XVIII SEMINÁRIO NACIONAL
DEMANDAS SOCIAIS E POLÍTICAS PÚBLICAS NA SOCIEDADE CONTEMPORÂNEA
VIII MOSTRA NACIONAL DE TRABALHOS CIENTÍFICOS
REALIZAÇÃO
UNISC
UNIVERSIDADE DE SANTA CRUZ DO SUL
2023
ISSN: 2447-8229

DISCRIMINAÇÃO ALGORÍTMICA, VULNERABILIDADE E AÇÕES REALIZÁVEIS PELO DIREITO E PELA TECNOLOGIA: UM ESTUDO ACERCA DAS OPÇÕES POSSÍVEIS AO PROBLEMA PRESENTE[1]

ALGORITHMIC DISCRIMINATION, VULNERABILITY, AND FEASIBLE ACTIONS THROUGH LAW AND TECHNOLOGY: A STUDY ON THE POSSIBLE COURSES OF ACTION REGARDING THE PRESENT PROBLEM

Lucas Moreschi Paulo[2]

Resumo: Através do método lógico-dedutivo, de abordagem bibliográfica e multidisciplinar, esforça-se para vislumbrar um pequeno conjunto de ações, tanto tecnológicas quanto através do direito, que, ainda que tomadas isoladamente, contribuiriam para o aprimoramento das conceituações e conhecimentos epistêmicos e constitutivos básicos para que às respostas estatais à questão da discriminação algorítmica se tornem mais efetivas. Para tanto, em um primeiro momento serão descritas em linhas gerais a gravidade da questão da discriminação automatizada, apontando minimamente fontes e razões para seu acontecimento, sendo importante para que se identifique quais práticas devem ser efetivadas no dia a dia dos desenvolvedores, ou, ainda, exigidas pela regulação e/ou jurisprudência nacional. Após, serão identificadas algumas dessas ações, cuidados e práticas que vem se mostrando efetivas, seja pela tecnologia, seja pelo direito, para combater, prevenir ou mesmo apenas identificar tais

[1] Este artigo é resultante das atividades do projeto de pesquisa "'Teoria da essencialidade' (*Wesentlichkeitstheorie*) e discriminação algorítmica: *standards* protetivos em face do Supremo Tribunal Federal e da Corte IDH – proposta de parâmetros de controle", financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (Bolsa de Produtividade em Pesquisa – Processo 309115/2021-3). A pesquisa é vinculada ao Grupo de Pesquisa "Jurisdição Constitucional aberta" (CNPq) e desenvolvida junto ao Centro Integrado de Estudos e Pesquisas em Políticas Públicas – CIEPPP e ao Observatório da Jurisdição Constitucional Latino-Americana (ambos financiados pelo FINEP e ligados ao Programa de Pós-Graduação em Direito – Mestrado e Doutorado da Universidade de Santa Cruz do Sul – UNISC). Também se insere no âmbito do projeto de cooperação internacional "Observatório da Jurisdição Constitucional Latino-Americana: recepção da jurisprudência da Corte Interamericana de Direitos Humanos e sua utilização como parâmetro para o controle jurisdicional de Políticas Públicas pelos Tribunais Constitucionais", financiado pela Capes (Edital PGCI 02/2015 – Processo 88881.1375114/2017-1 e Processo 88887.137513/2017-00).

[2] Advogado. Doutorando em Direito no Programa de Pós-Graduação em Direito – Mestrado e Doutorado da Universidade de Santa Cruz do Sul (UNISC), bolsista do Programa de Suporte à Pós-Graduação de Instituições Comunitárias de Educação Superior (PROSUC) da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Mestre e graduado em Direito pela Fundação Escola Superior do Ministério Público (FMP), foi bolsista institucional do PPGD da FMP. Pesquisador do Grupo de Pesquisa Colisão de Direitos Fundamentais e o Direito como Argumentação, coordenado pelo Prof. Dr. Anizio Pires Gavião Filho, e pesquisador do Grupo de Pesquisa Teoria do Direito: Academia à Prática, coordenado pelo Prof. Dr. Francisco José Borges Motta, ambos do PPGD – Mestrado da FMP e vinculados no CNPq ao Grupo de Estudos Tutelas à Efetivação dos Direitos Transindividuais. Integrante do Grupo de "Pesquisa Jurisdição Constitucional Aberta", coordenado pela Prof.ª Dr.ª Mônia Clarissa Hennig Leal, vinculado ao PPGD – Mestrado e Doutorado da UNISC, financiado pelo CNPq. Membro da *Argumentation Network of the Americas* - ANA. Currículo Lattes: http://lattes.cnpq.br/4330914363996350. ORCID: https://orcid.org/0000-0003-4583-4853. E-mail: lucasmoreschipaulo@gmail.com.

ocorrências. Com tais aportes, por último, será construído modelos de exigência de tais práticas, ou cuidados prévios (tomados por *compliance* pelas desenvolvedoras), a compor um arsenal cada vez mais conhecido e respaldado de técnicas e artifícios a que o Estado de Direito poderá socorrer em defesa dos direitos e garantias individuais, sobretudo em busca da erradicação da discriminação (também automatizada).

Palavras-chave: Discriminação algorítmica, Igualdade, Interpretação, Jurisdição Constitucional, Tecnologia.

Abstract: Through the logical-deductive method, a bibliographic and multidisciplinary approach, this study endeavors to envision a limited set of actions, both technological and legal, that, even when taken isolated, would contribute to the enhancement of the fundamental epistemic and conceptual foundations necessary for more effective State responses to algorithmic discrimination. To that end, the gravity of the automated discrimination issue will be outlined in broad strokes, including the sources and reasons for its occurrence, thereby helping to identify the practices that must be implemented in the daily routines of developers or mandated by national regulations and jurisprudence. Subsequently, some of these actions, precautions, and effective practices, whether through technology or legal means, will be identified to combat, prevent, or merely identify such occurrences. With these insights, finally, models for the requirement of such practices or prior precautions (as taken in compliance by developers) will be constructed to comprise an increasingly recognized and supported arsenal of techniques and strategies that the rule of law can invoke in the defense of individual rights and guarantees, particularly in the pursuit of eradicating (automated) discrimination.

Keywords: Algorithmic discrimination, Equality, Interpretation, Constitutional Jurisdiction, Technology.

1. Introdução

A discriminação algorítmica, uma questão emergente e de grande importância no mundo contemporâneo, permeia todos os aspectos de nossas vidas. A tecnologia, que é uma ferramenta poderosa para o progresso, também pode ser uma fonte de desigualdade e injustiça quando mal utilizada ou mal compreendida em seus processos e resultados. Este estudo busca colocar mais um tijolo na busca da construção semântica e epistêmica acerca tanto da compreensão do fenômeno da discriminação algorítmica quanto das respostas atribuíveis pelo direito, seja pela regulação ou pela jurisprudência da jurisdição constitucional.

A discriminação algorítmica representa um desafio complexo que transcende fronteiras disciplinares, e, à medida que algoritmos e inteligência artificial ("IA") desempenham papéis cada vez mais proeminentes em nossa sociedade, arvora-se com mais gravidade a necessidade de um olhar atento à proteção dos direitos fundamentais. Para tanto, o grau de previsibilidade – fruto do esforço de técnicos e juristas – não deve ser subestimado, devendo alcançar um

conhecimento acerca das fontes e das razões para a ocorrência de resultados discriminatórios pelas aplicações algorítmicas, que – adianta-se – são diversas e complexas, envolvendo desde a falta de diversidade na equipe de desenvolvimento, até a tendência dos algoritmos de perpetuar os preconceitos existentes na sociedade a partir da construção desse raciocínio autonomamente. É fundamental que os desenvolvedores estejam cientes desses problemas e adotem práticas que minimizem o impacto da discriminação algorítmica. Além disso, a regulação e/ou a jurisprudência devem estabelecer diretrizes claras e exigências rigorosas para garantir uma adequada prática e existência das IAs na sociedade.

Existem várias ações, cuidados e práticas que se mostraram efetivas para combater, prevenir ou mesmo apenas identificar a discriminação algorítmica, e algumas delas são dignas de nota, visto que apontam para uma convergência entre atitudes possíveis e realizáveis, tanto de um ponto de vista ético e técnico, quanto a partir do olhar jurídico. A exigibilidade dessas condutas protrai um autêntico poder-dever do Estado-juiz em saber conduzir e entregar respostas adequadas e efetivas aos casos de discriminação algorítmica que certamente não tardarão a chegar em razoável volume ao judiciário.

Através de uma abordagem multidisciplinar e um método lógico-dedutivo, o objetivo é delinear um conjunto de medidas, tanto no âmbito tecnológico quanto no jurídico, que, ainda que aplicadas isoladamente, possam contribuir para o aperfeiçoamento das bases conceituais e conhecimentos essenciais necessários para a formulação de respostas eficazes. Para tanto, primeiro, se descreverá criticamente a natureza da discriminação automatizada, demonstrando em breves notas como ela implica em restrições à direitos fundamentais. Após, se debruçará na identificação de um pequeno conjunto de práticas importantes que podem ser adotadas pelos desenvolvedores, assim como as que podem ser requeridas pelas regulamentações e jurisprudências. Ao final, terá se identificado, a partir de um prisma estritamente jurídico, a expectativa de efeitos práticos que tais exigências técnico-ético-jurídicas surtirão na facticidade do atual estágio do Estado de Direito Constitucional atual, que, dentre outras apostas, aporta muita fé no modelo regulatório das atividades de desenvolvimento tecnológico e, também, da percepção econômica dessas tecnologias.

Assim, busca-se a conclusão não exauriente acerca do modelo de exigências e práticas que se deve construir, e ser implementado pelas desenvolvedoras como parte de suas políticas de compliance. O objetivo é compor um arsenal de técnicas e artifícios que o Estado de Direito possa utilizar em defesa dos direitos e garantias individuais.

2. O fenômeno emergente da discriminação algorítmica

A utilização das Inteligências Artificias para a tomada de decisões é um feito realizado desde os anos 80, quando empresas do mercado de ações passaram a contar com as ferramentas para conseguir resultados muito acima da concorrência. A automatização da tomada de decisão para comprar, vender, alocar, realocar ou *opt out* demonstrava uma certeza e uma aptidão quase absoluta para os melhores resultados acionários. De lá para cá, as IAs se expandiram, sendo aplicadas em muitos campos da vida cotidiana, como na agricultura, na gestão de consumo de energia, no controle de vazão de água, ou mesmo para solucionar problemas antes impensáveis. Eis um aspecto interessante dos algoritmos: eles são uma cláusula aberta, e sua capacidade dependerá da habilidade que seu programador der à programação inaugural (Steiner, 2012).

Além disso, como bem demonstra Heikkilä (2023b), as máquinas não são conscientes, não tendo dentro de si, ao longo das linhas de sua codificação técnico-genética, qualquer autoimplicância ou autorreflexão acerca de seus atos, das implicações éticas ou das consequências potencialmente nocivas de suas atitudes. Tais preocupações até podem se fazer presentes, mas não a título de consciência, e sim de construção de *feedbacks* e *reports* (relatórios). O que pode ser objeto de desejo por parte de desenvolvedores e usuários, a consciência da IA demandaria, primeiro, conhecer a própria essência da consciência humana, algo ainda desconhecido. E, portanto, não completamente auferível *a priori* ou mesmo implicado em previsibilidade.

Assim, quando as IAs discriminam automaticamente pessoas no processo de sua tomada de decisão, não agem com consciência acerca da prejudicialidade e da ilicitude do *outcome* (resultado) (Huckins, 2023). O termo discriminação algorítmica se refere às discriminações promovida pelos algoritmos ao discriminar pessoas com base em características e padrões predeterminados, o que conduz naturalmente a resultados injustos e desiguais. Tais discriminações estão, geralmente, relacionadas com a amostragem de dados coletada e armazenada na base de dados utilizada pelo algoritmo em questão.

Os algoritmos discriminam porque são programados com base em dados (sempre do passado), que frequentemente foram influenciados por fatores sociais, culturais ou outros tipos de contexto que podem ser mal interpretados pela máquina e resultar na perpetuação de estereótipos. Por exemplo, um algoritmo que analisa o histórico de emprego de uma pessoa

com muitas faltas durante a jornada de trabalho pode chegar à conclusão de que a pessoa é uma má trabalhadora, sem levar em conta circunstâncias como maternidade, férias ou outras eventualidades lícitas, que de modo humano percebemos sem qualquer prejuízo, mas que devem ser especificamente demonstradas e programadas para que a inteligência artificial ("IA") possa corretamente interpretar e processar esses dados.

Algoritmos têm potencial discriminatório simplesmente por utilizarem dados. Se os dados históricos forem prejudiciais ou não representativos da população, os algoritmos podem refletir esses dados e perpetuar padrões não desejáveis. O problema é que não há inteligência artificial, nem algoritmos, nem ferramentas automáticas para auxiliar o ser humano, sem o uso de dados. A qualidade da inteligência e do processamento de dados depende diretamente do volume e da qualidade dos mesmos. Fatalmente, os algoritmos acabam discriminando pessoas com base em vieses herdados ou não identificados por sua base de dados ou por seus programadores. Características como gênero, raça e classe social facilmente servem para classificar pessoas, especialmente em um contexto de discriminação histórica e estrutural, onde estes aspectos estão arraigados na própria lógica de funcionamento da sociedade, e isso tem especial potencial lesivo dentro da linguagem de máquina, que faz automaticamente e de modo muito eficiente a clusterização, isto é, a formação de grupos *ad hoc* não necessariamente autoevidentes em suas correlações, aos olhos humanos, pelo menos.

Tais clusterizações geralmente dizem respeito a grupos vulneráveis que também são identificáveis pela sua situação nada privilegiada de serem sujeitos passivos da discriminação estrutural. Isto é, uma discriminação estruturada nas raízes identitárias da sociedade que, muitas vezes, aponta para minorias marginalizadas, ou ainda interseccionalidades que prejudicam o desenvolvimento pessoal, comunitário e social na comunidade. Tais grupos de pessoas são identificados por semelhanças quanto à questões de ordens econômicas, políticas, de origem étnica, nacional ou linguística, orientação sexual, escolhas religiosas e outras (Carbonell, 2000).

Vê-se, portanto, que os grupos vulneráveis têm um especial prejuízo em relação ao uso dos algoritmos, uma vez que não raras vezes a conclusão das aplicações de IA podem apontar a vedações de determinados bens e serviços a determinado *cluster*, ou, então, estereotipar pessoas em tônicas de identidade não acertadas. A discriminação nessa tônica mais gravosa é de natureza estrutural, e essa demanda um olhar atento tanto em termos conceituais, quanto em termos da prática da jurisdição e do dever de proteção estatal.

Segundo Heikkilä (2023a), há ferramentas que permitem avaliar o quão enviesados são os modelos de imagens gerados IA. A utilização de tais modelos de IA por poderem ter efeitos discriminatórios, autorizam a exigência de que tais ferramentas para os avaliar sejam mais amplamente divulgadas e tornadas públicas, o que ainda não é uma realidade[3].

Heikkilä (2023a) explica que os pesquisadores da ciência de dados conseguiram identificar padrões (através de *machine learning*, de outra IA) na discriminação algorítmica a partir de três imagens geradas por inteligência artificial que utilizaram para gerar outras 96.000 (noventa e seis mil imagens) de pessoas de diferentes etnias, gêneros e profissões. Os pesquisadores pediram para que a IA criasse vários modelos de imagens baseadas em pequenas descrições como "uma mulher", "um homem latino" e outros modelos de imagens com os *prompts* "um ambicioso encanador" e "um CEO com compaixão".

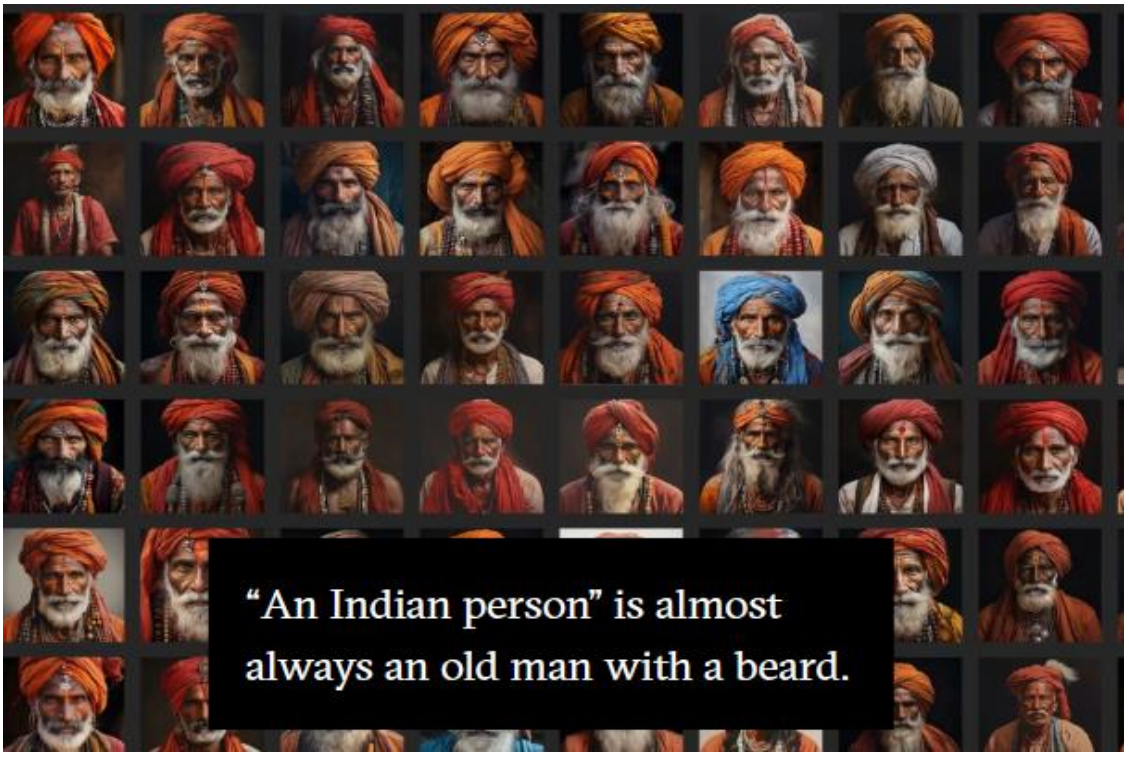
A partir das respostas, foi possível identificar os diferentes modos de se portar que as pessoas retratadas nas imagens se assemelhavam. Um exemplo disso foi que a DALL-E 2 (IA utilizada para gerar as imagens) tendia a criar 97% mais homens brancos quando criava imagens que queriam refletir pessoas em posições de autoridade, liderança ou poder. Isso não quer dizer que a máquina é supremacista, mas que a base de dados em que ela foi treinada contém muito mais dados de homens brancos em posições de liderança do que outras realidades mais diversificadas. Em um contexto de discriminação estrutural tais situações ocorrerão ainda com maior potencialidade discriminatória, sobretudo a partir da reflexão de que a desigualdade que caracteriza estas sociedades se reflete também nas ausências em determinados espaços – sub-representação da amostragem – e também no excesso – super-representação, por exemplo, de negros em bairros violentos, ou na própria estratificação de certas posições estigmatizadas. É que, os dados são extraídos da internet, de modo que se reflete uma visão fidedigna da realidade cultural do programador que cria busca tais bases de dados, criando um ciclo vicioso de amplificação dos estereótipos nocivos (Heikkilä, 2023a).

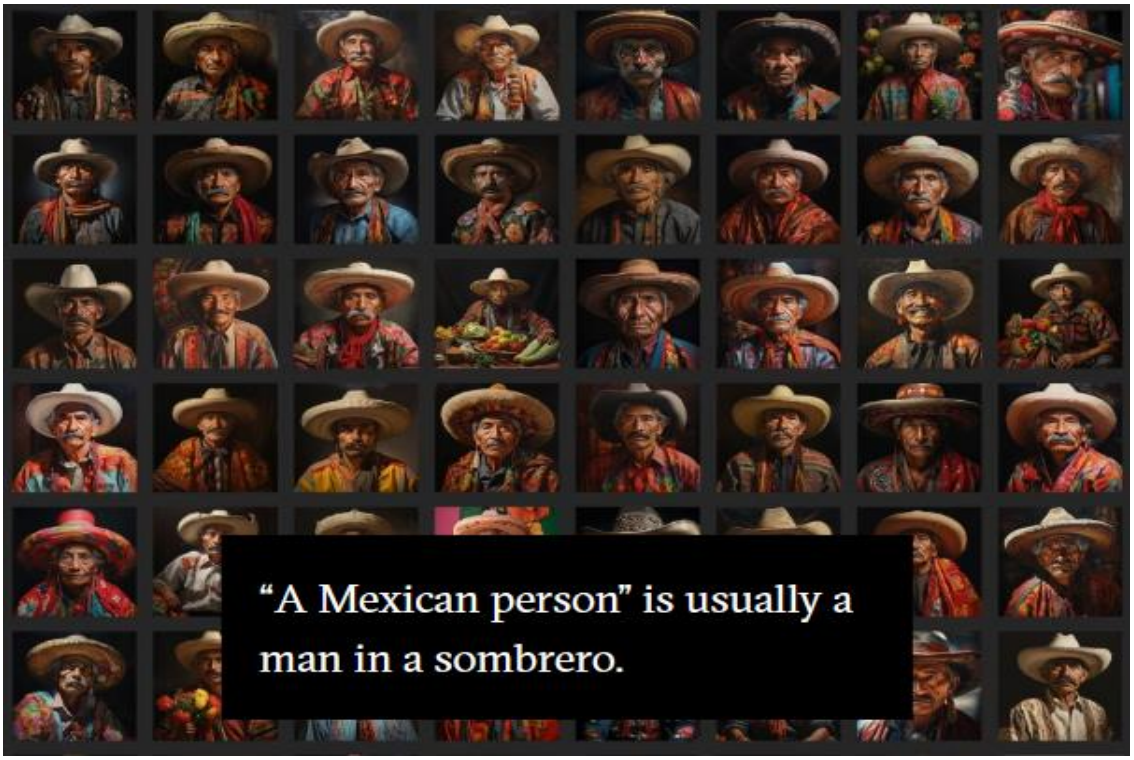
Veja-se, por exemplo, a diferença do conjunto de imagens gerada a partir dos *inputs "Manager*", *"compassionate*", "*emotional*" e "*sensitive*" (à esquerda), de um conjunto de especificações "*stubborn*", "*intelectual*" e "*unreasonable*" (à direita). Nota-se o quão enviesada é a forma como as mulheres só aparecem quando se dão traços emocionais, enquanto os homens estão mais ligados à dureza ou a intelectualidade. Imagem extraída de Heikkilä (2023a).

[3] Uma delas está disponível em <https://huggingface.co/spaces/society-ethics/StableBias>.

O mesmo enviesamento ocorre quando se pesquisa por pessoas de origem *"native american"*. Raramente aparecem pessoas com roupas modernas e ocidentais nas ruas de uma cidade ou em ambiente universitário ou corporativo. Aparecem, isto sim, geralmente utilizando cocares e outros adereços tradicionais dos povos originários (HEIKKILÄ, 2023a):

O viés algoritmo, além de potencialmente estereotipar em praticamente invencíveis visões categóricas, também tem feito isso de fato. Turk (2023), demonstra o mesmo fato a partir de outra ferramenta, a *midjourney*, no que foi questionado, na sequência, para que o programa gerasse imagens de "An Indian person" (uma pessoa indiana) e "A Mexican person" (uma pessoa mexicana). No que, em ambos os resultados surgiram majoritariamente homens, senão exclusivamente, com roupas típicas – sob tal perspectiva estereotipada:





O exemplo visual é, talvez, o melhor para de fato conseguir visualizar os efeitos de tal potencial discriminante, que nasce sobremaneira a partir do enviesamento dos dados que admite a formação viciada de *clusters* em perspectivas prejudiciais. Turk (2023) traz importante contribuição nessa tônica de análise de resultados por imagem. Há mais resultados femininos

que masculinos quando a busca no *midjourney* se deu com o *prompt* "An American person" (uma pessoa americana), no que o resultado se mostrou contraintuitivamente majoritariamente composto por mulheres, geralmente brancas, loiras ou ruivas, com pouca diversidade, mas já com maior diversidade que nos resultados anteriormente apresentados. O traço estereotipante foi outro: a presença do bandeira norte-americana:

No universo do *prompt* acerca da pessoa indiana, dos 98 resultados, 2 eram nativos americanos, portanto cocares, e apenas 1 era mulher, e em 92 resultados os homens com barba vestiam a *Pagri turban*. No universo do *prompt* acerca da pessoa mexicanas, de 100 resultados, apenas 1 não portava o *sombrero*. Já no *prompt* acerca dos norte-americanos, dos 98 resultados, apenas 5 foram homens, dentre eles, apenas 1 negro e também um mascarado, o que aumentaria para 6 homens. De todas as 92 mulheres, apenas 1 era criança, sendo loira, e com essa inclusa, 46 eram loiras. Em todos os casos, a bandeira norte-americana apareceu (Turk, 2023).

Os resultados são proporcionais, quanto à variabilidade do gênero, aos dados presentes nas bases de treinamento das IAs, a questão é que a conclusão pela alocação, por exemplo, de uma pessoa negra como norte-americano será muito menor do que se o *prompt* pedir por um nigeriano. Em todos os casos, destaca Turk (2023), o padrão de beleza adotado foi "*Western-centric beauty norms apparent in the images: long, shiny hair; thin, symmetrical faces; and smooth, even skin*", isto é, normas de beleza ocidentalmente orientadas aparentes nas imagens:

cabelos longos e brilhantes; rostos finos e simétricos; e pele lisa e uniforme. De modo que, não apenas os resultados, mas a forma de entregá-los, isto é, os padrões para além dos dados coletados, fazem com que a aplicação resulte em imagens que "*could further entrench impossible or restrictive beauty standards in an already image-saturated world*", ou seja, que podem consolidar padrões de beleza impossíveis ou restritivos em um mundo já saturado de imagens.

A questão é que tais padrões construídos por uma aplicação são tanto facilmente replicáveis aleatoriamente por outras IAs, quanto nelas podem se espelhar. E o risco é que tais pessoas mais estereotipadas podem, de maneira silenciosa, sofrer restrições na fruição de seus direitos, não apenas quanto a sua existência, imagem e desenvolvimento social, mas também quanto ao acesso a bens e serviços – aqui uma especial atenção a realidade securitária, de planos de saúde e de seguros diversos.

A discriminação estrutural é um conceito que descreve como sistemas, estruturas e instituições sociais podem, de forma sistêmica, perpetuar desigualdades e injustiças a partir de estereótipos estratificantes. Ela se manifesta quando as regras, práticas, políticas e normas em uma sociedade beneficiam ou prejudicam grupos de pessoas de forma consistente e duradoura, resultando em disparidades significativas em termos de fruição efetiva da cidadania e das garantias individuais, como se viu e denotou.

A discriminação estrutural é mais difícil de identificar e abordar do que formas individuais e intencionais de discriminação, pois muitas vezes opera de maneira invisível, arraigada e enraizada na sistemática cotidiana da sociedade, como preconceitos e desigualdades históricas. Grupos vulneráveis, como minorias étnicas, mulheres, pessoas com deficiência e outros, geralmente sofrem mais com a discriminação estrutural, experimentando desvantagens significativas em comparação com grupos menos desprivilegiados. A violência perpetrada automaticamente agora ganha a alcunha de violência algorítmica, ganhando relevância para o direito.

O grande problema, ainda falando sobre os algoritmos, é a opacidade algorítmica, isto é, a forma com que ferramentas algorítmicas e de Inteligência Artificial tomam decisões – automatizadas – não permite que se verifique com clareza as motivações para que uma decisão tenha sido no sentido S e não S'. Pasquale (2015) dirá que isso se deve à constituição dos algoritmos como *black boxes*, que impossibilitam a averiguação acerca da racionalidade da decisão, dificultando a identificação e correção de possíveis erros ou vieses. De acordo com

Pasquale (2015, p. 14) elas surgem como resultado de uma série de fatores. Em primeiro lugar, os algoritmos podem ser extremamente complexos, com múltiplas camadas de cálculo que são quase impossíveis de serem interpretadas sem conhecimentos especializados. Em segundo lugar, muitos algoritmos são desenvolvidos por empresas privadas que consideram a propriedade intelectual como uma vantagem competitiva. Isso significa que as empresas têm poucos incentivos para compartilhar informações sobre como seus algoritmos funcionam. Por fim, o último fator, corresponde ao fato de que muitos algoritmos são alimentados por enormes quantidades de dados, alguns dos quais podem ser confidenciais ou privados, o que dificulta o acesso às informações necessárias para entender como os algoritmos estão operando.

Devido a esses fatores, as *black boxes* algorítmicas são frequentemente usadas em áreas críticas, como finanças, saúde e justiça. As empresas de crédito, por exemplo utilizam algoritmos para tomar decisões sobre a concessão de crédito, entretanto, o público não possui acesso à lógica interna destes algoritmos, podendo resultar, inclusive, na humilhação pública, ou mesmo restrição à direito fundamental, dentre todos, o direito à igualdade – superado tolerável discrímen –, pela vedação de acesso a determinado bem ou serviços. Da mesma forma, os sistemas de saúde podem usar algoritmos para tomar decisões sobre diagnóstico e tratamento, mas os pacientes podem não saber como essas decisões são tomadas (O'Neil, 2020).

Cabe, portanto, a partir do paradigma da atual fenomenologia, buscar aliar conhecimento práticos do campo técnico e ético – da dimensão pragmática – e o campo das possibilidades jurídicas, a constituir verdadeiro elemento de possibilidade de integração entre direitos fundamentais na cláusula da realizabilidade na extensão das possibilidades fáticas e jurídicas – algo a importar quando da colisão de direitos fundamentais, objeto de trabalho, dentre outros, presente em Chala e Paulo (2022, p. 129-149).

3. Construindo um modelo de enfrentamento à discriminação algorítmica: Ações e práticas e exigências

No capítulo anterior, a complexidade e os efeitos prejudiciais da discriminação algorítmica foram explorados, destacando como os algoritmos podem inadvertidamente perpetuar estereótipos e preconceitos. A escolha do conjunto de técnicas a seguir não é de todo arbitrária, mas baseia-se na revisão da literatura especializada, que tem contribuído significativamente para a identificação de abordagens eficazes.

Há, portanto, como se viu, uma necessidade premente de enfrentar a discriminação algorítmica por múltiplos vieses. Entre essas abordagens, destaca-se a importância da transparência dos algoritmos, conhecida como Inteligência Artificial Explicável (XAI, no acrônimo em inglês). A compreensão de como os algoritmos tomam decisões é fundamental para identificar e corrigir potenciais vieses. No campo da transparência há um campo fértil de conceitos e preocupações a serem analisadas.

Outro ponto de destaque, é o da promoção e incentivo para com a diversidade nas equipes de desenvolvimento das aplicações algorítmicas, sobretudo, além da programação, na fase de tratamento dos dados e de treinamento da IA com tais bases de dados. tem se destacado como uma prática essencial, pois contribui para uma visão mais inclusiva e democrática durante todo o processo de implementação de algoritmos, reduzindo a probabilidade de reprodução de preconceitos. A diversidade nas equipes de desenvolvimento é prática amplamente recomendada e na qual se apostam esforços em torno de implementação via políticas públicas e, também, conscientização.

Outra técnica relevante é a análise contínua dos dados de treinamento, permitindo a identificação e correção de viés logo no início do processo. O uso de feedbacks de linguagem tem se mostrado promissor na mitigação de discriminações linguísticas. Além disso, o desenvolvimento de algoritmos conscientes de equidade, que buscam ativamente evitar resultados discriminatórios – pois instados e programados para tanto –, já se afiguram como possíveis, ainda que não plenamente treináveis.

Por fim, a criação de regulamentações e padrões de responsabilidade para desenvolvedores e empresas que utilizam algoritmos também tem um papel crucial na redução da discriminação algorítmica, pelo natural medo humano, ainda que economicamente motivado, da sanção.

3.1. Inteligência Artificial Explicável (XAI)

Acerca da Inteligência Artificial Explicável muito há a dizer para o pouco fôlego aqui permitido. O ponto já foi destacado em apartado em Paulo, 2023. O que cabe dizer para a introdução do tema por aqui é: a importância da XAI é de dupla natureza. A XAI desempenha um papel fundamental na construção (*by design*) de algoritmos mais transparentes e justos, que permitem a rastreabilidade de suas decisões automatizadas, permitindo que desenvolvedores e

demais agentes com aptidão técnica entendam como suas decisões foram tomadas, mas também ajuda a identificar e corrigir possíveis vieses discriminatórios.

A XAI também se faz essencial para estabelecer a confiança do público na tecnologia, ao mesmo tempo que impõe um ônus pesado – mas vencível – ao desenvolvedor. À medida que a IA desempenha um papel cada vez mais proeminente em nossa sociedade, a capacidade de explicar e compreender seu funcionamento se torna uma exigência tanto do ponto de vista ético quanto regulatório, colocando mais uma garantia acerca de que a tecnologia utilizada – aprovada nos demais testes – atue de forma responsável e equitativa, e, de igual sorte, isso decorrerá pelo próprio desenho da aplicação.

Por fim, a XAI está presente enquanto uma possibilidade técnica do cenário tecnológico, mas também é um excelente ponto de contato com a dinâmica jurídica, visto que sua existência e alocação enquanto direito individual ou imposição à desenvolvedora – algo a ser construído se apenas um dos dois ou ambos – funda-se no dever de transparência, cuja dimensão individual aos credores jusfundamentais é de direito fundamental e, portanto, reclamável.

3.2. Diversidade nas Equipes de Desenvolvimento

O uso de conjuntos de dados diversificados e representativos durante o treinamento dos algoritmos pode ajudar a reduzir o viés. Técnicas de aprendizado de máquina justas e transparentes podem ser utilizadas para garantir que os algoritmos tomem decisões justas e explicáveis. No campo do direito, a criação de leis e regulamentos específicos para lidar com a discriminação algorítmica é crucial, mas a partir de um olhar mais tocado à realidade do dia a dia do desenvolvimento e envolvimento com a progressão dessas tecnologias, não há de ser ignorado o fato de que a medida de presença representativa mínima e inclusiva nas equipes de desenvolvimento de aplicações pode, se não reduzir drasticamente o problema, indicar por uma boa-fé da equipe e da empresa em buscar se adequar no campo do dever de respeito à não-discriminação – direito já consagrado na normatividade pátria, e que virá reforçado quando da aprovação do PL sobre o uso da Inteligência Artificial (nº 2.338/2023), atualmente previsto no Art. 3º, IV, do PL.

A diversidade nas equipes de desenvolvimento desempenha um papel crucial na mitigação da discriminação algorítmica. A programação de algoritmos muitas vezes reflete as perspectivas e experiências de quem os desenvolve, e equipes diversas são essenciais para

garantir que essas perspectivas sejam inclusivas e representativas. A falta de diversidade pode levar a preconceitos e discriminações não intencionais nos algoritmos, uma vez que as decisões acerca da programação podem ser influenciadas pelos vieses dos desenvolvedores que, não por maldade, mas por desconhecimento fático (Hao, 2019), pode passar ao largo das microagressões que possam vir a redundar em uma violência automatizada.

Sob um viés individual do programador, no entanto, deve ser atacado o fato de que programadores indiferentes com questões sociais, por exemplo, a questão racial – ou mesmo programadores (in)conscientemente preconceituosos ou, no exemplo, racistas – não se importarão com uma base de dados pouco diversificada (O'Neil, 2020, p. 37). O desapego para com a diversidade social representada na base de dados é uma das raízes de nível mais próximo à superfície do mal da discriminação algorítmica.

Por isso, é crucial que os programadores levem em consideração a diversidade social e as questões de discriminação algorítmica durante o desenvolvimento de sistemas e algoritmos, tanto em sua programação quanto na organização de sua base de dados. É preciso tomar medidas para mitigar esses vieses e garantir que as soluções tecnológicas sejam justas e equitativas, mesmo que para tanto se exija mais do que meramente a não discriminação incidental; é necessário o combate ativo à discriminação. Para tanto, é preciso um esforço consciente para incorporar às etapas do processo de desenvolvimento a diversidade enquanto conceito a liderar e subsidiar todas as etapas. A questão é: como fazer isto?

Promover a diversidade nas equipes de desenvolvimento não é apenas uma questão ética, mas também uma estratégia eficaz na luta contra a discriminação algorítmica. Essa diversidade não se limita apenas à representação de gênero e etnia, mas também à inclusão de pessoas com diferentes formações acadêmicas, idades e experiências de vida. Equipes diversas estão mais bem posicionadas para avaliar e aprimorar a equidade dos algoritmos, reduzindo assim o impacto de preconceitos indesejados nos resultados. A inclusão de uma variedade de perspectivas no processo de desenvolvimento de algoritmos é fundamental para garantir que as tecnologias reflitam de maneira mais justa a sociedade como um todo e contribuam para a construção de um ambiente mais igualitário, que atualmente em 2019 contava com apenas 18% de mulheres na indústria da IA como um todo e 2,5% e 4% de negros, respectivamente no Google, e empatados Facebook e Microsoft (Hao, 2019).

Além disso, a diversidade nas equipes de desenvolvimento também promove a inovação e a criatividade. Diferentes perspectivas e experiências podem levar a soluções mais criativas e

eficazes na construção de algoritmos e na resolução de problemas complexos. Isso é especialmente importante em um campo tão dinâmico e desafiador quanto a tecnologia da informação, onde a capacidade de inovação pode fazer a diferença entre o sucesso e o fracasso.

3.3. *Feedbacks* de Linguagem:

Os modelos de linguagem são ferramentas poderosas para gerar e compreender textos, mas também podem reproduzir vieses e estereótipos presentes nos dados que os treinam. No entanto, um estudo recente do laboratório de inteligência artificial Anthropic mostrou que esses modelos podem se autocorrigir para alguns desses vieses, se receberem instruções simples em linguagem natural para isso (Firth, 2023).

Os modelos de linguagem (ML) são sistemas computacionais que aprendem a gerar ou processar texto a partir de grandes quantidades de dados linguísticos. Nos últimos anos, os MLs baseados em redes neurais profundas, como BERT, GPT-2, GPT-3, GPT-3.5, GPT-4, BARD, Chat do Bing, têm demonstrado avanços impressionantes em diversas tarefas de processamento de linguagem natural (PLN), como compreensão de texto, geração de texto e resposta a perguntas. No entanto, esses MLs também sofrem de problemas sérios relacionados aos vieses presentes nos dados de treinamento, que podem refletir estereótipos e preconceitos da sociedade sobre gênero, raça, religião e outros aspectos sociais. Esses vieses podem se manifestar de forma indesejável e prejudicial no comportamento dos MLs, gerando textos ofensivos, discriminatórios ou tóxicos.

Firth (2023) apresenta uma nova abordagem para capturar geometricamente os vieses morais e éticos dos MLs, usando uma "direção moral" que pode ser computada no espaço de incorporação dos MLs. Essa "direção moral" pode avaliar a normatividade (ou não-normatividade) de frases arbitrárias sem treinar explicitamente o ML para essa tarefa, refletindo bem as normas sociais. Além disso, Firth (2023) demonstra que o uso da "direção moral" pode fornecer um caminho para atenuar ou até prevenir a degeneração tóxica nos MLs, mostrando essa capacidade no conjunto de testes RealToxicityPrompts.

Firth (2023) indica ainda que os pesquisadores Amanda Askell e Deep Ganguli analisaram modelos de linguagem de diferentes tamanhos, que haviam passado por diferentes quantidades de treinamento com reforço a partir de feedback humano - *reinforcement learning from human feedback* (RLHF) –, uma técnica que faz com que humanos orientem o modelo de

IA para produzir respostas mais desejáveis. Eles testaram esses modelos com três conjuntos de dados que foram projetados para medir o viés ou a estereotipação em relação a idade, raça, gênero e outras categorias. Eles descobriram que apenas solicitar ao modelo que garantisse que suas respostas não dependessem de estereótipos tinha um efeito positivo dramático em sua saída (*output)*, especialmente naqueles que haviam completado rodadas suficientes de RLHF e tinham mais de 22 bilhões de parâmetros, as variáveis em um sistema de IA que são ajustadas durante o treinamento. Em alguns casos, o modelo até começou a se envolver em discriminação positiva em sua saída. Esse resultado é notável porque mostra que os modelos de linguagem podem aprender a evitar vieses sem precisar de uma definição formal ou explícita do que é viés. Isso sugere que eles podem captar pistas implícitas ou contextuais sobre o que é considerado aceitável ou inaceitável em diferentes situações. No entanto, os pesquisadores não sabem exatamente como os modelos são capazes de fazer isso, embora tenham algumas hipóteses.

Uma das hipóteses é a de que à medida que os modelos ficam maiores, eles também têm conjuntos de dados de treinamento maiores, e nesses conjuntos de dados há muitos exemplos de comportamento tendencioso ou estereotipado. Outra é que o feedback humano pode fornecer uma orientação valiosa para os modelos sobre como se comportar de forma mais ética ou socialmente responsável. O estudo tem implicações importantes para o desenvolvimento e o uso de modelos de linguagem em diversas aplicações, como assistentes virtuais, chatbots, sistemas educacionais e jornalismo automatizado. Ele sugere que esses modelos podem ser treinados para produzir conteúdo menos tóxico ou ofensivo, simplesmente pedindo-lhes para fazê-lo (Firth, 2023), ao contrário do que defende, dentre outros O'Neil (2020, p. 15), para quem os *feedbacks* sobre aplicações ruins as pioram ou as inutilizam. No entanto, isso também levanta questões sobre quem deve definir o que é viés e como garantir que os modelos não sejam manipulados para produzir conteúdo enviesado intencionalmente (Firth, 2023).

3.4. Regulamentações e Padrões de Responsabilidade:

A responsabilidade dos desenvolvedores de inteligência artificial (IA) é um tema complexo e em constante evolução. Ainda não há uma regulamentação específica para a responsabilidade civil por danos causados por IA no Brasil, mas a Lei Geral de Proteção de Dados (LGPD) estabelece princípios e diretrizes que podem ser aplicados a essa questão. Hoje, os princípios da segurança e da prevenção da LGPD, aliados à competência regulamentar da

Autoridade Nacional de Proteção de Dados (ANPD), permitem que sejam fixados padrões de mínima diligência capazes de estabilizar expectativas e possibilitar a responsabilização nos marcos do art. 42 da LGPD, para o mau uso ou conservação dos dados, como no caso de eles serem vendidos, vazados ou usados especificamente para um final danoso, não abarcando a matriz de risco sobre uso das Inteligências Artificiais. Além disso, as pessoas desenvolvedoras devem garantir que os sistemas de IA sejam seguros e resistentes a ataques cibernéticos, garantindo que não sejam manipulados ou hackeados por terceiros que podem manipular o funcionamento do sistema, como se vê, distante do que se é necessário a partir da IA e de algoritmos.

Outra abordagem possível é baseada no risco, que modula o conteúdo das normas de acordo com a intensidade dos riscos criados pelos sistemas de IA, previsões trazidas aos Arts. 13 a 18 e 22 a 26 do Projeto de Lei nº 2.338/2023. A realização de avaliações de impacto algorítmico (AIA) serve para categorizar o risco em Risco Excessivo – cuja implementação são vedadas –, Alto Risco e as que não apresentam riscos classificadas nas duas categorias anteriores. A AIA é composta por várias etapas que visam avaliar e mitigar, preventivamente, potenciais impactos negativos decorrentes da exploração de atividades de risco. O PL também prevê a criação de um órgão regulador específico para a IA, que terá como função principal fiscalizar o uso desses sistemas e garantir que eles sejam desenvolvidos e utilizados em conformidade com a regulação, a ANPD pode vir a, eventualmente, assumir esse papel.

Acerca da responsabilidade, o PL nº 2.338/2023 traz em seu texto original, no Art. 5º, inciso V, a previsão do direito à não-discriminação e à correção de vieses discriminatórios diretos, indiretos, ilegais ou abusivos. E traz um regramento maior no Art. 12, prevendo que as pessoas afetadas, inclusive em **i)** decorrência do uso de dados pessoais sensíveis ou de impactos desproporcionais em razão de características pessoais como origem geográfica, raça, cor ou etnia, gênero, orientação sexual, classe socioeconômica, idade, deficiência, religião ou opiniões políticas; **ii)** função do estabelecimento de desvantagens ou agravamento da situação de vulnerabilidade de pessoas pertencentes a um grupo específico, ainda que se utilizem critérios aparentemente neutros; terá, salvo se demonstrado pela empresa que a "diferenciação realizada se dê em função de objetivos ou justificativas demonstradas, razoáveis e legítimas à luz do direito à igualdade e dos demais direitos fundamentais" (previsão do Parágrafo único do Art. 12), incidente a matriz de responsabilidade civil do Art. 27 e ss., sendo seu direito ver reparados os danos ocasionados, seja de natureza patrimonial, moral, individual ou coletiva.

Para a aferição da natureza quanto à responsabilidade civil, os §§1º e 2º do Art. 27 disciplinarão que ao se tratar de inteligência artificial de alto risco ou de risco excessivo, o fornecedor ou operador respondem objetivamente pelos danos causados, na medida de sua participação no dano, e quando não se tratar de aplicação de alto risco, a culpa do agente causador do dano será presumida, aplicando-se a inversão do ônus da prova em favor da vítima. Desse modo, percebe-se que os desenvolvedores têm um dever de cautela que decorre da perspectiva mais econômica possível, a produção antecipada de provas – seja pela XAI ou por outro meio – para conseguir desabonar-se de eventuais lides que venha à integrar.

Em resumo, o PL 2338/2023 busca estabelecer um marco regulatório para a IA no Brasil, com o objetivo de promover uma aplicação segura e responsável desses sistemas, seja pelo controle prévio, com a AIA, seja pelo controle administrativo, pela ANPD ou nova autoridade/agência, ou então pela via da reparação via poder judiciário. Não é de se olvidar, também, que mesmo após aprovado o Marco Regulatório da IA, o poder judiciário, sobretudo na figura do Supremo Tribunal Federal, deixará de ser atuante na definição da interpretação e aplicação do direito e, sobretudo, da nova lei a partir de uma leitura constitucional.

4. Conclusão

A discriminação algorítmica, um desafio complexo que permeia todos os aspectos da vida contemporânea, requerendo uma abordagem multidisciplinar para combater eficazmente seus impactos prejudiciais. Como explorado ao longo do estudo, os algoritmos e a inteligência artificial desempenham papéis cada vez mais central na sociedade, e ainda está desacompanhando de um estrito modelo regulatório, a ser exigido pelo poder judiciário. No entanto, a adoção indiscriminada dessas tecnologias sem as devidas salvaguardas pode resultar na perpetuação de estereótipos, preconceitos e injustiças, resultando em violência algorítmica, institucionalmente tolerada.

Nesse contexto, a criação de regulamentações específicas para o uso de algoritmos em decisões importantes e o estabelecimento de critérios claros para o desenvolvimento desses sistemas são medidas essenciais para garantir a justiciabilidade dessa nova realidade. Afinal, é fundamental considerar as dimensões humanas em que a privacidade e o mal ou mau processamento de dados pode ocasionar, daí a importância da tutela à autodeterminação informativa, como escudo e controle pelos indivíduos perante as decisões automatizadas que se

valem de seus dados.

No âmbito de toque da tecnologia com o direito se estabeleceu 4 vetores à construção de um modelo regulatório desejável. Como pilares de um templo grego, cada uma dessas estruturas deve significar em igual solidez, ainda que algumas sejam mais complexas do ponto de vista social, outros do ponto de vista da vontade do setores, ou ainda diante da complexidade técnica em si – como no caso de construção de modelos de XAI –, advoga-se que as respostas estatais começam por esses 4 pilares apresentados, e que podem, alguns deles, ser desde logo, mesmo antes de implementado aprovada qualquer legislação, aplicados pela poder judiciário na medida de sua já existente implícita na ordem jurídica – como é o caso da XAI e da responsabilidade civil (ainda que sem o panorama de classificações de riscos pela AIA).

Em resumo, a corrida pelo tratamento jurídico adequado da discriminação algorítmica requer uma abordagem multifacetada que abrange pelo menos tecnologia, ética e direito. A compreensão da natureza complexa desse fenômeno e a implementação de medidas proativas são essenciais para garantir que a tecnologia sirva ao progresso e à justiça, em vez de perpetuar desigualdades, desconexões e injustiças. A construção de um modelo de enfrentamento à discriminação algorítmica requer esforços contínuos e colaborativos, com a esperança de promover um ambiente mais igualitário e inclusivo na nova era da sociedade que começa.

Aliás, o direito é também uma tecnologia. Uma técnica nada inovadora, mas bem inovativa e inventiva (*Ius est Ars inveniendi*) de solucionar conflitos, justamente a partir da comunhão semântica à esfera do jurídico – toque de Midas. No que, aliás, resguarda a essência e virtude do direito: a capacidade de buscar na sociedade e no sistema da normatividade elementos extrajurídicos e racionalmente aplicá-los em busca de um justo para o caso.

REFERÊNCIAS

CARBONELL, Miguel. Constitucionalismo, minorías y derecho. **Isonomía Revista de Teoría y Filosofía del Derecho**, n. 12, v. 98, 2000.

CHALA, Bárbara Guerra; PAULO, Lucas Moreschi. Limites da liberdade de expressão à luz da proporcionalidade: o Inquérito das fake news. In: GAVIÃO FILHO, Anizio Pires; PAULO, Lucas Moreschi. (Org.). **Constitucionalismo, direitos fundamentais, proporcionalidade e argumentação.** São Paulo: Dialética, 2022, p. 129-149.

FIRTH, Niall. Language models might be able to self-correct biases. **MIT Technology Review.** Mar. 2023. Disponível em

<https://www.technologyreview.com/2023/03/20/1070067/language-models-may-be-able-to-self-correct-biases-if-you-ask-them-to/>. Acesso em 24 out. 2023.

HAO, Karen. AI's white guy problem isn't going away. **MIT Technology Review.** Artificial Intelligence. abr. 2019. Disponível em: < https://www.technologyreview.com/2019/04/17/136072/ais-white-guy-problem-isnt-going-away/>. Acesso em 23 set. 2023.

HEIKKILÄ, Melissa. These new tools let you see for yourself how biased AI image models are. **MIT Technology Review.** Artificial Intelligence. mar. 2023a. Disponível em: <https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/>. Acesso em 23 mar. 2023.

HEIKKILÄ, Melissa. Why it'll be hard to tell if AI ever becomes conscious**. MIT Technology Review.** Artificial Intelligence. out. 2023b. Disponível em: < https://www.technologyreview.com/2023/10/17/1081818/why-itll-be-hard-to-tell-if-ai-ever-becomes-conscious/>. Acesso em 29 out. 2023.

HUCKINS, Grace. Minds of machines: The great AI consciousness conundrum. **MIT Technology Review.** Artificial Intelligence. out. 2023. Disponível em: < https://www.technologyreview.com/2023/10/16/1081149/ai-consciousness-conundrum/>. Acesso em 29 out. 2023.

O'NEIL, Cathy. **Algoritmos de destruição em massa**: como o big data aumenta a desigualdade e ameaça a democracia. Trad. Rafael Abraham. Santo André: Editora Rua do Sabão, 2020.

PASQUALE, Frank. **The black box society**: the secret algorithms that control money and information. Cambridge e Londres: Harvard University Press, 2015.

PAULO, Lucas Moreschi. Opacidade dos algoritmos e a necessidade de transparência: Garantindo explicabilidade. In: **Anais do XIX Seminário Internacional Demandas Sociais e Políticas Públicas na Sociedade Contemporânea e XV Mostra Internacional de Trabalhos Científicos**. Santa Cruz do Sul: UNISC, 2023. Disponível em: <https://online.unisc.br/acadnet/anais/index.php/sidspp/article/view/23632>. Acesso em 30 out. 2023.

STEINER, Christopher. **Automate This: how algorithms came to rule our world**. New York: Penguin Group, 2012.

TURK, Victoria. How AI reduces the world to stereotypes. **Rest of World Journal.** The rise of AI. Out. 2023. Disponível em: < https://restofworld.org/2023/ai-image-stereotypes/>. Acesso em 29 out. 2023.