UNIVERSIDADE PRESBITERIANA MACKENZIE

RAFAEL AFONSO CRISTINO SOUSA BARROS

DECISÕES AUTOMATIZADAS E VIÉS ALGORÍTMICO: IDENTIFICAÇÃO, RESPONSABILIDADE E PERSPECTIVAS

São Paulo

2021

RAFAEL AFONSO CRISTINO SOUSA BARROS

DECISÕES AUTOMATIZADAS E VIÉS ALGORÍTMICO: IDENTIFICAÇÃO, RESPONSABILIDADE E PERSPECTIVAS

Trabalho de Graduação Interdisciplinar apresentado como requisito para obtenção do título de Bacharel no Curso de Direito da Universidade Presbiteriana Mackenzie.

Orientador: Prof. Ivandick Cruzelles Rodrigues

São Paulo

2021

RAFAEL AFONSO CRISTINO SOUSA BARROS

DECISÕES AUTOMATIZADAS E VIÉS ALGORITMICO: IDENTIFICAÇÃO, RESPONSIBILIDADE E PERSPECTIVAS

Trabalho de Graduação Interdisciplinar apresentado como requisito para obtenção do título de Bacharel no Curso de Direito da Universidade Presbiteriana Mackenzie.

Orientador: Prof. Ivandick Cruzelles Rodrigues

Aprovado em: ___/___/___

BANCA EXAMINADORA

Examinador: Prof. Ivandick Cruzelles Rodrigues

Examinador(a):

Examinador(a):

Dedico este trabalho à causa do livre acesso ao conhecimento.

AGRADECIMENTOS

A realização desse Trabalho de Conclusão de Curso nunca teria sido possível sozinho, reservo essa página para agradecer alguns daqueles que, da sua forma, possibilitaram essa realização.

À minha família, meus pais, Magda e Reumir, e minha avó, Santa, pelo apoio incondicional e carinho, e por todo incentivo à minha curiosidade.

À Tomomi, pelo incentivo e amor, recebido mesmo quando distantes ou em fusos diferentes, sem o qual não seria o mesmo.

Ao Gabriel, Otávio, Mateus, George, Leonardo, pelos anos de amizade, confusões e sucessos que guardarei com saudades.

Ao Galdino e Juliana, amizades de longa data, cuja confiança e parceria sempre me motiva a ser melhor.

Ao Thiago, que continua a servir como fonte de inspiração em meu trabalho, em sua humanidade e atenção.

A todos aqueles que de alguma forma, me ensinando ou desafiando nesses anos de graduação, marcaram meus anos de formação em memórias e aprendizados que levarei adiante.

RESUMO: Diante da crescente implementação de algoritmos e de sua sempre ascendente complexidade há também uma preocupação generalizada sobre a natureza desses sistemas invisíveis que parecem reger nossas vidas, do consumo de mídia e nossos gostos pessoais, até a decisões importantes, como concessão de empréstimos ou cálculos de seguro, questiona-se como garantir a justeza dessas decisões tomadas por 'máquinas'. O presente trabalho busca apresentar perspectivas sobre a possibilidade de tomada de decisões enviesadas por algoritmos, utilizando-se de critérios discriminatórios, bem como, se confirmada a possibilidade, maneiras de se identificar, classificar e possivelmente garantir responsabilização por eventual dano causado. A pesquisa realizada adotou uma perspectiva inicialmente técnica, abordando o conceito, funcionalidade e aplicabilidade de algoritmos e redes neurais artificiais (RNA) para então abordar aspectos relacionados, como a discriminação algorítmica e aspectos legais e regulatórios na legislação brasileira, em linha com uma abordagem interdisciplinar.

Palavras-chaves: Algoritmos, discriminação, decisões automatizadas, LGPD.

ABSTRACT: With the increasing implementation of algorithms and their ever-increasing complexity there is also widespread concern about the nature of these invisible systems that seem to govern our lives, from media consumption and our personal tastes to important decisions such as granting loans or insurance calculations, the question is how to ensure the fairness of these decisions made by 'machines'. This paper seeks to present perspectives on the possibility of biased decision making by algorithms, using discriminatory criteria, as well as, if confirmed the possibility, ways to identify, classify and possibly ensure accountability for any damage caused. The research adopted an initially technical perspective, addressing the concept, functionality and applicability of algorithms and artificial neural networks (ANN) to then address subsidiary aspects, such as the algorithmic discrimination, also addressing legal and regulatory aspects in Brazilian Law, in line with an interdisciplinary approach.

Key-Words: Algorithms, Discrimination, Automated Decisions, LGPD.

“Tudo agora, temos que admitir, encontra-se em nossas mãos; não temos o direito de supor outra coisa.”

(James Baldwin. Da próxima vez, o fogo)

Sumário

1. Introdução

Diante da crescente implementação de algoritmos e de sua sempre ascendente complexidade, há também uma preocupação generalizada sobre a natureza desses sistemas invisíveis que parecem reger nossas vidas, do consumo de mídia e nossos gostos pessoais, até a decisões importantes, como concessão de empréstimos ou cálculos de seguro, como garantir a justeza dessas decisões tomadas por 'máquinas'?

Os exemplos na mídia são diversos, mais recentes como em séries populares como Black Mirror ou ainda no início da explosão da bolha dotcom, com Matrix. A tecnologia, especialmente a Inteligência Artificial, nos cativa, especialmente pelas suas possíveis consequências desastrosas. Apesar de se tratar de ficção, tais preocupações evidenciam a necessidade de melhor análise dessas ferramentas, não por óticas alarmistas, mas sim por meio de uma análise cuidadosa da possibilidade de tomada de decisões danosas por algoritmos, das consequências de decisões discriminatórias e os riscos sociais ali envolvidos.

O presente trabalho busca apresentar perspectivas sobre a possibilidade de tomada de decisões enviesadas por algoritmos, utilizando-se de critérios discriminatórios, bem como, se confirmada a possibilidade, maneiras de se identificar, classificar e possivelmente garantir responsabilização por eventual dano causado.

A própria questão do reconhecimento da personalidade jurídica, suposto impedimento para a responsabilização por danos causados por decisões tomadas de maneira totalmente automatizada encontra paralelo na cultura, toma por pressuposto que a responsabilização por danos causados por decisões tomadas por 'robôs' só seria possível quando I.As como a "Skynet", 'GLaDOS' ou 'H.A.L.' estivessem respondendo penal ou civilmente por danos causados por suas decisões. Esse trabalho também possuí como objetivo desmistificar tal possibilidade, apontando a realidade do desenvolvimento de Inteligência Artificial e possibilidades legais de responsabilização.

Ao longo deste trabalho de conclusão buscará se pontuar também as limitações das ferramentas que possuímos no momento para a identificação de decisões enviesadas e até mesmo no que poderá ser realizado para a prevenção de futuros episódios discriminatórios.

Para podermos nos aprofundar na discussão, é necessário delimitar precisamente o que é um Algoritmo e quais são suas características fundamentais. Nesse sentido, adiante foi preparada introdução técnica sobre o funcionamento e estrutura de algoritmos, bem como sobre a sua pretensa neutralidade.

2. O que são Algoritmos

Conforme a definição apresentada pelos autores Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest e Clifford Stein na imprensa do Instituto de Tecnologia de Massachusetts (MIT), um algoritmo pode ser definido como um procedimento computacional utilizado como ferramenta para a solução de um problema computacional específico.[1]

Um algoritmo, como processo computacional, recebe valores como *input* e retorna valores, ou grupo de valores como *output*. Portanto, um algoritmo é uma sequência de passos de transformação do *input* em *output*.

As possíveis aplicações de um algoritmo se estendem desde simples aplicações de ordenação e organização de dados, até algoritmos que auxiliam no de sequenciamento genético do DNA humano, comércio eletrônico, alocação de recursos escassos, punição criminal ou mesmo viabilidade de recebimento de benefícios sociais.

Há de se notar que, mesmo para as tarefas simples, como a mencionada organização de dados, há diversas opções de algoritmos que conseguem obter o resultado desejado, resolvendo o problema computacional, no entanto, utilizando de métodos distintos. No mesmo sentido, alguns algoritmos podem ser mais eficientes em reorganizar alguns poucos itens, ou se especializar na reorganização rápida, enquanto outros algoritmos podem ser mais adequados à reorganização de largas quantidades de dados, com pouca organização prévia.

Assim, um algoritmo, apesar de objetivar a solução de um problema computacional específico, pode assumir formas diversas com o mesmo objetivo, visando maior especialização, considerando questões como hardware e nuances de sua implementação.[2]

---

[1] CORMEN, Thomas H.; LEISERSON, Charles E.; RIVEST Ronald L.; STEIN Clifford. "**Introduction to Algorithms**" The MIT Press, Estados Unidos, Massachusetts, 2009 p. 5

[2] Idem. **"Introduction to Algorithms"** Estados Unidos, Massachusetts, 2009 p. 13

Algoritmos independem de linguagens específicas e funcionam como receitas para a solução de um problema computacional específico. A crescente complexidade dos problemas computacionais que necessitam de soluções impõe desafios que, por vezes, requerem o que é chamado de algoritmos de *machine learning*.

O *machine learning*, em contraste aos algoritmos programados diretamente por cientistas e engenheiros, são desenvolvidos de maneira automática a partir da inserção de dados de treinamento. Como explicam Michael Kearns e Aaron Roth[3], algoritmos de *machine learning* se assemelham à um processo de "autoprogramação" a partir dos dados inseridos, podendo servir enquanto base para o desenvolvimento de um algoritmo de previsão com base em tendências históricas:

> No design tradicional do algoritmo, embora o output possa ser útil (como uma lista ordenada do tempo de uso do Facebook, que poderia ajudar na análise das características demográficas dos usuários mais engajados), esse output não é, em si mesmo, outro algoritmo que pode ser aplicado diretamente a outros dados. Em contraste, no aprendizado de máquinas, esse é o cerne da questão. Por exemplo, pense em pegar um banco de dados de informações do ensino médio referentes a estudantes que já se encontram em faculdades, alguns dos quais se formaram na faculdade e outros não, e usá-lo para derivar um modelo que preveja a probabilidade de graduação para futuros estudantes desta escola de ensino médio. Em vez de tentar definir diretamente um algoritmo para fazer essas previsões - o que poderia ser bastante difícil e sutil - escrevemos um meta-algoritmo que usa os dados históricos para obter nosso modelo ou algoritmo de previsão. O aprendizado da máquina é às vezes considerado uma forma de "autoprogramação", uma vez que são principalmente os dados que determinam a forma pormenorizada do modelo aprendido.[4] (tradução nossa)

Algoritmos, inclusive aqueles desenvolvidos com base em *machine learning*, vêm assumindo uma posição crítica no desenvolvimento, organização e gerenciamento social de organizações, sejam elas particulares ou públicas, bem como o acesso à informação e diversos elementos de nossa vida diária. Essa importância crítica encontra contraste na ausência de

[3] KEARNS, Michael. ROTH, Aaron. **"The Ethical Algorithm",** Oxford University Press, New York, 2020. p.n.p

[4] In traditional algorithm design, while the output might be useful (like a sorted list of Facebook usage times, which could help in analyzing the demographic properties of the most engaged users), that output is not itself another algorithm that can be directly applied to further data. In contrast, in *machine learning*, that's the entire point. For example, think about taking a database of high school information about previously admitted college students, some of whom graduated from college and some of whom did not, and using it to derive a model predicting the likelihood of graduation for future applicants. Rather than trying to directly specify an algorithm for making these predictions—which could be quite difficult and subtle—we write a meta-algorithm that uses the historical data to derive our model or prediction algorithm. *Machine learning* is sometimes considered a form of "self-programming," since it's primarily the data that determines the detailed form of the learned model.

transparência ou compreensão sobre a tomada de decisões que levaram ao design específico da solução técnica.

Para os propósitos deste texto os termos Algoritmos, Inteligência Artificial e Deep Learning serão utilizados de maneira intercambiável, com os termos IA e algoritmos descrevendo um modelo matemático, utilizado como solução de problemas computacionais, e o *machine learning* como campo técnico inserido na área.

O *machine learning* utiliza de redes neurais artificiais (RNA), sistema inspirado na rede biológica de neurônios, com funções matemáticas servindo como neurônios artificiais do modelo. Esses neurônios são organizados em diversos níveis na RNA, reagindo à *inputs* bem como transmitindo sinais entre neurônios. Sendo assim, dados de entrada podem viajar diversas vezes entre cada neurônio artificial, que pode ser configurado, sendo atribuídos pesos e medidas diferentes para a sua ativação, bem como podem ser estruturados níveis ocultos entre o nível de *input* e *output*. Esses pesos, medidas e níveis atribuídos à cada neurônio artificial, que compõe a topologia de uma rede neural artificial, requerem otimização estendida para alcançar os objetivos desejados pela RNA, uma das maneiras comuns de se alcançar esses resultados se dá por meio da inserção de dados já tratados e identificados.

Especificamente no caso da utilização de *machine learning* para o desenvolvimento da solução técnica, somos confrontados com a possibilidade da ausência de transparência direta da solução implementada por parte dos engenheiros envolvidos em seu desenvolvimento, o que pode levar a consequências diretas sobre a vida diária de milhares de pessoas, assim, enquanto o texto trabalhara a Inteligência Artificial como um todo, será dado destaque especial às considerações envolvendo o *machine learning*.

Tal preocupação com transparência e explicabilidade das soluções para problemas computacionais pode parecer um excesso de zelo, mas, como mencionado anteriormente, algoritmos vêm tomando uma função crítica em nossa vida diária e como toda tecnologia ou solução técnica, especialmente uma tão avançada e complexa, não há como falar sobre sua ***neutralidade***.

3. Neutralidade

O código, enquanto ferramenta que possibilita o desenvolvimento de programas de diversas utilidades é costumeiramente tratado de maneira objetiva e técnica, sendo enxergado como uma ferramenta direta de resolução de problemas computacionais e apenas isso, busca-se nesse capítulo precisar as razões pelas quais tal tecnologia não pode ser enxergada ou tratada desta maneira.

Tratando-se de uma solução técnica, objetiva e supostamente imparcial, não haveria como algoritmos não serem, por falta de melhor termo, neutros.

Tal característica dos algoritmos é apresentada em resposta à questionamentos quanto a ética, sobre o processamento de dados, ou desenvolvimento de programas e aplicações com consequências sociais danosas.

Laura Denardis, em seu livro de 2014 "The internet in Everything"[5], aponta o erro em entender não só algoritmos, mas qualquer ponto técnico de controle como sendo neutros. Ao comparar as limitações físicas do desenvolvimento da tecnologia e como isso a influencia e afeta, com a realidade e construção social que contextualiza o desenvolvimento e aplicação desta tecnologia, Denardis explicita a necessária discussão sobre tecnologia e poder:

> Os pontos técnicos de controle não são neutros - eles são locais de luta por valores e arenas de poder para mediar interesses concorrentes. Ao mesmo tempo, o mundo natural e físico, é claro, existe. O processo científico e a inovação incorporam fatos sobre o mundo físico derivados da experiência material vivida. Do ponto de vista da engenharia, não é possível construir um propulsor de foguetes a partir de tufos de grama, não importa a vontade daqueles que assim desejam. Compreender a política da tecnologia requer o reconhecimento das realidades da engenharia material e a construção social da mesma.[6]

[5] DENARDIS, Laura "**The internet in Everything: freedom and security in a world with no off switch**", Yale University Book Press, Connecticut, 2020. p. 18

[6] "Technical points of control are not neutral—they are sites of struggle over values and power arenas for mediating competing interests. At the same time, the natural and physical world, of course, exists. The scientific process and innovation incorporate facts about the physical world derived from lived material experience. From an engineering perspective, it is not possible to construct a solid rocket booster out of lawn clippings, no matter what powerful values will it so. Understanding the politics of technology requires acknowledging both material engineering realities and also the social construction of the same".

Denardis referencia o ensaio “Artefatos tem política” de Langdon Winner. Segundo Winner, artefatos, incluindo “as máquinas, as estruturas e os sistemas da cultura material”, podem conter propriedades políticas em pelo menos duas maneiras. A primeira, em razão da sua utilização prática e implementação material em certa sociedade, e a segunda, em razão da sua própria estrutura, sendo uma “tecnologia inerentemente política.”

Algoritmos, enquanto tecnologia e invenção, utilizada para a solução de problemas de computação, advindos de uma comunidade particular, marcada por suas especificidades sociais, para Langdon Winner, e para os propósitos desta dissertação não são necessariamente neutros.

A discussão quanto à discriminação algorítmica depende da possibilidade de pontuar não apenas “erros” no desenvolvimento de programas em específico, mas também de se discutir e questionar o desenvolvimento e efeitos de programas a partir de uma perspectiva social, política e preocupada com a preservação de direitos coletivos, objetivando maior transparência quanto aos programas que cada dia mais afetam nossa vida diária.

Podemos traçar comparações entre os conflitos introduzidos pelo código e tecnologias de informação com os desafios enfrentados durante a Revolução Industrial. Na época tivemos avanços incríveis que eram inimagináveis até a introdução do maquinário e a indústria. No entanto, tivemos também uma intensificação da exploração de trabalhadores, poucas preocupações quanto à sua saúde e nenhuma preocupação em relação à degradação do meio ambiente, como argumenta Cathy O’Neil [7].

> "Em certo sentido, nossa sociedade está enfrentando uma nova revolução industrial. E podemos tirar algumas lições da última. A virada do século XX foi uma época de grande progresso. As pessoas podiam iluminar suas casas com eletricidade e aquecê-las com carvão. As modernas ferrovias trouxeram carne, vegetais e conservas de um continente distante. Para muitos, a boa vida estava ficando melhor. No entanto, este progresso tinha um lado negativo horripilante. Era alimentada por trabalhadores horrivelmente explorados, muitos deles crianças. Na ausência de regulamentos de saúde ou segurança, as minas de carvão eram armadilhas mortais. Só em 1907, 3.242 mineiros morreram. Os frigoríficos trabalhavam doze a quinze horas por dia em condições imundas e muitas vezes enviavam produtos tóxicos. A Armour and Co. despachou latas de carne podre em toneladas para as tropas do exército americano, usando uma camada de ácido bórico para mascarar o mau cheiro. Enquanto isso, monopolistas vorazes dominavam as ferrovias, empresas de

[7] O’ NEIL, Cathy “**Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy”** Crown, New York, 2016 - Pag. 162

> energia e serviços públicos e aumentavam as tarifas dos clientes, o que representava um encargo para a economia nacional". (tradução nossa)[8]

A Revolução Industrial nos oferece algumas lições quanto ao desenvolvimento de novas tecnologias. O seu início se dá na busca por ferramentas e instrumentos que permitiriam o aumento de produtividade do trabalho, o que por sua vez levou ao desenvolvimento do maquinário, resultando no barateamento de mercadorias e na redução do trabalho necessário a ser investido por parte dos trabalhadores para que a mesma quantidade de mercadoria fosse produzida.

Logicamente, a situação narrada durante a Revolução Industrial deveria levar à uma redução de horas de trabalho ou à melhores condições de vida. No entanto, como sabemos, o surgimento do maquinário não reduziu as horas de trabalho; ele, de fato, as estendeu.

O maquinário, enquanto poderia trazer à humanidade uma melhora generalizada de qualidade de vida resultou, de imediato, na melhoria de vida para uma pequena parcela, enquanto trabalhadores se encontravam em regimes cada vez mais intensos em fábricas insalubres e com baixos salários. A ferramenta que deveria aliviar a carga de trabalho a intensificou, a vitória do homem sobre a natureza resultou em maior submissão do homem às forças produtivas, no entanto, não de todos os homens.

Esse resultado contraditório não tem suas raízes em alguma característica essencialmente negativa do maquinário, mas é resultante da sua caracterização enquanto ferramenta, o seu desenvolvimento, aprimoramento e, principalmente, aplicação.

Na época, movimentos surgiram em revolta contra a introdução do maquinário, apontando os efeitos negativos que a sua aplicação de maneira a beneficiar apenas parcela da

[8] “In a sense, our society is struggling with a new industrial revolution. And we can draw some lessons from the last one. The turn of the twentieth century was a time of great progress. People could light their houses with electricity and heat them with coal. Modern railroads brought in meat, vegetables, and canned goods from a continent away. For many, the good life was getting better. Yet this progress had a gruesome underside. It was powered by horribly exploited workers, many of them children. In the absence of health or safety regulations, coal mines were death traps. In 1907 alone, 3,242 miners died. Meatpackers worked twelve to fifteen hours a day in filthy conditions and often shipped toxic products. Armour and Co. dispatched cans of rotten beef by the ton to US Army troops, using a layer of boric acid to mask the stench. Meanwhile, rapacious monopolists dominated the railroads, energy companies, and utilities and jacked up customers’ rates, which amounted to a tax on the national economy”

sociedade vinham trazendo, em 1821, até mesmo o autor John Stuart Mill, um dos primeiros pensadores do capital, em seu livro "Princípios da Política Econômica" reconhece que era de se duvidar se a introdução do maquinário teria aliviado o trabalho do ser humano.[9]

É claro que o maquinário que não pode ser culpado por sua utilização, mas também não podemos entendê-lo como neutro. A sua utilização, desenvolvimento e aprimoramento são decisões que devem ser compreendidas como sendo essencialmente políticas, pois possuíram consequências na forma em que organizamos nossa vida diária, nos forçando a questionar a sua ética, propósito e objetivo.

Também há semelhanças entre o desenvolvimento do maquinário e as tecnologias surgidas na revolução informacional: ambas surgiram de maneira incremental e correspondendo à uma demanda de aceleração produtiva, objetivando maior integração internacional do comércio, escalas e cadeias produtivas mais complexas.

Quanto a tais desenvolvimentos, Leonardo Parentoni, durante o painel "Discriminações algorítmicas: impactos na sociedade, perspectivas e soluções" realizado no Fórum da Internet do Brasil de 2020[10] apresentou sua hipótese quanto à presença de vieses em algoritmos e a maneira como a sua presença foi tolerada, correlacionando o seu surgimento com as prioridades estabelecidas para o desenvolvimento de protocolos da rede DARPA, que viria a se tornar a Internet.

Inclusive, Yasha Levine em sua obra, "Surveillance Valley: The Secret Military History of the Internet"[11] apresenta registros de arquivo demonstrando o processo de desenvolvimento da Internet, desde seu início enquanto ARPANET, e como as necessidades militares e de segurança influenciaram as prioridades estabelecidas em seu desenvolvimento:

> Mesmo o primeiro teste bem-sucedido da rede TCP/IP de nível Internet, realizado em 22 de novembro de 1977, simulou um cenário militar: o uso de

[9] "É questionável que todas as invenções mecânicas já feitas tenham servido para aliviar a faina diária de algum ser humano" (MILL apud MARX, MARX, Karl. Maquinaria e Grande indústria. In: MARX, Karl. **O Capital: Crítica da economia política**. Livro I: O processo de produção do capital. Trad. Rubens Enderle. São Paulo: Boitempo, 2013, pp. 548)

[10] NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR. Fórum Da Internet No Brasil **"Discriminações algorítmicas: impactos na sociedade, perspectivas e soluções**". 2020. Disponível em: <https://www.youtube.com/watch?v=FN0OT0hYp8U> Acesso em em: 10/11/2020

[11] LEVINE, Yasha **"Surveillance Valley: The Secret Military History of the Internet"** Perseus Books, LLC 2018, Nova York, n.p.)

> redes de rádio, satélite e com fio para se comunicar com uma unidade móvel ativa que lutava contra uma invasão soviética da Europa. Uma antiga van de entrega GMC equipada pelo SRI com um monte de equipamentos de rádio desempenhou o papel de uma divisão motorizada da OTAN, subindo e descendo a auto-estrada perto de Stanford e transmitindo dados através da rede de rádio da ARPA. [12] (tradução nossa)

O desenvolvimento de tal protocolo priorizou a resiliência da rede, a sua eficiência de custo, adequação com diversos outros tipos de rede local, listando apenas em 7º lugar a importância de "*accountability*"[13], ou seja, a possibilidade de auditoria foi relegada ao fim da lista de prioridades. Estes eram os focos principais da expansão da ARPANET:

> 1. A comunicação pela Internet deve continuar, apesar da perda de redes ou gateways.
>
> 2. A Internet deve suportar múltiplos tipos de serviços de comunicação.
>
> 3. A arquitetura da Internet deve acomodar uma variedade de redes.
>
> 4. A arquitetura da Internet deve permitir o gerenciamento distribuído de seus recursos.
>
> 5. A arquitetura da Internet deve ser rentável.
>
> 6. A arquitetura da Internet deve permitir a conexão do host com um baixo nível de esforço.
>
> 7. Os recursos utilizados na arquitetura da Internet ser auditáveis (*accountable*)" (tradução nossa) [14]

Assim como na Revolução Industrial, certas demandas condicionaram o desenvolvimento da tecnologia e suas prioridades. Na era da informação não havia preocupação

[12] Even the first successful test of the Internet-grade TCP/IP network, which took place on November 22, 1977, simulated a military scenario: using radio, satellite, and wired networks to communicate with an active mobile unit battling a Soviet invasion of Europe. An old GMC delivery van outfitted by SRI with a bunch of radio gear played the role of a motorized NATO division, driving up and down the freeway near Stanford and beaming data over ARPA's radio network

[13] CLARK, D. D. "The Design Philosophy of the DARPA Internet Protocols". Massachusetts Institute of Technology. Proc. SIGCOMM '88, Computer Communication Review Vol. 18, No. 4, 1988.pp. 106–114.

[14] "1. Internet communication must continue despite loss of networks or gateways. 2. The Internet must support multiple types of communications service. 3. The Internet architecture must accommodate a variety of networks. 4. The Internet architecture must permit distributed management of its resources. 5. The Internet architecture must be cost effective. 6. The Internet architecture must permit host attachment with a low level of effort. 7. The resources used in the internet architecture must be accountable"

de início com a privacidade de futuros usuários, a utilização de seus dados, ou mesmo considerações éticas quanto ao desenvolvimento de programas que poderiam afetar a o coletivo de maneira negativa. Durante décadas a programação e informatização de sistemas foi tratada de maneira neutra e sinal de avanço para uma sociedade baseada na tomada racional de decisões.

No entanto, o prometido não foi entregue e o código, assim como o maquinário, não veio sem os seus próprios problemas e dilemas, como a violação de privacidade, perpetuação de violências e pobreza, racialização e discriminação de minorias. Cathy O'Neil sintetiza tal sintoma ao tratar de processos de *Big Data*:

> "Os processos de big data codificam o passado. Eles não inventam o futuro. Fazer isso requer imaginação moral, e isso é algo que só os humanos podem fornecer. Temos que incorporar explicitamente melhores valores em nossos algoritmos, criando modelos de Big Data que seguem nossa liderança ética. Às vezes isso significa colocar a justiça à frente do lucro." [15] (tradução nossa)

Enquanto o código é uma ferramenta que nos oferece, de relance, mecânicas objetivas e imparciais ao garantir legitimidade completa à uma ferramenta, sem considerar a sua aplicação e possíveis danos sociais resultantes, evitamos confrontar as consequências éticas de sua utilização e que devem ser abordadas durante o processo de desenvolvimento de códigos que viram a afetar milhares, se não milhões, de pessoas pelo mundo.

Uma ferramenta com tal capacidade de efeito não pode ser simplesmente ignorada e categorizada como neutra. Precisamos, assim como fizemos em relação ao maquinário e aos desafios trazidos por ele, apresentar medidas que evitem ou ao menos remedeiem os seus possíveis efeitos negativos, reconhecendo tratar-se não apenas de uma matéria técnica, mas uma ferramenta social.

Há de se reconhecer que, majoritariamente, aqueles que desenvolvem as soluções técnicas necessárias partem de um grupo homogêneo, com vieses implícitos. É necessário, portanto, questionarmos quais devem ser os limites no desenvolvimento de programas, as suas

[15] "Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide. We have to explicitly embed better values into our algorithms, creating Big Data models that follow our ethical lead. Sometimes that will mean putting fairness ahead of profit"

consequências e a sua utilização, procurando mapear efeitos negativos que dele decorreram, intencionais ou não.

4. Discriminação e Algoritmos

Dessa forma, considerando que não há como evitar a questão da possibilidade da presença de vícios e reflexos sociais no que se refere à tecnologia, sendo esses conscientes ou não, devemos retornar nossa atenção ao aspecto técnico da questão e compreender como tais elementos podem surgir mesmo por meio do processo de aprendizado via redes neurais.

A presença de 'erros' ou desvios com consequências danosas em algoritmos, especialmente aqueles que envolvem certa complexidade, não deve ser surpreendente para a maioria. Há exemplos famosos, como o escândalo das emissões da Volkswagen de 2015[16], onde um software foi utilizado para trapacear em testes de emissão para 11 milhões de veículos, que, ao contrário do indicado pelo software, emitiam 40 vezes o autorizado pela agência de proteção ambiental americana.

Como no exemplo acima, os efeitos desses vieses, intencionais ou não, podem ser extremamente difundidos e de difícil percepção. No caso de vieses e discriminação contra grupos já estigmatizados, perpetuando a situação em que se encontram, verifica-se consequências sociais extremamente danosas:

> Grupos estigmatizados perdem respeito social e também oportunidades materiais, sendo que estereótipos negativos legitimam o tratamento discriminatório desses grupos ao atribuir a eles a responsabilidade pela situação na qual se encontram, um trabalho que formas de racismo cultural fazem todos os dias em diversas sociedades liberais [17]

No entanto, não abordaremos todo e qualquer exemplo de vícios ou desvios, propositais ou nãos em programação, mas sim especificamente resultados que podem ser categorizados como discriminatórios. A seguir, serão trabalhados os conceitos como viés algorítmico,

[16] MEARIN, Lucas **"A diesel whodunit: How software let VW cheat on emissions"** 23 de setembro, 2015, Computer World, Estados Unidos, Disponível em: <https://tinyurl.com/khnj3s7t> Acessado em: 30/03/2021

[17] MOREIRA, Adilson José . "Pensando como um negro: ensaio de hermenêutica jurídica" São Paulo: Editora Contracorrente, 2019, n.p.

discriminação negativa e positiva e a tarefa de conceituação do que pode ser considerado uma decisão justa.

4.1 O algoritmo deve discriminar?

A questão da discriminação deve ser tratada com a nuance que merece, principalmente quando abordada no campo dos algoritmos. Por óbvio, em sua operação regular o algoritmo funciona recebendo dados e os distinguindo de acordo com sua programação, com a distinção entre dados inseridos, organização de acordo com ordem, tamanho ou certas categorias, dentre outras funções. No entanto, na discussão de viés algorítmico o ponto a ser abordado não é da discriminação em sua conceituação mais genérica, sendo essencial, assim, esclarecer que esta conceituação básica do termo 'discriminação' não pode ser comparada com a discriminação contra qual a lei garante proteção.

Dessa maneira, é necessário esclarecer que diferenciações realizadas por algoritmos não necessariamente implicam em tratamento negativo de um grupo ou indivíduo. A discriminação a qual aqui se refere, e a qual deve se combater durante o processo de elaboração, desenvolvimento e implementação de algoritmos é aquela decorrente de categorias sociais, como raça, gênero, sexo e outras categorias. A proteção garantida por lei não aborda qualquer maneira de discriminação, mas sim aquelas nocivas ao indivíduo e à sociedade.[18]

Adiante será ainda abordado a conceituação legal de discriminação, buscando esclarecer onde se encaixa conceitos como discriminação positiva e negativa na avaliação de vieses e danos causados por algoritmos.

O risco presente na racionalização da discriminação realizada por algoritmos também não pode ser ignorado. Enquanto se trata de um processo comum em diversos modelos, há de se observar que vieses continuam presentes e influenciam a tomada de decisões que levaram à escolha, design e implementação de um modelo em específico.

[18] EDER, Nikes **"Privacy, Non-Discrimination and Equal Treatment: Developing a Fundamental Rights Response to Behavioural Profiling"** em **"Algorithmic Governance and Governance of Algorithms"**, Springer, Suíça, 2021

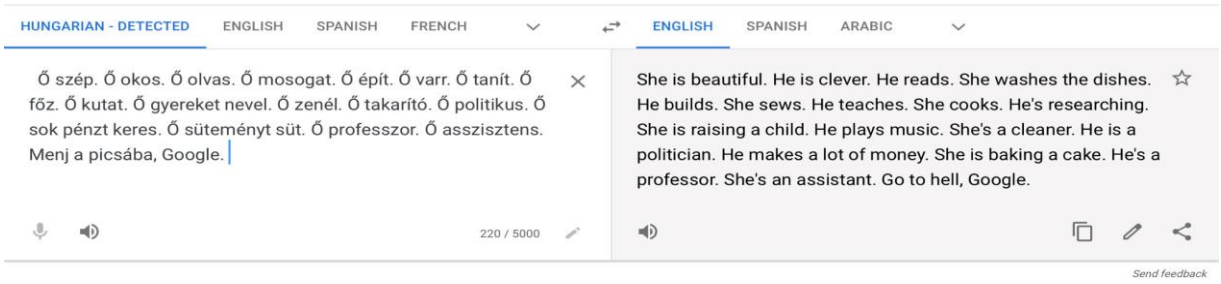
Virginia Eubanks, em sua obra “Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor”[19], sumariza os efeitos da discriminação danosa e sua contribuição na reprodução de uma sociedade desigual e como independe de esforço consciente por aqueles responsáveis pelo seu desenvolvimento:

> A discriminação racional não exige ódio de classe ou racial, ou mesmo preconceito inconsciente, para operar. Exige apenas que se ignore o preconceito que já existe. Quando ferramentas automatizadas de tomada de decisão não são construídas para desmantelar explicitamente as iniquidades estruturais, sua velocidade e escala as intensificam.[20] (tradução nossa)

A violência, segregação e discriminação reproduzidas por meio de métodos racionais, como algoritmos e outras ferramentas tecnológicas, servem para esconder e dificultar a tarefa de identificar e combater tais vieses. A seguir, serão explorados exemplos reais de vieses algorítmicos e como estes podem ocorrer.

4.2 O que é viés algorítmico

Ainda no início de 2021, foi reproduzido de maneira viral em redes sociais a discussão da “escolha” de pronomes do algoritmo de tradução automática quando traduzindo uma língua que possuí gêneros neutros, como o persa ou o húngaro, para uma língua onde o gênero de pronomes possuí mais peso, como o inglês ou português. Nos testes realizados, pode se verificar uma clara indicação de que certas atividades ou profissões tiveram pronomes associados correspondentes à estereótipos de cada gênero[21].



[19] **EUBANKS, Virgina. “Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor”. New York: St. Martin’s Press. 2018**

[20] “Rational discrimination does not require class or racial hatred, or even unconscious bias, to operate. It only requires ignoring bias that already exists. When automated decision-making tools are not built to explicitly dismantle structural inequities, their speed and scale intensify them.”

[21] CARR, Jemma “University lecturer slams 'sexist' Google Translate as gender neutral languages are translated into English with gendered pronouns suggesting men 'build' and women 'wash dishes'”24 de Março 2021, Disponível em: <https://www.dailymail.co.uk/news/article-9396937/University-lecturer-slams-sexist-Google-Translate.html> Acessado em: 30/03/2021

Coincidentemente, Michael Kearns e Aaron Roth, em sua obra "The Ethical Algorithim", de 2020, abordam tal exemplo como uma possibilidade de se identificar um viés de linguagem oculto no processo de aprendizagem de Redes Neurais:

> O problema aqui é que os dados de treinamento usados em aplicações de aprendizagem de máquinas podem frequentemente conter todos os tipos de vieses ocultos (e não tão ocultos), e o ato de construir modelos complexos a partir de tais dados pode tanto amplificar estes vieses quanto introduzir novos. Como discutimos na introdução, a aprendizagem de máquinas não lhe dará coisas como neutralidade de gênero "de graça" quando você não as pediu explicitamente. Assim, embora provavelmente muito poucos dos documentos usados para criar a palavra incorporação, se é que houve algum, exibiram um sexismo flagrante (e certamente nenhum deles realmente sugeriu que a dona de casa era a melhor análoga feminina para o programador de computador masculino), as minúsculas forças coletivas do uso da linguagem em todo o conjunto de dados, quando comprimidos em um modelo preditivo para analogias de palavras, resultaram em claro viés de gênero. E quando tais modelos então se tornam a base para serviços amplamente utilizados, tais como motores de busca, publicidade direcionada e ferramentas de contratação, o viés pode ser ainda mais propagado e até mesmo amplificado por seu alcance e escala.[22] (KEARNS, ROTH – Id. - 2020, n.p., Tradução nossa)

Como visto, neste caso a responsabilidade sobre os resultados das traduções não se tratou de um design específico ou resultado previsto do algoritmo, mas uma consequência, uma marca social, inesperada resultante da tecnologia preditiva, reproduzindo padrões passados. Tal fato não passou despercebido pelo Google, que há tempos trabalha na tentativa de reduzir vieses de gênero na tradução automática do Google Translate.[23]

O termo "Viés Algorítmico" se refere precisamente a situações como a relatada acima, no caso trata-se de uma consequência menor e mais visual de vieses implícitos presentes nos dados, mas há de se considerar a influência de tal possibilidade em algoritmos de maior escala

[22] "The problem here is that the training data used in *machine learning* applications can often contain all kinds of hidden (and not-so-hidden) biases, and the act of building complex models from such data can both amplify these biases and introduce new ones. As we discussed in the introduction, *machine learning* won't give you things like gender neutrality "for free" that you didn't explicitly ask for. Thus even though probably very few of the documents used to create the word embedding, if any, exhibited blatant sexism (and certainly none of them actually suggested that homemaker was the best female analogue for male computer programmer), the tiny collective forces of language usage throughout the entire dataset, when compressed into a predictive model for word analogies, resulted in clear gender bias. And when such models then become the basis for widely deployed services such as search engines, targeted advertising, and hiring tools, the bias can be further propagated and even amplified by their reach and scale."

[23] JOHNSON, Melvin "A Scalable Approach to Reducing Gender Bias in Google Translate" Google Ai Blog 22 de abril de 2020, Disponível em: <https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html> Acessado em: 30/03/2021

e efeito, como por exemplo aqueles que objetivam melhor e mais objetiva contratações, melhor verificação e precisão em sistemas de empréstimo e até mesmo na justiça criminal.[24]

Considerando o processo descrito anteriormente quanto o *machine learning*, há de se levantar questões quanto ao material de treinamento utilizado. Fundamentalmente esse será o fator definidor para o treinamento da maioria dos algoritmos que se utilizam de RNAs, a possibilidade de inserção de dados viciados, ou seja, que contenham vieses.

Mais grave é a possibilidade da geração de "*feedback-loops*", isso é, a aceitação e reprodução acrítica de precedentes históricos, reproduzindo estruturas discriminatórias, prologando os seus efeitos e logo após, por meio da reinserção dos resultados obtidos pelas ferramentas algorítmicas, condicionando futuramente o próprio modelo, Cathy O'Neil, em seu livro "Weapons of Math Destruction", exemplifica a questão se referindo ao sistema de combate ao crime "Predpol", que gera um modelo preditivo de crimes, que por sua vez recomenda patrulhas para policiais com bases geográficas. No entanto, O'Neil destaca que ao incluir o histórico de infrações consideradas como 'comportamento antissocial' ao modelo, o algoritmo estava direcionando mais polícias à bairros pobres e de populações marginalizadas, que, por sua vez, levou a um maior número de prisões e paradas por 'comportamento antissocial', o que logo era reinserido ao sistema, confirmando sua 'eficiência'. [25]

Com atenção, consegue-se enxergar nesse breve exemplo não só os problemas desses *feedback-loops*, internalizando e ao mesmo tempo estendendo tendências históricas, como também um exemplo claro de discriminação por *proxy*, onde não há intenção clara de discriminação contra um grupo, mas, onde por meio da utilização de um substituto, como no caso, bairro, um grupo será o principal afetado.

Em outro exemplo, levemos em conta o exemplo estabelecido pelo programa COMPAS, utilizado para análise de risco de reincidência de criminosos nos Estados americanos de Nova York, Wisconsin, California e Florida, em 2016, a agência de jornalismo investigativo ProPublica publicou matéria analisando os vieses e tratamento recebido por aqueles avaliados pelo algoritmo e pode concluir que:

[24] KEARNS, ROTH. "The Ethical Algorithm", 2020, p. n.p.
[25] O'NEIL, "Weapons of Math Destruction", 2016, n.p

> "Ao prever quem iria reincidir, o algoritmo cometeu erros com réus brancos e negros aproximadamente no mesmo ritmo, mas de maneiras muito diferentes. A fórmula era particularmente susceptível de falsamente assinalar os réus negros como futuros criminosos, rotulando-os erroneamente desta forma a quase o dobro da taxa dos réus brancos. Os réus brancos eram rotulados erroneamente como de baixo risco com mais frequência do que os réus negros".[26] (tradução nossa)

Não há como ignorar os processos discriminatórios que estão presentes na vida diária e marcam os dados que serão posteriormente utilizados para o treinamento e aprendizado de algoritmos, de traduções automáticas até predição de reincidência de crimes. A questão só se agrava quando consideramos a probabilidade dessas decisões automatizadas serem utilizadas como base para o treinamento de ainda mais algoritmos, reproduzindo e expandindo esse viés oculto.[27]

Virginia Eubanks, sobre modelos preditivos, os descreve como possuindo dano exponencial, ou seja, as suas consequências não se restringem apenas à um indivíduo ou a um pequeno grupo, mas sim se expandem para a sua inteira redes de contatos, os efeitos de decisões automatizadas, por escolha de design do próprio modelo, ou em razão de consequências sociais palpáveis, como o encarceramento, dificilmente são individuais, costumeiramente se espalhando como um vírus por sua rede.[28]

> Os impactos dos modelos preditivos são, portanto, exponenciais. Como a previsão depende de redes e abrange gerações, seu dano tem o potencial de se espalhar como um contágio, desde o ponto inicial de contato com parentes e amigos, até as redes de amigos, correndo através de comunidades inteiras como um vírus.[29] (tradução nossa)

A capacidade de perfilamento oferecida pelo tratamento em massa de dados, bem como a capacidade de utilizá-los para o desenvolvimento de novas ferramentas abre portas também

---

[26] "In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways. The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants. White defendants were mislabeled as low risk more often than black defendants." ANGWIN, e.al "Machine Bias" 23 de maio de 2016 Disponível em: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> Acessado em: 30/03/2021

[27] KEARNS, ROTH, 2020, p. n.p.

[28] EUBANKS, Virgina. "Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor". New York: St. Martin's Press. 2018

[29] The impacts of predictive models are thus exponential. Because prediction relies on networks and spans generations, its harm has the potential to spread like a contagion, from the initial point of contact to relatives and friends, to friends' networks, rushing through whole communities like a virus.

para a má-utilização dessas ferramentas, principalmente na reprodução de tendências socialmente danosas. Em caso recente nos Estados Unidos, a ausência de critérios e transparência permitiu que a companhia de vigilância Banjo recebesse mais de 223 milhões de dólares e um contrato público de 20 milhões de dólares, embora o seu CEO, Damien Patton, tenha possuído laços com a Ku Klux Klan.

Após tal revelação, em meio a preocupação de vieses existentes no algoritmo fornecido pela empresa, foram realizados processos de auditoria e não só não foram encontrados vieses, como não foi possível encontrar qualquer Inteligência Artificial.[30]

> "Em meio à crescente conscientização pública sobre o viés algorítmico, o estado de Utah suspendeu um contrato de US$ 20,7 milhões com o Banjo, e a Procuradoria Geral de Utah abriu uma investigação sobre questões de privacidade, viés algorítmico e discriminação". Mas em uma reviravolta surpresa, uma auditoria e relatório divulgados na semana passada não encontraram nenhum viés no algoritmo porque não havia nenhum algoritmo para avaliar em primeiro lugar. (...) "A Companhia Banjo declarou expressamente à Comissão que não usa técnicas que atendam à definição da indústria de Inteligência artificial. Banjo indicou que eles tinham um acordo para reunir dados do Twitter, mas não havia evidência de qualquer dado do Twitter incorporado ao Live Time", lê uma carta do Auditor Estadual de Utah, John Dougall, divulgada na semana passada."[31]

Esta comédia de erros exemplifica a total e completa ausência, mesmo em países considerados avançados tecnologicamente, de diligência em verificar a confiabilidade de algoritmos, mesmo quando estes podem trazer consequências extremamente relevantes, a curto e a longo prazo na vida de milhares de pessoas.

---

[30] JOHNSON, Khari. "Government audit of AI with ties to white supremacy finds no AI" VentureBeat, 5 de Abril de 2021, disponível em: <https://tinyurl.com/hmh7utz2> Acessado em: 11 de Abril de 2021.

[31] "Amid growing public awareness about algorithmic bias, the state of Utah halted a $20.7 million contract with Banjo, and the Utah attorney general's office opened an investigation into matters of privacy, algorithmic bias, and discrimination. But in a surprise twist, an audit and "Amid growing public awareness about algorithmic bias, the state of Utah halted a $20.7 million contract with Banjo, and the Utah attorney general's office opened an investigation into matters of privacy, algorithmic bias, and discrimination. But in a surprise twist, an audit and report released last week found no bias in the algorithm because there was no algorithm to assess in the first place. "Banjo expressly represented to the Commission that Banjo does not use techniques that meet the industry definition of artificial Intelligence. Banjo indicated they had an agreement to gather data from Twitter, but there was no evidence of any Twitter data incorporated into Live Time," reads a letter Utah State Auditor John Dougall released last week.

Em conclusão, há a reconhecida possibilidade de implementação de vícios e vieses em algoritmos, seja por meio do banco de dados utilizado para o treinamento do algoritmo ou até mesmo introduzido durante a fase de elaboração de modelos e topologia de um algoritmo, especialmente durante a otimização do algoritmo e em definir quais resultados são os desejados.

4.3 O que são decisões justas?

A definição de justiça em seu sentido mais amplo está fora do escopo do que será discutido neste capítulo. Aqui buscaremos apenas definir, sob a ótica de certos conceitos como paridade, transparência e privacidade, o que pode ser considerado como justo.

Cathy O'Neil entende que algoritmos danosos, chamados por ela de "Armas de Destruição Matemática" (ADM)[32], não vêm considerando elementos de justeza na tomada de decisões, entre as razões listadas estão a difícil categorização de justiça e imparcialidade por algoritmos, não sendo levados em conta na construção de modelos por cientistas de computação e ignorados de maneira generalizada.

> "As ADMs, pelo contrário, tendem a favorecer a eficiência. Pela sua própria natureza, elas se alimentam de dados que podem ser medidos e contados. Mas a justiça é delicada e difícil de quantificar. É um conceito. E os computadores, por todos os seus avanços em linguagem e lógica, ainda lutam poderosamente com conceitos. Eles "entendem" a beleza apenas como uma palavra associada ao Grand Canyon, ao pôr-do-sol oceânico e a dicas de beleza na revista Vogue. Eles tentam em vão medir "amizade" contando gostos e conexões no Facebook. E o conceito de justiça lhes escapa completamente. Os programadores não sabem como codificar, e poucos de seus chefes lhes pedem isso. Portanto, a justiça não é calculada em ADMs. E o resultado é uma produção industrial maciça de injustiça. Se você pensa em uma ADMs como uma fábrica, injustiça é o material negro que sai das chaminés de fumaça. É uma emissão, uma emissão tóxica. A questão é se nós, como sociedade, estamos dispostos a sacrificar um pouco de eficiência no interesse da justiça."[33] (tradução nossa)

[32] O' NEIL, 2016, "Weapons of Math Destruction" p. n.p.

[33] WMDs, by contrast, tend to favor efficiency. By their very nature, they feed on data that can be measured and counted. But fairness is squishy and hard to quantify. It is a concept. And computers, for all of their advances in language and logic, still struggle mightily with concepts. They "understand" beauty only as a word associated with the Grand Canyon, ocean sunsets, and grooming tips in Vogue magazine. They try in vain to measure "friendship" by counting likes and connections on Facebook. And the concept of fairness utterly escapes them. Programmers don't know how to code for it, and few of their bosses ask them to. So fairness isn't calculated into WMDs. And the result is massive, industrial production of unfairness. If you think of a WMD as a factory, unfairness is the black stuff

Acompanhando a preocupação de Cathy O'Neil, em "The Ethical Algorithm", Kearns e Roth dedicam considerável tempo à exploração do que significa uma decisão justa. Uma das noções mais simples sugeridas[34] é a chamada "paridade estatística": este seria o conceito mais direto para definir o que pode ser considerado como uma decisão justa. Primeiro, haveria de se definir um grupo de indivíduos cujo buscaria se proteger; por exemplo, se há preocupação que um grupo que durante a concessão de empréstimos esteja sendo discriminado contra, a paridade estatística requer apenas que a fração de membros do grupo discriminado que receberá o empréstimo seja igual à fração do grupo cujo não há preocupação contra discriminação.

Apesar da aparência de se tratar de uma decisão justa, afinal, estatisticamente ambos grupos receberam o mesmo tratamento, há falhas quando aplicados à realidade dos algoritmos, especialmente considerando as vastas quantidades de informações disponíveis sobre cada indivíduo, de modo que o algoritmo estaria cego à certas propriedades específicas dos indivíduos.

Assim, podem ser considerados outros aspectos auxiliares, levando em consideração elementos relevantes para a análise em questão, como por exemplo pagamentos passados e outros dados relevantes, mas ainda assim levando em consideração a concessão dos empréstimos. Apesar dessa possibilidade, sempre estará presente tal conflito entre a precisão do algoritmo e a justeza da decisão, em outras palavras, a otimização e eficiência de algoritmos sempre estarão em conflito com a justiça envolvida na tomada de uma decisão.

Há maneiras de remediar a falta de precisão da paridade estatística. Por exemplo, é sugerida por Kearns e Roth uma maneira sistêmica de explorar o conflito entre precisão e justiça de maneira algorítmica. Uma das maneiras consideradas seria estabelecer uma métrica comparativa entre os erros cometidos por um algoritmo e a justeza das decisões tomadas:

> "Qual é melhor - o corte "otimizado", que comete sete erros e tem uma pontuação de 4 injustiças, ou o corte "mais justo", que comete oito erros e tem uma pontuação de injustiças de 2? Não há uma resposta universalmente correta, porque cada um desses modelos é melhor em um critério e pior no outro. Eles são assim incomparáveis, e devemos considerar ambos como

belching out of the smoke stacks. It's an emission, a toxic one. The question is whether we as a society are willing to sacrifice a bit of efficiency in the interest of fairness.

[34] KEARNS, ROTH. "The Ethical Algorithm", 2020, p. n.p.

> candidatos razoáveis"[35] (KEARNS, ROTH. "The Ethical Algorithm", 2020, p. n.p., tradução nossa)

Há de se destacar que Kearns e Roth buscam soluções no contexto da legislação americana, onde em alguns Estados há receio na aplicação de medidas afirmativas, em sentido contrário à já bem consolidada doutrina de 'discriminação positiva' ou ação afirmativa aplicada no Brasil. Tal horizonte nos garante também maior liberdade para pensarmos métricas diferentes para grupos de usuários de certas aplicações.

Como sociedade é necessário por vezes descartar eficiência, como na redução de erros, por justiça, ignorar a necessidade de se "prejudicar" o algoritmo para garantir que as suas decisões estão sendo tomadas de maneira justa, é relevante para impedir as piores consequências de vieses.[36]

No caso inicialmente tratado, sobre a concessão de empréstimos, enquanto é possível elaborar um modelo que considere os aspectos individuais de cada cliente, é relevante considerar uma métrica onde ainda exista proporcionalidade referente aqueles que tiveram seus pedidos aceitos e até mesmo entre aqueles que teriam recebido falsos-negativos, apesar de ser impreciso.[37]

Mais especificamente, é necessário estabelecer modelos próprios para cada propósito. Como bem explorado em "The Ethical Algorithm", não há como tratar da mesma maneira algoritmos de tradução automática, enquanto e algoritmos que tratam de empréstimos, criminologia ou moradia dos indivíduos, durante o desenvolvimento deve-se considerar o modelo compatível.

5. Definições de Discriminação

A tarefa de delimitar a definição de discriminação a ser aplicada na avaliação e auditoria de algoritmos é importante não apenas para, como delineado anteriormente, conseguir

[35] "Which is better—the "optimal" cutoff, which makes seven mistakes and has an unfairness score of 4, or the "more fair" cutoff, which makes eight mistakes and has an unfairness score of 2? There is no universally right answer, because each of these models is better on one criterion and worse on the other. They are thus incomparable, and we should consider both to be reasonable candidates"
[36] O' NEIL, 2016, p. n.p.
[37] KEARNS, ROTH, 2020, p. n.p.

diferenciar entre discriminação em sentido amplo e abstrato, como função mecânica e a sua definição aplicada popularmente.

Além de reforçar a diferença entre essas duas definições, iremos explorar mais profundamente as nuances entre discriminação direta e indireta e, principalmente, entre discriminação positiva e negativa. Para tanto, foi utilizada a recente obra de Adilson José Moreira, Tratado De Direito Antidiscriminatório, que sistematicamente explora cada um dos pontos a serem abordados.

De início, a maneira como é abordada a pluralidade de significados associados com a própria palavra discriminação não é apenas elucidativa, como também introduz um ângulo de análise jurídico, que será importante mais a diante, ao tratar-se de responsabilização e consequências:

> "A palavra discriminação possui uma pluralidade de significados, embora tenha adquirido um sentido bem específico no mundo atual. Ela designa, por um lado, a ação de classificar objetos a partir de um determinado critério. Essa acepção genérica passou a segundo plano por causa da preponderância de sua dimensão moral e jurídica nos dias atuais. Hoje o termo discriminar tem conotações claramente negativas, pois sugere que alguém foi tratado de forma arbitrária. Os dois sentidos dessa palavra estão presentes no vocabulário jurídico. Sabemos que instituições estatais classificam indivíduos a partir de critérios necessário para o alcance de algum interesse público. O vocábulo discriminar significa aqui caracterizar pessoas ou situações a partir de uma característica para atribuir a ela alguma consequência. Contudo, a palavra discriminação tem também outro significado no mundo jurídico: ela indica que uma pessoa impõe à outra um tratamento arbitrário a partir de um julgamento moral negativo, o que pode contribuir para que a segunda esteja em uma situação de desvantagem."[38]

A definição jurídica de discriminação apresentada por Adilson Moreira, como a imposição de tratamento arbitrário a partir de um julgamento moral negativo, se alinha com os exemplos anteriormente apresentados de discriminação e viés algoritmo em concreto. Sendo que, em situações de discriminação, a partir de processos específicos e sua própria lógica interna

[38] MOREIRA, Adilson José. "**Tratado de Direito Antidiscriminatório**", São Paulo, Editora Contracorrente, 2020

estes algoritmos chegaram à certa conclusão, desejável ou não, e de maneira injusta impuseram um tratamento arbitrário, maiores taxas de reincidência, *red-flags* e outras consequências.

Preliminarmente, devemos tratar dos conceitos de discriminação negativa e discriminação positiva. A primeira dessas representa exatamente a arbitrariedade e injustiça a qual se refere acima:

> "A discriminação negativa designa um tratamento que viola o princípio segundo o qual todos os membros de uma comunidade política devem ser igualmente respeitados. Ela acontece quando um agente público ou privado trata uma pessoa ou grupo de pessoas de forma arbitrária, o que é frequentemente motivado por estigmas culturais."[39]

Enquanto a segunda categoria, de discriminação positiva, lida exatamente com medidas que buscam igualar ou amenizar situações de desvantagem histórica, tal possibilidade é legalmente reconhecida no Brasil, sendo possível de ser utilizada em uma diversidade de contextos:

> "A discriminação positiva pode ser distinguida da discriminação negativa porque cria uma distinção temporária ou permanente para membros de um determinado grupo que possuem uma história de desvantagem ou que estão em uma situação de vulnerabilidade."[40]

Quanto ao conceito da discriminação negativa, pode-se questionar se não se veria restringida a aplicabilidade de tal conceito, isso é, a discriminação em seu nível jurídico ou moral em razão da ausência de qualquer motivação discernível, afinal, é difícil acreditar que há uma quantidade expressiva de companhias, ou mesmo programadores que buscam discriminar contra um grupo específico de maneira consciente.

Diante de tal questionamento, há de se esclarecer que processos discriminatórios não requerem vontade, intenção ou consciência daquele que o realiza, ou no caso, implementa por meio da tecnologia. Aqui se delineiam as primeiras características da discriminação direta e indireta:

[39] MOREIRA, 2020, p. n. p.
[40] Idem. p. n.p

> "A discriminação direta envolve a intencionalidade: o agente discrimina outro de forma consciente porque está motivado por interesses que não podem ser justificados por estarem baseados em estereótipos ou preconceitos ou porque está motivado por algum interesse estratégico. A discriminação sofrida causa danos à pessoa, danos que estão relacionados com os critérios a partir dos quais ela foi discriminada. Vemos então que a discriminação direta constitui uma violação do preceito da justiça simétrica presente no texto constitucional. Esse mandamento requer o tratamento igual entre os que estão igualmente situados, um elemento básico da moralidade do regime democrático. A discriminação direta pode ser vista como uma manifestação e uma forma mais genérica de discriminação, que é a discriminação negativa, porque produz danos às pessoas discriminadas."[41]
>
> (...)
>
> "Em resumo, a discriminação direta está baseada nos seguintes elementos: a arbitrariedade, a intencionalidade, um tratamento desvantajoso e a utilização de um critério proibido por lei."[42]

Enquanto a discriminação direta se baseia na intencionalidade daqueles que perpetuam a discriminação, a discriminação indireta se sustenta na aparente neutralidade da tecnologia, na suposta imparcialidade do algoritmo e das redes neurais. Essas estruturas podem, em meio à ofuscação, de maneira implícita discriminar contra grupos estigmatizados e indivíduos, por vezes de maneira mais ampla e insidiosa do que a simples discriminação direta. A discriminação indireta é definida da seguinte maneira:

> "Além da ausência de intencionalidade aberta a discriminação indireta também requer a existência de um impacto desproporcional sobre um grupo, elemento que viola o interesse estatal da eliminação de hierarquias sociais. Uma sociedade democrática requer que práticas sociais não contribuam para a deterioração das condições de vida das pessoas e, por isso, ações que as impactam desproporcionalmente devem ser eliminadas."[43]

Como já indicado anteriormente, o risco presente na utilização de algoritmos que possuem, de maneira inerente, mas oculta, vieses contra um grupo específico, pode ocasionar na introdução de desvantagens consideráveis. Nesse sentido, até que se fossem identificado vieses em um algoritmo ele continuaria operando de maneira discriminatória, impactando na

[41] Idem. p. n.p
[42] Idem p. n.p.
[43] Idem p. n.p

deterioração não apenas imediata das condições de vidas dos membros desse grupo, como também perpetuando sistemas que de longa data antecipam tais ferramentas.

Portanto, até o momento foi possível delinear tanto a existência da prática de discriminação direta e indireta, utilizando o critério da intencionalidade como parâmetro para a sua classificação, mais adiante, ao tratar de responsabilidade, discutiremos as complexidades oferecidas por esta categorização.

Quanto ao aspecto jurídico e jurisprudencial, há de se destacar que a própria Constituição Federal Brasileira, estabelecendo como um dos seus princípios fundamentais a 'promoção do bem de todos, sem preconceitos de origem, raça, sexo, cor, idade e quaisquer outras formas de discriminação'[44], prevendo também a punição de qualquer discriminação que atente contra os direitos e liberdades fundamentais.[45]

A determinação jurídica da ocorrência de discriminação acompanha os parâmetros e critérios já apresentados, requerendo a violação do princípio de igualdade:

> "A jurisprudência dos tribunais constitucionais indica outro elemento muito importante desse conceito: ela deve ser determinada a partir da comparação de um grupo em relação a outro. Por significar uma violação do princípio da igualdade, precisa ser estabelecida a partir de um critério; assim um grupo estará em uma situação de desvantagem em relação a outro em função de um ou mais critérios relevantes. Ela se torna notória quando grupos que vivem em uma sociedade governada pelos mesmos princípios se encontram uma situação e desigualdade durável ou permanente."[46]

Tal perspectiva confirma a possibilidade jurídica da instituição de leis como o Estatuto de Igualdade Racial de 2003, que não apenas busca garantir acesso comum e igualitário a serviços públicos, como visa também a promoção da inclusão de grupos discriminados no

[44] Art. 3º Constituem objetivos fundamentais da República Federativa do Brasil: IV - promover o bem de todos, sem preconceitos de origem, raça, sexo, cor, idade e quaisquer outras formas de discriminação.

[45] Art. 5º Todos são iguais perante a lei, sem distinção de qualquer natureza, garantindo-se aos brasileiros e aos estrangeiros residentes no País a inviolabilidade do direito à vida, à liberdade, à igualdade, à segurança e à propriedade, nos termos seguintes: XLI - a lei punirá qualquer discriminação atentatória dos direitos e liberdades fundamentais;

[46] MOREIRA, Adilson José. "**Tratado de Direito Antidiscriminatório**", 2020, p. n.p.

mercado de trabalho, garantindo proteção contra tratamento discriminatório inclusive contra entidades particulares.

Assim, há clara possibilidade de discutir os efeitos implícitos do viés discriminatório de algoritmos tanto com base em nossa própria constituição como em relação à jurisprudência recente, que enxerga a proteção de direitos e liberdades fundamentais não apenas em nível individual e direto como também de maneira indireta, atingindo direitos difusos da coletividade.

Como já magistralmente notado, como não poderia deixar de ser, Adilson Moreira realiza tal conexão ao trabalhar a questão da discriminação e tecnologia, indicando não apenas a possibilidade de discriminação direta e indireta, bem como a sua associação com a discriminação estrutural, por meio da qual a Inteligência Artificial "promove a opressão racial porque atua sobre uma realidade estruturada a partir de sistemas de dominação, uma combinação responsável pela manutenção de disparidades entre grupos sociais."[47]

Embora o reconhecimento das práticas discriminatórias a partir de categorias como discriminação direta, indireta, institucional e estrutural nos permita melhor análise imediata da presença de decisões discriminatórias em algoritmos, se faz imperativo expandir essa perspectiva para uma interpretação das estruturas discriminatórias que se apresentam como pilares de nossa civilização.

Como já exposto anteriormente, a tecnologia não é aplicável de maneira neutra, a história e estruturas políticas de nossa sociedade condicionam a sua aplicação. É necessário aqui descartar qualquer ilusão de que se trata de comportamentos isolados de indivíduos ou erros ocasionais que levam à discriminação. A própria história recente de nossa civilização explica os comportamentos aqui reproduzidos: a racialização, opressão, marginalização e estigmatização da população negra, mulheres e da comunidade LGBT se deu não por erro ou por comportamentos isolados, mas de modo institucional e até verdadeiramente estrutural.

Sobre a questão, James Baldwin[48], se utilizando de Dostoevsky, confronta o pretenso humanismo, a aparência de igualdade e liberdade individual proposta pelo ocidente, com a

[47] MOREIRA, Adilson José. "**Tratado de Direito Antidiscriminatório**", 2020, p. n.p.
[48] I AM NOT YOUR NEGRO; Direção Raul Peck. Produção: Velvet Film, Estados Unidos: Magnolia Pictures, 2016. (95 min.)

própria ausência de justificação moral para a sua história de dominação, fazendo referência a um trem que traz comida apenas para metade da humanidade:

> Todas as nações ocidentais foram pegas em uma mentira, a mentira de seu pretenso humanismo. Isto significa que sua história não tem justificativa moral, e que o Ocidente não tem autoridade moral. "Vil como eu sou", diz um dos personagens de O Idiota de Dostoievski, "Eu não confio nos vagões que trazem pão para a humanidade". Pois os vagões que trazem pão para a humanidade, podem friamente excluir uma parte considerável da humanidade de desfrutar do que é trazido.[49]

O presente capítulo, embora reconhecidamente superficial quanto à questão da racialização e as dinâmicas da discriminação estrutural enfrentadas por grupos estigmatizados, espera ter evidenciado a possibilidade de, além do comportamento individual, direto, ser possível identificar a discriminação, com efeitos tão ou até mais danosos e perniciosos quanto uma ofensa direta.

6. Recursos disponíveis

Diante do já explorado entende-se ser adequado iniciar uma exploração direta sobre as possibilidades, ferramentas e recursos disponíveis para não apenas evitar, como também diretamente identificar e combater algoritmos que venham a reproduzir vieses, neste capitulo abordaremos conceitos como o da auditoria de algoritmos, a responsabilização por decisões discriminatórias e a revisão de decisões automatizadas, além de também propor uma discussão interdisciplinar entre as áreas técnicas do direito, ciência da computação e a ética e filosofia.

6.1 Identificação de Decisões enviesadas/discriminatórias.

A chamada auditoria de algoritmos possuí origem correlacionada com a prática de auditorias em outras áreas, como por exemplo a financeira, tendo surgido inicialmente durante o período da revolução industrial e expansão de firmas, servindo a função de garantir segurança

[49] All of the Western nations have been caught in a lie, the lie of their pretended humanism. This means that their history has no moral justification, and that the West has no moral authority. "Vile as I am," states one of the characters in Dostoevsky's The Idiot, "I don't believe in the wagons that bring bread to humanity. For the wagons that bring bread to humanity, may coldly exclude a considerable part of humanity from enjoying what is brought.

contra riscos financeiros inesperados, a expansão de formas e áreas abordadas por auditoria, como auditorias de *compliance*, ambiental e de práticas de trabalho são crescentemente comuns.

Do mesmo princípio, isso é, a garantia da legitimidade das operações financeiras, operacionais ou de *compliance*, a auditoria de algoritmos busca garantir a regularidade de algoritmos, Inteligências Artificiais e Redes Neurais. Essas ferramentas requerem maior cuidado e detalhe na sua análise, envolvendo não apenas uma análise dos resultados, mas também uma avaliação de impactos, verificação de banco de dados e até mesmo da topologia de redes neurais, quando estas estão disponíveis aos auditores.

Além de se tratar de um campo recente no contexto de auditorias, há também questões quanto à confidencialidade dos algoritmos e redes neurais utilizados por tais decisões, terceiros interessados em realizar uma auditoria independente do algoritmo de uma companhia não podem, apesar de possíveis impactos sociais, obter acesso irrestrito a esses algoritmos, requerendo maior criatividade e inovação para poder seguir adiante com a auditoria.

O acesso muitas vezes é restringido em razão de segurança das operações realizadas por esses algoritmos, sendo muitas vezes fundamentais para a operação financeira de instituições sendo enxergada qualquer abertura para auditorias independentes como a possibilidade de explicitar fragilidades para futura exploração por agentes maliciosos.

Há de se considerar também o regime de proteção de Propriedade Intelectual dessas ferramentas, diversas companhias optam pela não utilização de sistemas de copyright e de patentes, mas sim a da utilização do segredo evitando o registro público dessas invenções, o que dificulta consideravelmente os processos de avaliação desses algoritmos, os benefícios associados com a adoção de segredos de negócio, invés do registro público de Redes Neurais e modelos algoritmos, estão não apenas na maior flexibilidade do segredo industrial, como também maior facilidade de *enforcement*, como explicado por Jasper Siems.[50]

> Ao contrário de uma patente cujo período de concessão pode levar vários anos, a proteção do segredo comercial ocorre por padrão. Os pedidos de patente podem tornar-se rapidamente caros, especialmente para pequenas e médias

[50] SIEMS, Jasper. "Protecting Deep Learning: Could the New EU-Trade Secrets Directive Be an Option for the Legal Protection of Artificial Neural Networks?" Editorial: "Algorithmic Governance and Governance of Algorithms", 2021, Springer, Umea Sweden. Pag. 153

> empresas. A proteção do segredo comercial se aplica sem custos substanciais. Assim, ela requer apenas medidas de sigilo adequadas, algumas das quais já existem na maioria das empresas. Na lei de segredo comercial, às vezes é mais fácil detectar violações em comparação com a lei de direitos autorais. Particularmente, o segredo comercial também é protegido se o produto final não for o mesmo, havendo proteção em todos os casos, contanto que o segredo comercial tenha sido usado em sua criação. [51] (tradução nossa)

Por óbvio, quando busca-se garantir que não há impactos sociais negativos na operação de algoritmos são necessárias auditorias independentes e o cumprimento de padrões gerais de indústria para a avaliação de princípios éticos em algoritmos, a discussão sobre a possibilidade dessa forma de auditoria ainda encontra resistência, tanto no campo do acesso, como mencionando anteriormente, mas também em relação às definições claras quanto aos princípios éticos que devem ser aplicados, Ruha Benjamin em Race After Technology[52] descreve a importância dos processos de auditoria da seguinte maneira:

> É fundamental que tais auditorias sejam independentes e exequíveis. Atualmente, não há sequer padrões de impacto social em toda a indústria que contabilizem totalmente a forma como os algoritmos são usados para "alocar moradia, saúde, contratação, bancos, serviços sociais, bem como a entrega de bens e serviços". Os princípios éticos de AI do Google, criados após a controvérsia sobre o contrato da empresa no Pentágono, são um bom começo, mas se concentram de forma muito restrita em tecnologias militares e de vigilância e, ao confiar em "princípios amplamente aceitos do direito internacional e dos direitos humanos", eles contornam a prática comum dos governos que fiscalizam seus próprios cidadãos. Esses princípios também não garantem uma revisão independente e transparente; eles seguem, ao invés disso, um padrão atual na governança corporativa que mantém "processos internos e secretos" que impedem a responsabilidade pública.[53] (tradução nossa)

---

[51] Unlike a patent whose granting period may take several years, trade secret protection occurs by default. Patent applications can quickly become expensive, especially for small and medium-sized enterprises. Trade secret protection applies without any substantial costs. Thus, it only requires appropriate secrecy measures, some of which already exist in most company. In trade secret law, it is sometimes easier to detect infringements compared to copyright law. Particularly, the trade secret is also protected if the end product is not the same, but in all cases, if the trade secret has been used in its creation

[52] BENJAMIN, Ruha. **"Race After Technology",** Medford, Polity Press, United Kingdom. p. n.p

[53] Crucially, such audits need to be independent and enforceable. Currently there are not even any industry-wide standards for social impact that fully account for the way in which algorithms are used to "allocate housing, healthcare, hiring, banking, social services as well as goods and service delivery." Google's AI ethics principles, created in the aftermath of the controversy over the company's Pentagon

Há na área, portanto, a necessidade de se estabelecer critérios generalizados além de construir aceitação quanto a realização de auditorias independentes de auditorias de algoritmos. No entanto, não se deve enxergar a auditoria de algoritmos com ceticismo. Anteriormente mencionamos o conhecido caso da avaliação de reincidência criminal noticiado pela agência Propublica - esse e outros exemplos só foram possíveis de serem alcançados pela colaboração e trabalho empenhado de grupos independentes realizando extensa análise das *caixas pretas* que hoje definem a vida de milhões de pessoas diariamente.

Avanços vêm sendo feitos também na implementação de auditorias públicas de algoritmos, merece especial destaque os esforços sendo desprendidos pelas SAIs ("Instituições Supremas de Auditoria") Europeias, que ainda em novembro de 2020 divulgaram o portal 'Auditing Algorithms' <https://www.auditingalgorithms.net/> disponibilizando um esquema unificado de avaliação técnica de algoritmos por auditores governamentais:

> "Este documento descreve as auditorias potenciais dos sistemas de IA pelas Instituições Superiores de Auditoria (SAIs), cobrindo riscos relacionados ao uso de modelos ML em agências governamentais, bem como possíveis testes para obter provas de auditoria. Inclui ainda uma lista de verificação de capacidade de auditoria que resume os pré-requisitos mínimos que uma organização auditada deve reter da fase de implementação para permitir qualquer auditoria subsequente."[54] (tradução nossa)

Tamanha é a relevância da auditoria de algoritmos e a preocupação com a discriminação que recentemente o CNJ editou normas, por meio da Resolução Nº 332 de 21/08/2020[55] prevendo a realização de verificação de vieses discriminatórios, condicionando a adoção de algoritmos à possibilidade de eliminação de vieses no modelo:

---

contract, are a good start but focus too narrowly on military and surveillance technologies and, by relying on "widely accepted principles of international law and human rights," they sidestep the common practice of governments surveilling their own citizens. Nor do these principles ensure independent and transparent review; they follow instead a pattern current in corporate governance that maintains "internal, secret processes" that preclude public accountability.

[54] "This paper outlines potential audits of AI systems by Supreme Audit Institutions (SAIs), covering risks related to the use of ML models in government agencies as well as possible tests to gain audit evidence. It further includes an auditablity checklist which summarises the minimum prerequisites an auditee organisation should retain from the ML implementation phase to enable any subsequent audit."

[55] PRESIDENTE DO CONSELHO NACIONAL DE JUSTIÇA, Resolução Nº 332 de 21/08/2020, CNJ, Disponível em: https://atos.cnj.jus.br/files/original191707202008255f4563b35f8e8.pdf Acessado em: 10/05/2021.

> Art. 7o As decisões judiciais apoiadas em ferramentas de Inteligência Artificial devem preservar a igualdade, a não discriminação, a pluralidade e a solidariedade, auxiliando no julgamento justo, com criação de condições que visem eliminar ou minimizar a opressão, a marginalização do ser humano e os erros de julgamento decorrentes de preconceitos.
>
> § 1o Antes de ser colocado em produção, o modelo de Inteligência Artificial deverá ser homologado de forma a identificar se preconceitos ou generalizações influenciaram seu desenvolvimento, acarretando tendências discriminatórias no seu funcionamento.
>
> § 2o Verificado viés discriminatório de qualquer natureza ou incompatibilidade do modelo de Inteligência Artificial com os princípios previstos nesta Resolução, deverão ser adotadas medidas corretivas.
>
> § 3o A impossibilidade de eliminação do viés discriminatório do modelo de Inteligência Artificial implicará na descontinuidade de sua utilização, com o consequente registro de seu projeto e as razões que levaram a tal decisão.[56]

Na mesma resolução, o CNJ busca estreitar uma definição de transparência em relação à utilização de dados, incluindo a necessidade de capacidade de fornecimento de explicação satisfatória e passível de auditoria, em conformidade com o chamado direito de explicabilidade:

> Art. 8o Para os efeitos da presente Resolução, transparência consiste em:
>
> I – divulgação responsável, considerando a sensibilidade própria dos dados judiciais;
>
> II – indicação dos objetivos e resultados pretendidos pelo uso do modelo de Inteligência Artificial;
>
> III – documentação dos riscos identificados e indicação dos instrumentos de segurança da informação e controle para seu enfrentamento;
>
> IV – possibilidade de identificação do motivo em caso de dano causado pela ferramenta de Inteligência Artificial;
>
> V – apresentação dos mecanismos de auditoria e certificação de boas práticas; VI – fornecimento de explicação satisfatória e passível de auditoria por autoridade humana quanto a qualquer proposta de decisão apresentada pelo modelo de Inteligência Artificial, especialmente quando essa for de natureza judicial[57]

[56] Idem. P. 5
[57] Idem.

Por fim, é importante destacar que mesmo no desenvolvimento dos modelos houve foco sobre a importância do desenvolvimento de práticas para o desenvolvimento de equidade das decisões:

> Art. 21. A realização de estudos, pesquisas, ensino e treinamentos de Inteligência Artificial deve ser livre de preconceitos, sendo vedado: I – desrespeitar a dignidade e a liberdade de pessoas ou grupos envolvidos em seus trabalhos; II – promover atividades que envolvam qualquer espécie de risco ou prejuízo aos seres humanos e à equidade das decisões; III – subordinar investigações a sectarismo capaz de direcionar o curso da pesquisa ou seus resultados[58]

Apesar de positivo avanço no sentido de promover tais condições para a adoção de algoritmos sensíveis, ainda é necessário reconhecer que, até o momento, a realização de auditorias independentes não é generalizada, devendo continuar a ser promovida em razão de sua destacada importância e de fato implementada.

Há também de pontuar que certamente se verão discussões nos próximos anos quanto a realização de auditorias públicas de algoritmos frente à crescente adoção de modelos computacionais tanto por entidades particulares como órgãos governamentais e fundações.

6.2 Revisão de decisões discriminatórias

Uma das questões principais que definem o tema em discussão é a possibilidade não apenas de identificação de decisões enviesadas, mas também de sua reforma e da revisão de suas decisões, especialmente quando tomadas de maneira automatizada. A LGPD em seu Art. 20 garante ao titular dos dados "o direito de 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.".

Há de se destacar que tal previsão originalmente previa que a revisão de tal decisão automatizada fosse realizada por pessoa natural. Ouvidos sobre a previsão os Ministérios da Economia, da Ciência, Tecnologia, Inovações e Comunicações, a Controladoria-Geral da União e o Banco Central do Brasil se manifestaram pelo veto da provisão, por entenderem ser

[58] Idem. P. 9

excessivamente limitadora aos "modelos atuais de planos de negócios de muitas empresas", agindo contrário ao interesse público, por também impactar "na análise de risco de crédito e de novos modelos de negócios de instituições financeiras", o que acabaria por gerar um "efeito negativo na oferta de crédito aos consumidores".

Essa alterações assumidas pela Medida Provisória 869, aprovada em 2018, apontam para uma maior tendência legislativa que busca submeter qualquer tentativa legislativa de transparência à suposta viabilidade econômica e ao interesse público. De fato, ao contrário das razões descritas no veto, esse processo de revisão não apenas não afetaria desproporcionalmente modelos empresariais, como também auxiliaria o Brasil à se adequar às normas e padrões internacionais de transparência em tomadas de decisões automatizadas:

> A garantia de revisão por pessoa natural de decisão automatizada atenderia ao princípio da transparência e colocaria a Legislação Brasileira no mesmo patamar de outras legislações internacionais que garantem a efetividade deste direito através da intervenção humana. O objetivo de garantir o direito de revisão de decisão automatizada por pessoa natural gira em torno de corrigir eventuais discriminações decorrentes de processos algorítmicos, a fim de conferir maior transparência e responsabilidade nos processos de perfilamento dos cidadãos.[59]

Embora à primeira vista a racionalização econômica por trás da Medida que sugeriu o veto do requerimento de pessoa natural pareça demonstrar preocupação com a viabilidade do setor, acaba por prejudicá-lo, permitindo que práticas como a revisão automatizada de decisões automatizadas sejam viáveis, o que a longo termo pode ser prejudicial para a própria justeza de decisões, precisão dos algoritmos e a confiança coletiva que temos dessas ferramentas.

De maneira esclarecedora em obra recente, Caitlin Mulholland e Isabella Z. Frajhof[60] expõem elementos de destaque a serem observados na discussão e aplicação da previsão de

[59] SILVA, Priscilla. MEDEIROS, Juliana, **"A polêmica da revisão (humana) sobre decisões automatizadas"** Disponível em: https://feed.itsrio.org/a-pol%C3%AAmica-da-revis%C3%A3o-humana-sobre-decis%C3%B5es-automatizadas-a81592886345 Acessado em: 10/05/2021

[60] MULHOLLAND, Caitlin. FRAJHOF Isabella Z. **"Inteligência Artifical e a Lei Geral de Proteção de Dados Pessoais: breves anotações sobre o direito à explicação perante a tomada de decisões por meio de machine learning."** Em "Inteligência Artifical e direito: ética, regulação e responsabilidade" Coord: Ana Frazão e Caitlin Mulholland – 2. Ed. rev. Atual. E ampl. – São Paulo, Thomsom Reuters Brasil, 2020.

revisão de decisões automatizadas, destacando inicialmente a diferenciação que há de ser feita entre a possibilidade de se revisar uma decisão automatizada e a sua reforma:

> "Merecem ser feitas duas notas importantes sobre este artigo. A primeira refere-se ao fato de que a lei autoriza o pedido de revisão, mas isto não significa que, após a análise pelo controlador, o resultado final necessariamente será alterado. A segunda reconhece, à primeira vista, a discricionariedade da autoridade nacional para realizar a auditoria apenas quando o controlar se negar a fornecer as informações elencadas no parágrafo primeiro."[61]

Ana Frazão escreve, ainda antes da Medida Provisória que removeu a previsão de revisão de decisão por pessoa natural, os elementos essenciais do direito garantido pelo Artigo 20 da LGPD, que vão além também da simples possibilidade de requisição de explicação ou revisão, abordando também a possibilidade de se peticionar a autoridade nacional para a realização de auditoria:

> "(i) o direito de acesso e informação em relação a respeito dos critérios e procedimentos utilizados para a decisão automatizada, (ii) o direito de oposição quanto à decisão automatizada e de manifestar o seu ponto de vista, (iii) o direito de obtenção da revisão da decisão automatizada por uma pessoa natural e (iv) o direito de petição à autoridade nacional para realização de auditoria, em caso de não prestação de informações. [62]

No entanto, para que possa ser levada adiante a implementação do Artigo 20, é necessário definir o que é uma decisão totalmente automatizada e, por consequência, quando o titular do direito à explicação poderá fazer jus à previsão da LGPD. Segundo Caitlin Mulholland, conforme a LGPD, isto uma decisão totalmente automatizada se dá quando "a decisão automatizada é tomada sem qualquer interferência humana que seja capaz de alterar seu resultado final, bastando simplesmente que negue a eficácia ou deixe de promover seus direitos.".[63]

Assim, podemos concluir que o Artigo 20 da LGPD permite a revisão de decisões automatizadas tomadas por modelos algorítmicos quando não há presença de interferência

[61] Idem. p. 274
[62] Idem. p. 275
[63] Idem. p. 277

humana que, não apenas possa verificar o seu resultado, mas que tenha capacidade de alterar a sua conclusão e quando a decisão afetar a eficácia ou promoção de direitos do titular de dados, consonância com legislação internacional, como a GDPR, que prevê em seu artigo 22:

> ARTIGO 22.O
>
> Decisões individuais automatizadas, incluindo definição de perfis
>
> O titular dos dados tem o direito de não ficar sujeito a nenhuma decisão tomada exclusivamente com base no tratamento automatizado, incluindo a definição de perfis, que produza efeitos na sua esfera jurídica ou que o afete significativamente de forma similar.[64]

Por fim, quanto à chamada explicabilidade, Ana Frazão sintetiza os aspectos que deverão ser esclarecidos pelo controlador do algoritmo da seguinte maneira:

> "(i) os dados que são coletados, de que fonte e de que maneira, (ii) quais as linhas gerias de programação dos algoritmos e seus objetivos, (iii) como se deu a programação e o desenvolvimento do algoritmo, (iv) se o algoritmo pode ou não modificar seu próprio código, (v) se tais modificações são previsíveis ou ao menos verificáveis, (vi) quais as categorias relevantes dos personagens e os critérios para cada uma delas, (vii) quais são os outputs do processo decisório e como avaliar a sua adequação e acurácia, (viii) se há mecanismos de feedback, (viii) se há intervenção humana e em que nível, (ix) quais são os principais impactos e riscos para os titulares de danos, (x) que medidas forma tomadas para conter tais riscos" (Frazão, 2018a). [65]

Apesar das alterações por motivos de 'interesse público', a possibilidade de revisão de decisões automatizadas, mesmo em sua redação atual, é um importante reconhecimento e garantia oferecida pela LGPD que permite que indivíduos requeiram informações, explicação e revisão de decisões automatizadas que lhe tenha afetado. Há ainda de se verificar a implementação prática da provisão, bem como a maneira que o direito à explicabilidade se efetivara diante da prática comum de segredo industrial e comercial dos controladores de algoritmos de decisões automatizadas.

Apesar da dificuldade de prever a eficiência e aplicação do Artigo 20, há consequências lógicas que decorrem da possibilidade de identificação de decisões automatizadas que podem

[64] EUROPEAN PARLIAMENT AND OF THE COUNCIL **"Regulamento Geral sobre a Proteção de Dados"**, 2016, Disponível em: <https://gdprinfo.eu/pt-pt/pt-pt-article-22> Acessado em: 10/05/2021.
[65] MULHOLLAND, FRAJHOF, 2020, p. 278.

ter afetado negativamente, inclusive de maneira enviesada, titulares de direitos, isto é, há a possibilidade de responsabilização e indenização por danos causados por decisões automatizadas.

6.3 Responsabilização

Identificada uma decisão enviesada, ou decisão que violou direito de titular de maneira genérica, como poderá ser identificado o responsável pela decisão, se ela se deu de maneira totalmente automatizada e como poderá se proceder com o processo de responsabilização. Além da já mencionada eliminação total de revisão ou possibilidade de reforma de decisões automatizadas para a classificação de uma decisão como totalmente automatizada, Caitlin Mulholland[66] também oferece destaque à imprevisibilidade dos efeitos obtidos, essa categoria se mostra fundamental para considerações quanto a possibilidade de se antecipar resultados, relevante para aferir responsabilidade:

> "duas são as características essenciais da automação total de mecanismos decisionais em IA: (i) a absoluta independência da interferência humana para alcançar resultados e, como consequência (ii) a imprevisibilidade dos efeitos obtidos. Considerando que a IA permite associar a tecnologia de aprendizado por máquinas à criatividade – no sentido estrito de criação -, o resultado dessa capacidade de decidir autonomamente, sem mediação humana no seu desenvolvimento, impossibilita a previsibilidade antecipada dos resultados, seja porque desconhecidos seja porque difíceis de explicar. De acordo com Vladeck, nas tomadas de decisão totalmente autônomas por IA, a tecnologia abandona a tradicional função de ferramenta a serviço dos humanos para transforma-se em um sistema que atuará independente de instruções diretas humanas, sendo baseado na informação que ele próprio adquire e analisa. Esse sistema por sua veze tomara decisões altamente significantes em circunstâncias que podem ou não ser antecipadas ou diretamente associadas aos seus desenvolvedores."[67]

A representação de modelos algorítmicos como uma caixa-preta que toma decisões sem qualquer interferência, com base em informações assimiladas autonomamente parece ser um desafio intransponível para a identificação de responsabilidade, especialmente por não haver

[66] MULHOLLAND **"Responsabilidade civil e processos decisórios autônomos em sistemas de Inteligência Artificial (IA): autonomia, imputabilidade e responsabilidade."** Em "Inteligência Artifical e direito: ética, regulação e responsabilidade" Coord: Ana Frazão e Caitlin Mulholland – 2. Ed. rev. Atual. E ampl. – São Paulo, Thomsom Reuters Brasil, 2020.

[67] Idem. P. 332 – 333

categoria jurídica para esses modelos. Não há, se gostaríamos de responsabilizar modelos, a possibilidade de identificá-lo como pessoa dotada de personalidade, mesmo que jurídica:

> "Como consequência dessa expansão de estruturas tecnológicas baseadas na autonomia da IA – verdadeiras caixas-pretas -, a potencialidade e probabilidade danosas serão incrementadas, em decorrência da imprevisibilidade dos resultados alcançados pela IA e da inimputabilidade da tecnologia, duas características essenciais que poderiam, em tese, afastar a obrigação de indenizar." [68]

Essa certamente será a posição adotada por alguns daqueles que concordam com a oposição à revisão por pessoa natural, havendo a possibilidade de responsabilização por decisões tomadas de maneira totalmente automatizada, tal possibilidade resultaria em uma série de limitações e riscos às empresas que estão no limiar do desenvolvimento desta tecnologia. A impossibilidade de se responsabilizar individualmente a IA significaria a irresponsabilidade por qualquer dano sofrido. No entanto, apesar de não agradar a todos, não há espaço para tamanha lacuna no direito brasileiro.

Novamente, Caitlin Mulholland magistralmente explica que a virada conceitual da doutrina moderna de responsabilidade civil oferece perspectiva importante para a solução do presente problema, a inversão assumida na priorização da reparação do dano sofrido pela vítima, no lugar de manter seu foco na obrigação de indenização nos oferece uma abordagem pela qual "a necessidade amparada socialmente – e constitucionalmente – de reparar os danos injustamente sofridos, sejam eles resultados de um agir culposo, sejam consequência de uma atividade lícita qualquer" assume posição central no raciocínio legal.

Há então de se avaliar as alternativas existentes para o reconhecimento de responsabilidade. A primeira alternativa à completa irresponsabilidade seria trabalhar sobre a premissa de responsabilização civil da própria IA, estabelecendo para tanto uma personalidade jurídica para o algoritmo. Embora essa possa parecer uma alternativa moderna, há de se enxergar que há entraves consideráveis, tanto em nível administrativo para o registro dessas novas personalidades, o que levaria também à questionamentos, quanto em nível prático, referente ao processo de responsabilização. Apesar de poder oferecer um interessante debate, a

[68] Idem. P. 333 - 334

presente alternativa não se mostra viável, resultando na prática em efeitos semelhantes à alternativa de irresponsabilidade sobre a ação da IA.

Podemos então optar por responsabilizar aqueles responsáveis pela programação desses modelos, afinal, foram eles que elaboraram o código, o modelo, o peso e até mesmo o banco de dados utilizados para o treinamento de eventual rede neural, haveria alguém tão intimamente associado com o desenvolvimento quanto o programador? Bem, aqui devemos retornar ao aspecto da imprevisibilidade dos resultados de tomada de decisões da IA, como provar a responsabilidade de um programador em específico pelos danos resultantes, além do que, seria claramente desproporcional a responsabilização de programadores, muitas vezes contratados ao longo do desenvolvimento, sem controle não apenas sobre o resultado da decisão automatizada, mas como de todo processo de desenvolvimento de uma Inteligência Artificial, sinal disso é a habitualidade do "modo *crunch*" na indústria de software.

A desproporcionalidade de responsabilização dos programadores, aqueles diretamente envolvidos com o desenvolvimento da Inteligência Artificial, nos indica um critério importante a ser considerado, o investimento e o ganho econômico. Esta teoria de responsabilidade, comum nos Estados Unidos[69], é também conhecida como a responsabilização de *deep pockets*, ou seja, aqueles que estão engajados em atividade lucrativa e socialmente útil, mas que oferece risco, devem ser responsáveis pelo dano causado. Em outras palavras, aqueles com os bolsos mais fundos devem ser responsabilizados pela atividade de risco, como sumariza Mulholland:

> "Responsabilidade civil objetiva da sociedade que utiliza se beneficia e aufere lucros por meio da exploração da IA, objetivamente, por risco criado. Nesse sentido, uma interpretação possível do Art. 927 P.U. do Código Civil é de que, quando o legislador se refere a atividade que por sua natureza, implica risco aos direitos de outrem, poder-se ia interpretar extensivamente o conceito de IA como bens perigosos – por gerarem, potencialmente, danos qualitativamente graves e quantitativamente numerosos -, o que justificaria a responsabilidade por risco (rectius, perigo). De outro lado, poderia ser considerada a aplicação do CDC, sob os mesmos fundamentos já apresentados, aplicando-se a responsabilidade civil ao fornecedor – no caso, aquele que insere o sistema de IA no mercado de consumo – pelo fato do produto ou do serviço, amparada na presunção da existência de um defeito que ocasionou o dano, ainda que esse defeito

[69] CERKA, Paulis, GRIGIENE, Jurgita, SIRBIKYTE, Gintare **"Liability for damages caused by artificial Intelligence",** 2015, Computer Law & Security Review 31, Kaunas, Lithuania

> fosse desconhecido no momento em que o sistema de IA iniciou seu processo de desenvolvimento e autoaprendizagem."[70]

A possibilidade de responsabilização da sociedade controladora da Inteligência Artificial, utilizando como critério essencial o risco criado pela atividade econômica que se decidiu explorar não só se mostra mais adequada em relação à responsabilização do programador enquanto individuo, como também incentiva o desenvolvimento de ferramentas de aprimoramento de segurança e transparência, buscando evitar responsabilizações semelhantes. A questão da responsabilização vem sendo discutida a fundo na área da automação veicular e no que tange decisões automatizadas, podem oferecer também perspectivas de importância para a compreensão do cenário:

> Para cada algoritmo e componente de processamento de dados adicionado, os fabricantes também precisarão fazer uma análise de custo-benefício para saber se vale a pena o investimento para adicionar um recurso que aumente a segurança. O padrão para esta escolha terá que ser determinado pelos reguladores e possivelmente mais aperfeiçoado através de processos judiciais. Os fabricantes podem renunciar à adição de uma certa característica de segurança desde que a característica existente seja menos arriscada do que no carro tradicional e seja, portanto, uma melhoria em relação ao padrão de segurança atual? Existe uma avaliação monetária da vida humana que pode ser usada para determinar se o aumento marginal da segurança que resultará do investimento de fundos corporativos valeria a pena? É imoral para uma empresa lucrar se esse dinheiro poderia, de outra forma, ter acrescentado uma característica extra que poderia ter evitado um acidente e salvo uma vida? As respostas a estas perguntas determinarão diretamente o grau de segurança no automóvel de direção autônoma.[71] (tradução nossa)

No tópico de procedimentos preventivos, estes devem ir devem ir além também de soluções aparentemente simples para a questão do desenvolvimento de algoritmos. Vieses

[70] Idem p. 347

[71] For each added algorithm and data processing component, manufacturers will also need to do a cost-benefit analysis of whether it is worth the investment to add a safety-enhancing feature. The standard for this choice will have to be determined by regulators and possibly further refined through lawsuits. Can manufacturers forego the addition of a certain safety feature as long as the existing feature is less risky than in the traditional car and is thus an improvement on the current safety standard? Is there a monetary valuation of human life that can be used to determine whether the marginal increase in safety that will result from investment of corporate funds would be worth it? Is it immoral for a company to profit if that money could have otherwise gone into adding an extra feature that could have prevented an accident and saved a life? The answers to these questions will directly determine the degree of safety in the self-driving car.

históricos não serão solucionados de maneira simples ou individual, devendo se organizar um esforço coletivo não apenas por melhores práticas empresariais, como no caso de contratações e transparência, mas também no sentido de construção de espaços para a elevação em agentes políticos desses grupos discriminados.

7. Processos preventivos

Em linha com a proposta inicial deste projeto, deve-se esclarecer que não se buscará aqui apenas delinear programas de contratações e programas de diversidade, mas sim repensar algoritmos como pontos de disputa e de controle, como ferramentas que, ao contrário do comumente aceito, estão longe de serem neutras e que, portanto, podem ser políticos. Lidar com discriminação, especialmente quando implementada de maneira "neutra" e técnica, requer que não tratemos aqueles indivíduos e grupos discriminados como apenas definidos pela sua discriminação. Asad Haider, em sua obra "Armadilhas da Identidade", propõe a existência de um paradoxo do liberalismo, onde a defesa de uma identidade de lesão acaba por reduzir esse grupo à sua própria vitimização:

> Isso implica um "paradoxo" ao liberalismo, que persiste nos dias de hoje. Quando os direitos são concedidos a indivíduos "vazios", abstratos, eles ignoram as formas sociais reais de desigualdade e opressão que parecem estar fora da esfera política. No entanto, quando as especificidades das identidades lesadas são trazidas ao conteúdo dos direitos, Brown aponta que elas são "mais propensas a se tornarem lugar de produção e regulação da identidade como lesão do que veículos de emancipação". Em outras palavras, quando a linguagem liberal dos direitos é usada para defender uma identidade de grupo concreta da lesão física ou verbal, esse grupo acaba definido pela sua vitimização e os indivíduos acabam reduzidos a seu pertencimento como vítimas.[72]

Esse é o caminho que se toma quando nos restringimos à simples indicação da legislação necessária ou vigente ou quando indicamos processos de avaliação técnica para defender os direitos desse grupo. Há de se reconhecer que há uma necessária discussão que nos requer ir além do nível legislativo e técnico, rejeitando uma concepção estritamente legalista e essencialmente condescendente, tratando esses grupos como vítimas permanentes de sistemas imutáveis. É necessário pensar uma alternativa onde mulheres, negros, pobres, pessoas com

[72] HAIDER, Asad **"Armadilha da identidade: raça e classe nos dias de hoje"** Tradução de Leo Vinícius Liberato. – São Paulo: Veneta, 2019.

deficiência (PcDs) e outras categorias marginalizadas assumam a posição de agentes ativos, capazes de ação política e social.

7.1. Mudanças e Interdisciplinaridade

Como exemplo das limitações das propostas que propõem simplesmente maiores garantias ou apenas maior integração no desenvolvimento de tecnologias como solução única para a presença de vieses e discriminação em algoritmos, Ruha Benjamin[73] descreve a experiência de um ex-funcionário da Apple que, mesmo não sendo negro ou hispânico, propôs o treinamento da 'Siri' para identificar também inglês vernáculo afro-americano. Em resposta, escutou apenas que 'os produtos da Apple são destinados para o mercado premium':

> Um ex-funcionário da Apple que observou que ele não era "negro nem hispânico" descreveu sua experiência em uma equipe que estava desenvolvendo o reconhecimento da fala para Siri, o programa de assistente virtual. Como eles trabalhavam em diferentes dialetos ingleses - inglês australiano, cingapuriano e indiano - ele perguntou a seu chefe: "E o inglês afro-americano?". A isto seu chefe respondeu: "Bem, os produtos Apple são para o mercado premium". E isto aconteceu em 2015, "um ano após [o rapper] Dr. Dre ter vendido Beats do Dr. Dre à Apple por um bilhão de dólares". A ironia, o antigo funcionário parecia implicar, era que a empresa poderia de alguma forma desvalorizar e valorizar a Negritude ao mesmo tempo. Uma coisa é aproveitar-se do destaque de um artista negro para vender produtos (superfaturados) e outra bem diferente é envolver a especificidade cultural do povo negro o suficiente para melhorar o design subjacente de uma tecnologia amplamente utilizada. É por isso que a noção de que o preconceito tecnológico é "não intencional" ou "inconsciente" obscurece a realidade - que não há como criar algo sem alguma intenção, ou usuários específicos em mente.[74] (tradução nossa)

[73] BENJAMIN, Ruha "Race After Technology Abolitionist Tools for the New Jim Code" Polity Press, Medford, 2019

[74] A former Apple employee who noted that he was "not Black or Hispanic" described his experience on a team that was developing speech recognition for Siri, the virtual assistant program. As they worked on different English dialects – Australian, Singaporean, and Indian English – he asked his boss: "What about African American English?" To this his boss responded: "Well, Apple products are for the premium market." And this happened in 2015, "one year after [the rapper] Dr. Dre sold Beats by Dr. Dre to Apple for a billion dollars." The irony, the former employee seemed to imply, was that the company could somehow devalue and value Blackness at the same time.60 It is one thing to capitalize on the coolness of a Black artist to sell (overpriced) products and quite another to engage the cultural specificity of Black people enough to enhance the underlying design of a widely used technology. This is why the notion that tech bias is "unintentional" or "unconscious" obscures the reality – that there is no way to create something without some intention and intended user in mind.

É difícil imaginar que apenas programas de contratação seriam capazes de destrinchar da cultura dessa empresa a noção racista de que seus produtos não são nem mesmo destinados para a população negra. Novamente, não se trata de rejeitar tais possibilidades, elas são fundamentais para oferecer perspectivas diferentes no desenvolvimento de algoritmos, apenas não podemos assumi-la como uma 'saída fácil'.

Na mesma linha, a recente resolução do CNJ, já referenciada anteriormente, aponta no mesmo sentido da participação de equipes diversas no desenvolvimento de modelos de Inteligência Artificial, mas permite no mesmo artigo a sua dispensa em casos fundamentados, como a ausência de profissionais no quadro de pessoal dos tribunais, o que, considerando o objetivo da medida, pode se mostrar contraproducente:

> Art. 20. A composição de equipes para pesquisa, desenvolvimento e implantação das soluções computacionais que se utilizem de Inteligência Artificial será orientada pela busca da diversidade em seu mais amplo espectro, incluindo gênero, raça, etnia, cor, orientação sexual, pessoas com deficiência, geração e demais características individuais.
>
> § 1o A participação representativa deverá existir em todas as etapas do processo, tais como planejamento, coleta e processamento de dados, construção, verificação, validação e implementação dos modelos, tanto nas áreas técnicas como negociais.
>
> § 2o A diversidade na participação prevista no caput deste artigo apenas será dispensada mediante decisão fundamentada, dentre outros motivos, pela ausência de profissionais no quadro de pessoal dos tribunais.
>
> § 3o As vagas destinadas à capacitação na área de Inteligência Artificial serão, sempre que possível, distribuídas com observância à diversidade.
>
> § 4o A formação das equipes mencionadas no caput deverá considerar seu caráter interdisciplinar, incluindo profissionais de Tecnologia da Informação e de outras áreas cujo conhecimento científico possa contribuir para pesquisa, desenvolvimento ou implantação do sistema inteligente.[75]

Apesar da escolha de possibilitar a ausência de tais profissionais no quadro de desenvolvimento, a resolução mostra sua disposição para a construção de meios

[75] PRESIDENTE DO CONSELHO NACIONAL DE JUSTIÇA, Resolução Nº 332 de 21/08/2020, CNJ, Disponível em: https://atos.cnj.jus.br/files/original191707202008255f4563b35f8e8.pdf Acessado em: 10/05/2021

interdisciplinares para o desenvolvimento e demonstrou que há preocupação em relação à diversidade nesse processo.

Assim como quando se lida com código, não é possível se limitar à uma simples análise e remediação de seus "*outputs*", é necessário ir além, destrinchando o código, os seus bancos de dados e modelos que levaram a um resultado, da mesma forma, é necessário, no contexto discriminatório, como diz Silvio Almeida em seu prefácio ao livro "Armadilha da Identidade" de Asad Haider, "questionar o próprio maquinário social que produz as identidades sociais." [76]

Embora um passo fundamental, não basta apenas mais mulheres, mais imigrantes, mais negros na indústria de tecnologia, essas pessoas, acima de tudo, precisam ser escutadas, tomar a frente dessas lutas e, por que não, desenvolver também ferramentas pensadas nessa necessária reestruturação social. Ir além das garantias legais antidiscriminatórias significa conquistar agência e potência e conquistar agência e potência significa a capacidade de avançar contra as próprias estruturas discriminatórias.

Essa noção simples de mudança, que Ruha Benjamin descreve como inocente, já mostra sinais de desgaste, com milhares de funcionários do Google tendo se manifestado contra a colaboração da empresa com programas do pentágono visando a implementação de inteligência artificial para usos militares, já não há como tratar separadamente a ética da tecnologia e a sua utilização efetiva pelas estruturas que nos cercam.

Além da criatividade necessária para o desenvolvimento de inovações tecnológicas, Ruha Benjamin[77], pontua a necessidade de 'criatividade moral' e uma abordagem socialmente consciente do desenvolvimento tecnologia, treinamento e desenvolvendo inteligência artificial com um foco na promoção dessa igualdade, sendo necessário romper com a simples reprodução do passado.

> Se, como Cathy O'Neil escreve, "os grandes processos de dados codificam o passado. Eles não inventam o futuro. Fazer isso requer imaginação moral, e isso é algo que só os humanos podem proporcionar", então o que precisamos é de maior investimento

[76] ALMEIDA, Silvio **"Prefácio da edição brasileira de Armadilha da identidade"** IN **HAIDER, Asad "Armadilha da identidade: Raça e Classe nos dias de Hoje**" Tradução de Leo Vinícius Liberato. – São Paulo: Veneta, 2019. Pag. 18

[77] BENJAMIN, Ruha. "Race After Technology Abolitionist Tools for the New Jim Code". Medford, MA: Polity Press. 2019.

> em imaginários socialmente justos. Isto, creio, teria que implicar uma abordagem socialmente consciente do desenvolvimento tecnológico que exigiria priorizar a equidade sobre a eficiência, o bem social sobre os imperativos do mercado. Dada a importância dos conjuntos de treinamento no aprendizado de máquinas, outro conjunto de intervenções exigiria a criação de programas de computador a partir do zero e o treinamento de inteligência artificial "como uma criança ", a fim de nos conscientizar sobre os preconceitos sociais. (tradução nossa)[78]

É necessário a elaboração de uma visão crítica sobre a influência que essas ferramentas possuem sobre nossa vida e questionar quais instituições elas reproduzem, se, como Cathy O'Neil pontua diversas vezes em sua obra, aqueles que desenvolvem algoritmos partem de um grupo majoritariamente homogêneo, temos também de pensar sobre os grupos que em nosso país decidiram sobre essas questões e sobre suas interpretações, tanto sobre a tecnologia e sua atuação, quanto das necessidades atuais no combate à discriminação, quais elementos vão ser autorizados e quais proibidos?

8. Perspectivas e Conclusão

Buscou-se por meio desse trabalho discutir tecnologia e em específico algoritmos em relação às suas consequências sociais, as apresentando como tecnologias que, apesar das aparências e intenções, não podem ser lidas como neutras, mas sim importantes pontos de disputa. Como Laura Denardis já referência em sua obra, Andre Feenberg, filósofo da tecnologia, sugere que "a tecnologia é poder nas sociedades modernas, um poder de maior domínio do que o sistema político em muitas áreas".[79]

Consequentemente, há necessidade de enfrentar a questão de algoritmos a partir de uma perspectiva que não os considera como meras ferramentas técnicas, requerendo simples abordagem técnica regulatória, mas como uma ferramenta que pode vir a causar dano direcionado à grupos estigmatizados. Esse aspecto de direcionamento de danos, como se espera

[78] If, as Cathy O'Neil writes, "Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide,"50 then what we need is greater investment in socially just imaginaries. This, I think, would have to entail a socially conscious approach to tech development that would require prioritizing equity over efficiency, social good over market imperatives. Given the importance of training sets in *machine learning*, another set of interventions would require designing computer programs from scratch and training AI "like a child," so as to make us aware of social biases.

[79] DENARDIS, 2020, p. n.p.

ter demonstrado ao longo do texto, não decorre de um simples acidente computacional, mas de uma reprodução específica de nosso passado e sociedade.

Esse peso, que hoje carregamos, e que as nossas próximas gerações terão também de carregar, apesar de se expressar como sombras computacionais de nosso histórico, não foi de forma alguma extirpado de nossa sociedade. As estruturas que perpetuam violências sistêmicas continuam em voga; não há como se superar opressões históricas apenas por medidas legislativas.

No que tange a perspectivas, há de se esperar que, com a intensificação da utilização de algoritmos e a variedade que se espera com novas implementações da tecnologia, teremos de lidar, mais cedo ou mais tarde, com os seus desdobramentos. Se não houver atividade política desses grupos estigmatizados, se optarmos apenas pelo tratamento estritamente técnico ou legalista renunciaremos possibilidades verdadeiramente emancipatórias.

Além disso, esse trabalho busca também tatear as previsões legais que podem ser utilizadas para lidar com a questão da responsabilização e identificação de decisões enviesadas, será importante acompanhar a implementação da Lei Geral de Proteção de Dados na identificação e reforma dessas decisões e se teremos, assim como se discute em relação à GDPR, possíveis auditorias públicas de algoritmos em um futuro não tão distante.

Apesar do esforço, deve-se dizer que os temas retratados possuem profundidade considerável e que merecem aprofundamento compatível, o que nem sempre foi possível no escopo deste projeto, de modo que não há esgotamento de qualquer questão trabalhada, destaca-se por exemplo as discussões relacionadas com o tratamento de modelos algoritmos como dados sensíveis em razão da utilização e tratamento de dados e a sua possível recuperação via ataques assimétricos. [80]

Há, sem dúvidas, necessidade de se assumir melhores práticas de contratação, visando diversidade no próprio desenvolvimento de algoritmos de modo que se busque impedir vieses de surgirem, combatendo a homogeneidade. No entanto, será apenas conjuntamente com a

[80] VEALE M, BINNS R, EDWARDS L. “Algorithms that remember: model inversion attacks and data protection law.” Phil. Trans. R. Soc. Disponível em: http://dx.doi.org/10.1098/rsta.2018.0083 Acessado em 10/05/2020.

atividade política de grupos estigmatizados que garanta, não apenas no desenvolvimento, mas também na própria luta social contra a estigmatização, a voz ativa desses grupos na luta contra a discriminação estrutural, que será possível verdadeiramente enfrentar esses vieses. Tudo está em nossas mãos, precisamos enfrentar nossa história.

Referências

ALMEIDA, Silvio “Prefácio da edição brasileira de Armadilha da identidade” IN HAIDER, Asad “Armadilha da identidade: Raça e Classe nos dias de Hoje” Tradução de Leo Vinícius Liberato. – São Paulo: Veneta, 2019.

BENJAMIN, Ruha. “Race After Technology Abolitionist Tools for the New Jim Code”. Medford, MA: Polity Press. 2019.

BINNS, Reuben “Data protection impact assessments: a meta-regulatory approach”, International Data Privacy Law, Vol. 7(1), p. 22-35, University of Oxford, Inglaterra, 2017

BINNS, Reubens. “What Can Political Philosophy Teach Us about Algorithmic Fairness?”, IEEE Security & Privacy, University of Oxford, Inglaterra, 2018.

BURKOV, Andriy. “The Hundred-Page Machine Learning Book”, [s.n], [s.l.], 2019.

CARR, Jemma “University lecturer slams 'sexist' Google Translate as gender neutral languages are translated into English with gendered pronouns suggesting men 'build' and women 'wash dishes'”24 de Março 2021, Disponível em: <https://tinyurl.com/yy4yan8c> Acessado em: 30/03/2021

CERKA, Paulis, GRIGIENE, Jurgita, SIRBIKYTE, Gintare “Liability for damages caused by artificial Intelligence”, 2015, Computer Law & Security Review 31, Kaunas, Lithuania

CHUNG J. “Mind, Machine, and Society: Legal and Ethical Implications of Self Driving Cars,” tese de doutoramento, 2018

CLARK, D. D. "The Design Philosophy of the DARPA Internet Protocols”. Massachusetts Institute of Technology. Proc. SIGCOMM ‘88, Computer Communication Review Vol. 18, No. 4, 1988.

CLARK, David D. “The Design Philosophy of the DARPA Internet Protocols”, Massachusetts Institute of Technology, Proc. SIGCOMM ‘88, Computer Communication Review Vol. 18, No. 4, August 1988, pp. 106–114

CORMEN, Thomas H., LEISERSON, Charles. E., RIVEST, Ronald. L., & STEIN, Clifford. “Introduction to Algorithms”. Estados Unidos, Massachusetts, The MIT Press. 2009.

DENARDIS, Laura. “The internet in Everything: freedom and security in a world with no off switch”. Connecticut: Yale University Book Press, 2020.

EDER, Nikes “Privacy, Non-Discrimination and Equal Treatment: Developing a Fundamental Rights Response to Behavioural Profiling” em “Algorithmic Governance and Governance of Algorithms”, Springer, Suíça, 2021

EDER, Nikes. “Privacy, Non-Discrimination and Equal Treatment: Developing a Fundamental Rights Response to Behavioural Profiling” em “Algorithmic Governance and Governance of Algorithms”, Springer, Suíça, 2021

EUBANKS, Virgina. “Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor”. New York: St. Martin’s Press. 2018

EUROPEAN PARLIAMENT AND OF THE COUNCIL “Regulamento Geral sobre a Proteção de Dados”, 2016, Disponível em: <https://gdprinfo.eu/pt-pt/pt-pt-article-22> Acessado em: 10/05/2021.

HAIDER, Asad “Armadilha da identidade: raça e classe nos dias de hoje” Tradução de Leo Vinícius Liberato. – São Paulo: Veneta, 2019.

JOHNSON, Gabbrielle M. “Algorithmic bias: on the implicit biases of social technology” [s.n.], [s.l.], 2020.

JOHNSON, Khari. “Government audit of AI with ties to white supremacy finds no AI” VentureBeat, 5 de Abril de 2021, disponível em: <https://tinyurl.com/hmh7utz2> Acessado em: 11 de Abril de 2021.

JOHNSON, Melvin “A Scalable Approach to Reducing Gender Bias in Google Translate” Google Ai Blog 22 de abril de 2020, Disponível em: < https://tinyurl.com/4r4c5tej> Acessado em: 30/03/2021

KEARNS, Michael, ROTH, Aaron. “The Ethical Algorithm”. New York: Oxford University Press 2020.

LEVINE, Yasha “Surveillance Valley: The Secret Military History of the Internet” Perseus Books, LLC 2018, Nova York. 2018

LEVINE, Yasha. “Surveillance Valley: The Secret Military History of the Internet”. Nova York: Perseus Books, LLC. 2018

MARX, Karl. “O Capital: Crítica da economia política. Livro I: O processo de produção do capital.” Trad. Rubens Enderle. São Paulo: Boitempo, 2013, pp. 548

MEARIN, Lucas “A diesel whodunit: How software let VW cheat on emissions” 23 de setembro, 2015, Computer World, Estados Unidos, Disponível em: < https://tinyurl.com/khnj3s7t> Acessado em: 30/03/2021

MOREIRA, Adilson José. “Pensando como um negro: ensaio de hermenêutica jurídica” São Paulo: Editora Contracorrente, 2019 n.p.

MOREIRA, Adilson José. “Tratado de Direito Antidiscriminatório”. São Paulo: Editora Contracorrente. 2020

MULHOLLAND “Responsabilidade civil e processos decisórios autônomos em sistemas de Inteligência Artificial (IA): autonomia, imputabilidade e responsabilidade.” Em “Inteligência Artifical e direito: ética, regulação e responsabilidade” Coord: Ana Frazão e Caitlin Mulholland – 2. Ed. rev. Atual. E ampl. – São Paulo, Thomsom Reuters Brasil, 2020.

MULHOLLAND, Caitlin. FRAJHOF Isabella Z. “Inteligência Artifical e a Lei Geral de Proteção de Dados Pessoais: breves anotações sobre o direito à explicação perante a tomada de decisões por meio de machine learning.” Em “Inteligência Artifical e direito: ética, regulação e

responsabilidade" Coord: Ana Frazão e Caitlin Mulholland – 2. Ed. rev. Atual. E ampl. – São Paulo, Thomsom Reuters Brasil, 2020.

NEUBAUER, Andre. FREUDENBERGER, Jurgen. KUHN, Volker. "Coding Theory - Algorithms, Architectures, and Applications", John Wiley & Sons, Ltd. Chincester, Inglaterra, 2007

NEWMAN, David T., FAST, Nathanael J., HARMON, Derek. J. "When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decision" Organizational Behavior and Human Decision Processes Volume 160, 2020, Pag. 149-167

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR. Fórum Da Internet No Brasil "Discriminações algorítmicas: impactos na sociedade, perspectivas e soluções". 2020. Disponível em: < https://tinyurl.com/3hxn3fpb> Acesso em em: 10/11/2020

O' NEIL, Cathy. "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy". New York: Crown. 2016

PRESIDENTE DO CONSELHO NACIONAL DE JUSTIÇA, (Presidência CNJ), "Resolução Nº 332 de 21/08/2020, CNJ," Disponível em: https://tinyurl.com/25shte73 Acessado em: 10/05/2021

ROUGHGARDEN, Tim. "Algorithms Illuminated Part 1 The Basics" San Francisco, CA, Soundlikeyourself Publishing, LLC, 2014.

SEDGEWICK, Robert, WAYNE, Kevin. "Algorithms Part I, 4th Edition", Princeton University, 2014.

SHALEV-SHWARTZ, Shai., BEN-DAVID Shai. "Understanding Machine Learning From Theory to Algorithms" Cambridge University Press, 2014.

SIEMS, Jasper. "Protecting Deep Learning: Could the New EU-Trade Secrets Directive Be an Option for the Legal Protection of Artificial Neural Networks?" Editorial: "Algorithmic Governance and Governance of Algorithms", 2021, Springer, Umea Sweden. P. 153

SILVA, Priscilla. MEDEIROS, Juliana, “A polêmica da revisão (humana) sobre decisões automatizadas” Disponível em: <https://tinyurl.com/7sy8pp8n> Acessado em: 10/05/2021

SKIENA, Steven S. “The Algorithm Design Manual” Third Edition, Springer, NY, USA. 2020.

VEALE, Michael, BINNS, Reuben, EDWARDS, Lilian. “Algorithms that remember: model inversion attacks and data protection law.” Phil. Trans. R. Soc. Disponível em: <https://tinyurl.com/w2v9m86d> Acessado em 10/05/2020

VEALE, Michael. BINN, Reuben. EDWARDS, Lilian. “Algorithms that remember: model inversion attacks and data protection law.” Philosophical Transactions of the Royal Society, [s.l.] 2018

VEALE, Michael. BINNS, Reuben. AUSLOOS, Jef. “When data protection by design and data subject rights clash” International Data Privacy Law, Volume 8, Edição 2, Inglaterra, 2018

I AM NOT YOUR NEGRO; Direção Raul Peck. Produção: Velvet Film, Estados Unidos: Magnolia Pictures, 2016. (95 min.)

TERMO DE AUTENTICIDADE DO TRABALHO DE CONCLUSÃO DE CURSO

Eu, Rafael Afonso Cristino Sousa Barros
discente regularmente matriculado(a) na disciplina TCC II, da 10ª etapa do curso de Direito, matrícula nº 41631331, período noturno, turma S, tendo realizado o TCC com o título: Decisões Automatizadas e Viés Algorítmico: Identificação, Responsabilização e Perspectivas
sob a orientação do(a) Professor(a) Ivandick Cruzelles Rodrigues
declaro para os devidos fins que tenho pleno conhecimento das regras metodológicas para confecção do Trabalho de Conclusão de Curso (TCC), informando que o realizei sem plágio de obras literárias ou a utilização de qualquer meio irregular.

Declaro ainda que, estou ciente que caso sejam detectadas irregularidades referentes às citações das fontes e/ou desrespeito às normas técnicas próprias relativas aos direitos autorais de obras utilizadas na confecção do trabalho, serão aplicáveis as sanções legais de natureza civil, penal e administrativa, além da reprovação automática, impedindo a conclusão do curso.

São Paulo, 21 de maio de 2021.

Assinatura do discente

TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DO TRABALHO DE CONCLUSÃO DE CURSO

Material Bibliográfico: ( ) Artigo Científico (X)Monografia

Graduação em Direito

Título do Trabalho: Decisões Automatizadas E Viés Algorítmico: Identificação, Responsabilização e Perspectivas

Nome do Autor(a): Rafael Afonso Cristino Sousa Barros

E-mail: RafaelBarros@protonmail.com

Este e-mail pode ser divulgado (X) SIM ( ) NÃO

Orientador(a): Ivandick Cruzelles Rodrigues

Na qualidade de titular dos direitos autorais da publicação supracitada, de acordo com a Lei nº 9.610/98, (X) AUTORIZO ( ) NÃO AUTORIZO a Universidade Presbiteriana Mackenzie – UPM, a disponibilizar gratuitamente, sem ressarcimento dos direitos autorais, o documento, em meio eletrônico, no *site* da base de dados Adelpha, para fins de leitura pela internet, a título de divulgação da produção científica gerada pela Universidade, a partir desta data. Igualmente, declaro que a versão do Trabalho de Conclusão de Curso entregue em meio eletrônico corresponde fielmente e na integra à versão similar depositada de forma impressa em papel para a defesa ou apresentação.

Motivos no Caso de Não Autorização

( )Exigência de periódico de não divulgação até a publicação (exige justificativa, informe e nome do periódico)

( )Outros (justificar):

São Paulo, 21 de Maio de 2021.

Assinatura do(a) Autor(a)