non-Biasing*