Inteligência artificial, vieses algorítmicos e racismo: o lado desconhecido da justiça algorítmica

Recebido: 10 de janeiro de 2022 • Aprovado: 6 de julho de 2023
https://doi.org/10.22395/ojum.v23n50a49

Alexandre Morais da Rosa
Universidade do Vale do Itajaí (UNIVALI), Itajaí, Brasil
alexandremoraisdarosa@gmail.com
https://orcid.org/0000-0002-3468-3335

Bárbara Guasque
Universidade do Vale do Itajaí (UNIVALI), Itajaí, Brasil
barbara@guasque.adv.br
https://orcid.org/0000-0003-0633-8363

Resumo

O presente artigo se propõe a identificar algumas externalidades negativas oriundas da não observância de determinados padrões éticos em modelos de Inteligência Artificial (IA). O estudo objetiva esclarecer a importância de se voltar uma rigorosa atenção aos dados que são utilizados na construção de modelos de IA, tal como elencar possíveis soluções para reduzir a incidência de algoritmos enviesados e mitigar suas consequências danosas. A metodologia utilizada tem natureza exploratória e descritiva, abordando casos práticos e, também, como procedimento técnico, utilizou-se de pesquisa bibliográfica. A conclusão principal aferida é a de que os algoritmos enviesados produzem nefastas consequências sociais, violando direitos fundamentais e operando como catalisadores, o que aumenta e perpetua preconceitos e segregações inerentes à sociedade na qual se baseiam, contribuindo com a manutenção e intensificação do racismo estrutural que permeia a sociedade e o sistema de justiça criminal.

P*alavras-chave*: inteligência artificial; vieses algorítmicos; sistema de justiça; racismo; discriminação; direitos fundamentais.

Artificial Intelligence, Algorithm Biases and Racisms: The Dark Side of Algorithm Justice

Abstract

This article aims to identify some negative externalities arising from the failure to comply with specific ethical standards in Artificial Intelligence (AI) models. This study highlights the importance of paying rigorous attention to the data used in building AI models, such as listing potential solutions to reduce the incidence of skewed algorithms and mitigate their harmful consequences. This article followed an exploratory descriptive methodology, addressing practical cases, and turned to bibliographic review as a technical procedure. The main finding is that skewed algorithms cause disastrous social consequences, violating fundamental rights and actings as catalyzers, increasing and perpetuating prejudice and segregation inherent to their society, thus contributing to the structural racism that permeates society and the criminal justice system.

K*eywords*: artificial intelligence; algorithm skews; justice system; racism; discrimination; fundamental rights.

Inteligencia artificial, sesgos algorítmicos y racismo: el lado desconocido de la justicia algorítmica

Resumen

El presente artículo se propone identificar algunas externalidades negativas originadas por la inobservancia de determinados patrones éticos en modelos de Inteligencia Artificial (IA). El estudio pretende aclarar la importancia de prestar una atención rigurosa a los datos utilizados en la construcción de modelos de IA, así como enumerar posibles soluciones para reducir la incidencia de algoritmos sesgados y mitigar sus consecuencias perjudiciales. La metodología empleada es de naturaleza exploratoria y descriptiva, abordando casos prácticos, y también se utilizó la investigación bibliográfica como procedimiento técnico. La principal conclusión obtenida es que los algoritmos sesgados producen nefastas consecuencias sociales, vulnerando derechos fundamentales y operando como catalizadores, que aumentan y perpetúan prejuicios y segregaciones inherentes a la sociedad en la que se basan, contribuyendo al mantenimiento e intensificación del racismo estructural que permea a la sociedad y el sistema de justicia penal.

P*alabras clave*: inteligencia artificial; sesgos algorítmicos; sistema de justicia; racismo; discriminación; derechos fundamentales.

Introdução

O presente artigo é fruto da pesquisa acadêmica dos autores que são líderes e pesquisadores do grupo de pesquisa SpinLawLab, vinculado ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) do Ministério da Ciência, Tecnologia e Inovações para incentivo à pesquisa no Brasil. Mencionado grupo de pesquisa visa proporcionar instrumentos para a constatação de diversas perspectivas do fenômeno jurídico, sob a óptica da inteligência artificial e demais inovações tecnológicas. Por fim, cumpre registrar que este estudo faz parte das pesquisas empreendidas durante o estágio de pós-doutoramento da pesquisadora Bárbara Guasque, com a supervisão do pesquisador Alexandre Morais da Rosa (UNIVALI), e contou com o fomento do CNPq mediante a bolsa de pós-doutorado júnior.

Permeando todas as esferas da vida em sociedade, a utilização de inteligência artificial (IA) está presente também no Poder Judiciário. Até junho de 2020, já existiam 64 projetos de IA em 47 tribunais brasileiros, em diferentes fases de implementação (Fundação Getulio Vargas, 2020, p. 26). As novas tecnologias estão apontando no Judiciário brasileiro com o fito de automatizar atividades repetitivas, classificatórias e organizacionais, além de servir de apoio à tomada de decisão judicial, o que incrementa a produtividade e angaria eficiência e redução da morosidade processual e da insegurança jurídica — os grandes gargalos do Poder Judiciário nacional.

As funcionalidades são muitas e o rol se amplia rapidamente. A inserção da disrupção nos tribunais brasileiros constitui uma grande aliada no aprimoramento do ambiente institucional judicial nacional e traz a esperança de uma justiça mais célere, efetiva e estável.

Um Poder Judiciário asfixiado por um acervo de 79,7 milhões de processos em tramitação, e com um custo aproximado de 91 bilhões de reais com servidores (Conselho Nacional de Justiça, 2020), pode vislumbrar na tecnologia e na disrupção uma esperança para tentar reverter o panorama de excessiva morosidade, alto custo e insegurança jurídica.

Em geral, todos os sistemas atuais desenvolvidos nos tribunais brasileiros estão efetuando tarefas simples e repetitivas, nas quais não se enquadra o uso problemático dos algoritmos e seus possíveis vieses. Todavia, como a relação entre os cidadãos, o Poder Judiciário e as tecnologias disruptivas ainda é algo recente e crescente, existem algumas questões relevantes que necessitam ser abordadas. É o caso dos vieses algorítmicos que levantam preocupantes questionamentos de natureza ética.

À vista disso, o presente artigo tem natureza exploratória e descritiva, e se justifica para identificar e descrever os impactos negativos que a inteligência artificial e os algoritmos enviesados podem produzir na sociedade e no sistema de justiça, quando não bem desenvolvida e fiscalizada. Este estudo também se propõe a elencar

alternativas hábeis para atenuar a possibilidade de enviesamento dos modelos, mitigando as externalidades negativas oriundas de algoritmos enviesados, para que não venham a contribuir com a manutenção e intensificação do racismo estrutural que permeia a sociedade e o sistema de justiça criminal.

1. Inteligência Artificial e vieses algorítmicos

Consoante Navarro (2017), a IA é um campo da ciência e da engenharia, encarregado de compreender o comportamento inteligente do cérebro humano e, além disso, criar artefatos que simulem dito comportamento de maneira automatizada. Para a autora, a IA se ocupa de "*emular las diversas capacidades del cerebro humano para presentar comportamientos inteligentes sintetizando y automatizando tareas intelectuales*" (Navarro, 2017, p. 24).

Isso quer dizer que as máquinas, em alguma medida, imitam o processo cognitivo humano, após um desenvolvimento de aprendizado baseado em dados que fornecem generalizações sobre dado assunto. Porém, a IA alcançará sempre resultados superiores aos que poderia conseguir qualquer ser humano porque "*el sistema planteará alternativas que ni siquiera habíamos pensado previamente al no poder abarcar todos los datos un cerebro humano*". E "*a mayor cantidad de datos, más posibilidades de relacionarlos y obtener, por tanto, mejores resultados*" (Nieva Fenoll, 2018, p. 15).

A palavra-chave na IA é "algoritmo" (Nieva Fenoll, 2018, p. 15). Isso porque, "*un sistema de* IA *necesita de una secuencia de instrucciones que especifique las diferentes acciones que debe ejecutar el computador para resolver un determinado problema*". Esse esquema executivo que contempla as instruções, o caminho a ser percorrido, é desempenhado pela estrutura algorítmica (Navas, 2017, p. 24).

Um algoritmo é, portanto, um esquema executivo, uma sequência de ações para resolver um problema ou responder uma questão. Assim, "*una receta de cocina es un algoritmo para humanos*". Uma sequência de instruções para resolver o problema de como fazer determinada receita (Rodríguez, 2018, p. 109).

Todavia, computadores e máquinas precisam de algoritmos mais complexos, que os auxiliem a efetuar as tarefas definidas e que sempre produzam o mesmo resultado, com base nos mesmos parâmetros. Nesse sentido, encontra-se um subconjunto específico de algoritmos usados para aprendizado de máquina (Rodríguez, 2018, pp. 109-111).

A particularidade desses algoritmos é que eles aprendem por conta própria, fazendo inferências a partir dos dados. Portanto, o aprendizado de máquina é a capacidade destas de aprender com os dados, identificando tendências e padrões em eventos aparentemente aleatórios. "E*stos patrones, si se basan en series de datos lo suficientemente largas y en datos de suficiente calidad, pueden usarse para predecir el futuro*" (Rodríguez, 2018, p. 105). Dessa maneira, o aprendizado de máquina olha para o passado para encontrar padrões e tentar prever o futuro. "E*n esencia, el aprendizaje automático es pura predicción*"

(Rodríguez, 2018, p. 105). Cabe salientar que a expressão "prever o futuro" se refere a probabilidades, pois a IA efetua, tão somente, cálculos matemáticos e probabilísticos de que algo aconteça no futuro com base em dados do passado.

Esses algoritmos vêm sendo utilizados em diversas tarefas dentro do Poder Judiciário. Um exemplo disso são os sistemas Victoria, do Tribunal de Justiça do Rio de Janeiro, e Elis, do Tribunal de Justiça de Pernambuco, que automatizaram o rito das execuções fiscais. Também o sistema Mandamus, que utiliza de técnicas de inteligência artificial para auxiliar na automação do processo de elaboração, distribuição e gerenciamento do cumprimento de mandados judiciais. Ainda, o mais conhecido deles, o Victor, é utilizado no Supremo Tribunal Federal com o objetivo de otimizar a análise da Repercussão Geral (Da Rosa & Guasque, 2021).

Em que pesem os modelos[1] de IA desenvolvidos nos Tribunais brasileiros estarem servindo para ajudar na automação de atividades repetitivas e classificatórias, além de ser um instrumento de apoio à tomada de decisão judicial, é importante destacar os efeitos nocivos que o aprendizado de máquina, alimentado com *big datas* portadores de preconceitos e discriminações, pode gerar na sociedade e no sistema de justiça, catalisando injustiças sociais com a falsa aparência de neutralidade matemática, o que gera, conforme preleciona, O'Neil, "ciclos destrutivos de *feedback*" (O'Neil, 2020, p. 22).

Os vieses algorítmicos ocorrem quando o algoritmo adquire e reflete os valores humanos, ou seja, ele incorpora os mesmos desvios culturais que estão implícitos nos dados que são utilizados para treinamento do modelo. Isso acaba enviesando o resultado final que é obtido.

Panch et al. (2019) definem o viés algorítmico como a aplicação de um algoritmo que amplifica as desigualdades existentes em *status* socioeconômico, raça, origem étnica, religião, gênero, deficiência ou orientação sexual. Consoante os autores, a utilização dos algoritmos não somente reflete as desigualdades sociais pré-existentes, mas pode, em última análise, agravá-las. Segundo os pesquisadores, "se o mundo tiver uma determinada aparência, isso se refletirá nos dados, seja diretamente ou por meio de proxies, e, portanto, nas decisões algorítmicas." (Panch et al., 2019).

Isso quer dizer que os modelos algorítmicos, como elucida O'Neil, apesar de desfrutarem de uma reputação de imparcialidade, os algoritmos são "*opinions embedded in math*" (O'Neil, 2019). Eles expressam os objetivos e as ideologias de seus criadores e "seus pontos cegos, refletem as prioridades e o julgamento de quem os alimentou". Dependendo de quem construir esses modelos, quais variáveis levar em conta e com quais dados os alimentar, o resultado será um ou outro (O'Neil, 2020, pp. 33-35).

[1] Um modelo nada mais é do que a representação abstrata de algum processo. Esteja ele rodando dentro de um computador ou na nossa cabeça, o modelo absorve o que sabemos e usa isso para prever respostas em situações variadas (O'Neil, 2020, pp. 30-31).

Logo, algoritmos não são neutros, não são puramente matemáticos. Por outro lado, o algoritmo não cria o preconceito, somente repete um padrão contemplado nos dados utilizados para o treinamento do modelo. Como um algoritmo comporta uma sequência de regras para se chegar a um resultado, e embutido nessa sequência há um desvio, teremos um problema no resultado final. m geral, isso não surge dessa sequência de regras, que costuma ser objetiva, mas dos dados que reproduzem padrões e preconceitos sociais.

Isso acontece porque os modelos algorítmicos são, em grande medida, estatísticos. Então, a princípio, o algoritmo não faz nenhum pré-julgamento, ele não estabelece nenhuma regra, tampouco realiza valoração. Ele somente aprende com os dados mediante análises estatísticas, ou seja, ele aponta o que acontece habitualmente. O algoritmo reflete uma dinâmica social predominante.

Portanto, o cerne da questão é impedir que as ferramentas tecnológicas reproduzam limitações, falhas e preconceitos presentes na sociedade e no sistema de justiça.

É imperioso, assim, identificar os impactos negativos que os modelos enviesados podem produzir na sociedade e no sistema de justiça se não lançarmos uma atenção rigorosa sobre os dados que serão utilizados para treinamento e validação dos modelos.

2. Algoritmos como a perpetuação e massificação de práticas sociais discriminatórias

Em 2018, a Amazon decidiu descartar um modelo que a empresa utilizava para fazer a seleção dos currículos para novas contratações. Ele apresentava um viés sexista que desprezava os currículos femininos, atribuindo uma pontuação alta apenas para currículos masculinos. A discriminação da ferramenta contra candidatas do sexo feminino acontecia porque os dados utilizados para treinamento do modelo foram currículos enviados para a empresa nos últimos 10 anos, os quais são, em sua imensa maioria, de homens, como acontece na maior parte da indústria de tecnologia (Dastin, 2018). Segundo o AI I*ndex*, relatório independente sobre o setor da tecnologia, os homens representam 71% dos candidatos a vagas de empregos na área da IA nos Estados Unidos (Paul, 2019). Diante desse contexto, o algoritmo passou a entender que os candidatos homens eram mais aptos para as vagas.

O'Neil explica que, como o aprendizado de máquina se baseia nos dados do passado, o algoritmo constatou, estatisticamente, que não somente os homens eram maioria, como eram promovidos com mais frequência e recebiam mais aumento salarial. Em contrapartida, as mulheres eram minoria, tendiam a sair rapidamente e recebiam menos aumento. Logo, o algoritmo entendeu que os homens eram as melhores contratações, perpetuando e ampliando, por meio do modelo algorítmico, um viés histórico, sexista e misógino (O'Neil, 2018).

Ainda, um estudo conduzido pelo N*ational* I*nstitute of* S*tandards and* T*echnology* (NIST), agência governamental norte-americana, testou 189 algoritmos de 99 desenvolvedores. A constatação foi que a tecnologia de reconhecimento facial é, para a correspondência um para um, entre 10 e 100 vezes menos eficaz para rostos asiáticos e afro-americanos, em comparação com as imagens de caucasianos. Para a correspondência um para muitos, a equipe observou taxas mais altas de falsos positivos para mulheres afro-americanas (NIST, 2019). Diferenças em falsos positivos, na correspondência um para muitos, são particularmente importantes, porque as consequências podem incluir acusações falsas — o que vem acontecendo com indesejável frequência (Hill, 2020).

Esses casos têm, como raízes e denominadores comuns, a predominância maciça de homens brancos na indústria da tecnologia e a consequente ausência de diversidade nas equipes desenvolvedoras. Uma pesquisa publicada pelo AI N*ow* I*nstitute*, instituto de pesquisa que estuda as implicações sociais da inteligência artificial, constatou que a ausência de diversidade dentro das equipes vem contribuindo para a criação de sistemas falhos e que perpetuam preconceitos de gênero e raça. Apenas 15% dos pesquisadores de IA do Facebook e 10% do Google são mulheres. As mulheres representam somente 18% dos graduados em ciência da computação nos Estados Unidos (Whittaker et al., 2018).

Com relação à população negra, essa estatística é ainda menor, correspondendo a apenas 2,5% dos funcionários do Google, 4% do Facebook e da Microsoft e 6% do Twitter. A força de trabalho da Apple contempla 9% de negros, todavia, essa estatística inclui funcionários de varejo. Assim como a Amazon, que tem 26,5% de funcionários negros; no entanto, a maioria ocupa empregos de baixa remuneração e apenas 8,3% deles estão em cargos de gerência (Whittaker et al., 2018).

A participação de negros na indústria da tecnologia ainda é inexpressiva. De acordo com a U.S. E*qual* E*mployment* O*pportunity* C*ommission*, no ano de 2014, apenas 8,6% dos graduados com bacharelado em ciência da computação e informática eram negros (U.S. Equal Employment Opportunity Commission, 2015). "Evidências sugerem que a IA, como um campo, é ainda menos diversificada do que a ciência da computação como um todo" (Whittaker et al., 2018).

As estatísticas revelam um setor dominado por homens brancos e asiáticos. Sendo que os demais grupos permanecem sub-representados na seara da tecnologia e da IA. A limitação de raça, gênero, áreas e contextos no desenvolvimento do modelo conduz a uma cegueira do algoritmo quanto aos elementos que não estão naquele enquadramento. Então, por exemplo, nos algoritmos de reconhecimento facial, como as equipes de desenvolvimento são majoritariamente compostas por homens brancos, existe uma incapacidade do algoritmo em identificar negros; mormente mulheres negras.

Isso ocorre porque não há componentes negros dentro da equipe, tampouco no banco de dados, que ajudem a treinar o modelo. Não há representatividade estatística.[2]

Conforme os cientistas da computação constroem e treinam os algoritmos, eles se concentram nas características faciais que são mais visíveis em determinada raça, mas não nas demais. Além disso, normalmente se baseiam em bibliotecas de códigos pré-existentes, que costumam ser escritas por outros pesquisadores brancos. O resultado final é voltado para focar em rostos caucasianos. Isso restou evidente no estudo do NIST (2019), que demonstrou que algoritmos de reconhecimento facial desenvolvidos por empresas asiáticas são muito mais precisos em rostos asiáticos, comprovando que as características de gênero e raça dos criadores dos modelos detêm forte influência sobre o seu resultado final.

Mesmo que o software seja desenvolvido para aprender com a experiência e tornar-se mais preciso com técnicas de aprendizado de máquina, os conjuntos de dados de treinamento são geralmente compostos por pessoas brancas. Portanto, o algoritmo especializa-se e aprimora-se somente na identificação de indivíduos brancos.

Isso pode acontecer com relação a raças e gêneros, mas também pode ocorrer com determinados contextos e classes sociais que, se não forem bem representados no treinamento do modelo, não terão uma representatividade estatística; formarão um provável ponto cego no modelo, capaz de carrear danosas consequências sociais e violar direitos fundamentais.

O software R*ekognition*, desenvolvido pela Amazon para identificar pessoas em vídeos e imagens, é utilizado por inúmeros órgãos policiais dos Estados Unidos. Contudo, esse software apresenta falhas graves no reconhecimento de pessoas como falso positivo. Um estudo efetuado pela União Americana pelas Liberdades Civis, em 2018, testou o sistema submetendo fotos dos congressistas norte-americanos. O sistema identificou 28 dos deputados e senadores como criminosos. Ainda, "as falsas correspondências eram desproporcionalmente de pessoas de cor". Cerca de 40% dos congressistas, erroneamente identificados como criminosos, eram negros, embora eles representassem apenas 20% dos membros do Congresso (Snow, 2018).

Ocorre que, por ser um sistema muito acessível, o R*ekognition* vem sendo facilmente adquirido e comercializado para segurança pública, o que deve exacerbar o falso reconhecimento positivo de pessoas negras e intensificar o racismo que permeia o nosso sistema penal.

Também é importante destacar os efeitos nocivos que o aprendizado de máquina, alimentado com *big datas* que retratam preconceitos e discriminações, podem gerar

[2] Representatividade estatística refere-se à "determinação da qualidade de uma amostra constituída de modo a corresponder à população no seio da qual ela é escolhida" (Freitas, 2020).

na sociedade e no sistema de justiça, catalisando injustiças sociais com a falsa aparência de neutralidade matemática (O'Neil, 2020, p. 22).

Um dos exemplos mais conhecido é o sistema Compas (*Correctional Offender Management Profiling for Alternative Sanctions*), utilizado nos EUA. O Perfil de Gerenciamento Corretivo de Infratores para Sanções Alternativas, em português, é uma ferramenta desenvolvida e comercializada pela Northpointe, foi elaborada com o intuito de prever a reincidência criminal de condenados, auxiliando os juízes na definição de quais programas de tratamento ou de condicional o réu teria, com base nas informações extraídas pelo sistema. Quando os réus são autuados na prisão, eles respondem a um questionário. Suas respostas são inseridas no Compas e geram uma pontuação, incluindo previsões de "risco de reincidência" e "risco de reincidência violenta" (Larson et al., 2016).

Assim, o algoritmo passou a ser utilizado não somente para negar liberdade condicional, com base na alta possibilidade de reincidência demonstrada na pontuação do algoritmo, como passou a servir de base para o aumento da pena, a partir desse mesmo critério.

O primeiro grande problema dessa circunstância é que a utilização desse algoritmo para justificar um aumento de pena solapa importantes garantias fundamentais, como a legalidade e o devido processo legal. Como bem delineou O'Neil, o questionário utilizado para avaliar o réu inclui informações sobre o seu local de moradia, seus familiares e amigos. São detalhes que passam ao largo do fato típico e das circunstâncias previstas legalmente para justificar um aumento de pena. Nessa conjuntura, o réu tem sua liberdade condicional negada, assim como a sua pena-base aumentada, não por circunstâncias legais ou por causa do crime praticado, mas devido ao seu local de moradia, suas relações de amizade, seus familiares etc. Por meio da pontuação determinada pelo algoritmo, os réus são julgados por serem quem são e pelo meio em que vivem, e não pela conduta praticada (O'Neil, 2020, p. 43).

Mas a teratologia vai além. Um estudo feito por Larson et al. (2016) analisou mais de 10 mil réus criminais no Condado de Broward, na Flórida. A pesquisa comparou as taxas de reincidência previstas pelo algoritmo com as taxas reais de reincidência dos réus durante dois anos após terem sido avaliados. Os resultados da análise demonstraram que os réus negros tinham 45% mais chances de receber pontuações de risco mais altas do que os réus brancos. "Os réus negros também tinham duas vezes mais probabilidade de serem classificados erroneamente com um risco maior de reincidência violenta" (Larson et al., 2016).

Ou seja, o algoritmo detinha um viés que considerava réus negros mais perigosos que os demais. Isso porque ele extraiu esse padrão dos dados que foram utilizados para treinar o sistema. No entanto, não existiam erros nos dados fornecidos;

eles estavam corretos. Apesar de constituírem somente 13% da população, os negros ocupam 40% das vagas em presídios nos Estados Unidos (O'Neil, 2020, p. 39). Diante disso, o modelo efetuou uma análise estatística, levando-o a adotar o padrão de que os negros eram mais perigosos.

Ao receberem pontuações mais altas do que os réus brancos, os réus negros não somente têm negada a liberdade condicional, mas também lhes são atribuídas condenações mais altas. Consoante a Associação Americana para Liberdades Civis (ACLU), as sentenças impostas a homens negros são cerca de 20% maiores do que aos condenados brancos por crimes similares (O'Neil, 2020, p. 40).

Em que pese o modelo não contemplar o aspecto "raça" no questionário, constavam entre as perguntas o local de moradia do réu. Esse fator, diante do contexto social e segregador dos Estados Unidos, era um forte indicativo da raça do acusado. Então, ainda que a raça não fosse um parâmetro utilizado no julgamento, as perguntas realizadas acabaram por extrair uma informação racial, e os dados estatísticos refletiram o racismo estrutural, prejudicando o resultado final do modelo, pois o torna racista. Dessa maneira, perpetua-se, com ares de cientificidade, o racismo que contamina o sistema de justiça criminal norte-americano.

Outro retrato pernicioso que estampa essa conjuntura se encontra na ferramenta R*ing*, da Amazon. Os dispositivos R*ing* consistem em campainhas inteligentes que são conectadas às câmeras de vigilância e que, tem por fito, combater o crime. Referido sistema é disponibilizado para ser utilizado em bairros residenciais, lançando mão da rede cada vez maior de câmeras de segurança, para alertar quando um indivíduo considerado suspeito for enquadrado por alguma câmera da região. No entanto, a Amazon estabeleceu parcerias secretas com departamentos de polícia dos Estados Unidos, munindo a polícia com "câmeras onipresentes" e um sistema de vigilância muito mais amplo e poderoso (Sadowski, 2020, p. 153). Algumas cidades criaram até incentivos pecuniários para colocar as câmeras R*ing* em bairros pobres e de alta criminalidade, permitindo aos departamentos de polícia amplo acesso às filmagens. Contudo, a utilização dessa ferramenta pelas forças policiais permite uma vigilância excessiva e invasiva em comunidades de baixa renda e de maioria negra, o que gerou um policiamento desproporcional que aumentou a violência policial nessas comunidades (Sadowski, 2020, p. 154).

Ainda, a ferramenta permite e incentiva que policiais, em conjunto com moradores dos bairros e condomínios, criem listas de pessoas indesejáveis que circulam pelas áreas residenciais nobres. "Permite que os moradores, de bairros de alta renda, façam julgamentos sobre pessoas que consideram desagradáveis, convocando a polícia para confrontá-los" (Sadowski, 2020, p. 154).

"O aplicativo N*eighbours*, do R*ing*, já firmou parceria com mais de 1.300 forças policiais nos Estados Unidos — um aumento de 300% em relação a agosto de 2019" (Biddle, 2019). E uma pesquisa efetuada pelo site Motherboard, que é a seção de tecnologia da VICE, revelou que pessoas negras são mais propensas a serem vigiadas pelo aplicativo N*eighbours* (Haskins, 2019).

Segundo Kelley e Guariglia, da E*lectronic* F*rontier* F*oundation*,[3] as campainhas de vigilância R*ing* e o aplicativo N*eighbours* têm papel protagonista em "permitir e perpetuar o assédio policial aos negros americanos". Essas ferramentas incentivaram "os piores instintos de muitos residentes", incitando-os a espionar pedestres, vizinhos e trabalhadores (Kelley & Guariglia, 2020).

Nessa mesma linha está o software PredPol, utilizado por departamentos de polícia para prever lugares onde crimes futuros podem acontecer. Programas como esse vêm sendo amplamente utilizados em departamentos de polícia que dispõem de reduzido orçamento nos Estados Unidos. O PredPol é baseado em software sísmico, ou seja, à medida que vê um crime acontecer em dado local, ele incorpora em padrões de histórico a fim de prever os locais onde novos crimes poderão acontecer futuramente. Diferentemente do Compas, esse tipo de software não se concentra no indivíduo, uma vez que leva em conta apenas a localização geográfica. No entanto, ao configurar o sistema PredPol, os departamentos de polícia inserem, como dados de treinamento, relatórios policiais de crimes anteriores. Estes contemplam majoritariamente crimes de perturbações (já que os policiais foram todos treinados na ortodoxia da tolerância zero). Esses tipos de notificações de práticas delitivas são endêmicos em bairros de baixa renda, o que alimenta o sistema com inúmeros casos e gera alta pontuação no PredPol. A consequência é que o sistema acaba prevendo como pontos críticos para cometimento de crimes sempre os mesmos lugares, mormente bairros de baixa renda e de maioria negra, angariando um ciclo contínuo de policiamento excessivo nessas áreas. Logo, o PredPol acaba por automatizar os preconceitos. E pior, traz ares de evidência matemática e a falsa percepção de legitimidade, pois é operado mediante algoritmos "imparciais" (O'Neil, 2020, pp. 135-137).

Esse enviesamento acontece porque os sistemas algorítmicos inferem qual será a próxima resposta com base em dados anteriores. Consequentemente, eles não prospectam um futuro diferente. Acontece que, além de os registros históricos certamente contemplarem preconceitos inerentes à sociedade e ao sistema de justiça, eles também podem não refletir os valores de hoje e não permitem a modificação. Isso quer dizer que transformar dados coletados em informações para treinamento dos modelos de IA torna difícil afastar-se do racismo e de outras injustiças históricas. Por isso, a sociedade civil, o governo e os envolvidos no sistema de justiça precisam ter

[3] Fundada em 1990, a Electronic Frontier Foundation é a principal organização sem fins lucrativos que defende as liberdades civis no mundo digital (EFF, 2020).

plena ciência das limitações e dos problemas da tecnologia ao considerarem utilizar ferramentas de inteligência artificial opacas e sem explicabilidade e transparência, sobretudo na segurança pública e em outras áreas do setor público.

Diante dessas circunstâncias, os presídios estão repletos de condenados por crimes sem vítimas (crimes de perturbação), sendo a maioria provenientes de bairros de baixa renda, negros ou imigrantes hispânicos. Enquanto isso, os crimes de fraude contra o mercado imobiliário e o mercado de ações — que trouxeram profundas repercussões nefastas para milhares de pessoas que perderam suas casas, empregos e planos de saúde na crise financeira de 2008 — não preenchem os relatórios policiais nem as estatísticas do PredPol. Consoante O'Neil, "criminalizamos a pobreza, acreditando o tempo todo que nossas ferramentas não são apenas científicas, mas justas" (O'Neil, 2020, p. 144).

Recentemente, impulsionados pela onda de protestos que tomou os Estados Unidos após a morte de George Floyd, a Amazon decidiu suspender, por um ano, o seu sistema de reconhecimento facial R*ekognition*. Porém, manteve intacta as parcerias para a utilização do aplicativo N*eighbours*. Sem prestar explicações detalhadas sobre a real motivação da iniciativa, a companhia afirmou apenas que defende que os governos implementem regulamentações mais rígidas para o uso ético da tecnologia de reconhecimento facial, principalmente nos Estados Unidos. A IBM aproveitou-se da ocasião para declarar que cancelou, de maneira definitiva, o desenvolvimento de softwares de reconhecimento facial (Paul, 2020). Isso se traduz em um forte indicativo de que as ferramentas de reconhecimento facial são ineficientes, além de produzirem graves danos sociais.

Além da problemática da ausência de diversidade, todos esses exemplos de modelos enviesados também se devem ao fato de que os softwares não são auditados, não estão expostos à legislação que os regulamente e fiscalize, não são obrigados a submeterem-se a revisão para verificar se há enviesamento embutido, tampouco são abertos para serem revisados e auditados por pessoas de fora. Na maioria dos casos, inclusive, sequer é dado ciência à população de que está sendo submetida a um algoritmo.

A empresa CyberExtruder, por exemplo, criadora de um software de reconhecimento facial para ser utilizado na segurança pública, admite não haver realizado testes para aferir vieses em sua ferramenta. Apenas confirmou que certas cores de pele são simplesmente mais difíceis para o software lidar, dadas as limitações atuais da tecnologia (Breland, 2017).

A Northpointe, que fornece o software Compas, contestou as conclusões do relatório publicado pela ProPublica, mas se recusou a revelar o funcionamento interno do programa, o qual considera comercialmente sensível (Buranyi, 2017). E assim,

mesmo diante do enviesamento dos modelos e da baixa efetividade das opções disponíveis, as Secretarias de Segurança Pública de vários estados brasileiros adquiriram e estão utilizando tecnologia de reconhecimento facial para auxiliar na prisão de procurados pela polícia.

Consoante dados levantados por Damasceno e Fernandes para a Folha de S. Paulo, por meio das secretarias estaduais de segurança e das polícias civil e militar, 20 estados brasileiros utilizam ou estão implementando a tecnologia de reconhecimento facial na segurança pública local. Outros três estudam sua implementação e apenas quatro estados não utilizam, não tiveram contato com o sistema ou planejam utilizar (Damasceno & Fernandes, 2021).

Após cerca de um ano de utilização da tecnologia de reconhecimento facial na segurança pública brasileira, as estatísticas relativas ao enviesamento do modelo já começaram a aparecer. Um levantamento realizado pela Rede de Observatórios da Segurança constatou que, das 151 prisões por reconhecimento facial que aconteceram no país, 90% eram de pessoas negras (Ramos, 2019).

Desprezando a intensa movimentação mundial, que escancara a discriminação algorítmica e requer o banimento da utilização de softwares de reconhecimento facial,[4] a adoção dessa ferramenta está se alastra pelos estados e municípios brasileiros, incentivada e financiada pelo Governo Federal (Ministério da Justiça e Segurança Pública, 2019).

Também merece menção o software Crime Radar, recentemente ofertado no Brasil pelo Instituto Igarapé, para servir de auxílio às polícias, fornecendo mapeamento e predição de crimes baseados em dados. Aos moldes do PredPol norte-americano, a plataforma digital se utiliza de algoritmos de aprendizagem profunda para prever taxas de crimes em diferentes bairros e horários. Segundo a empresa, "a plataforma pode ajudar as polícias a reduzir homicídios em até 10%, diminuir registros criminais em até 40% e encurtar drasticamente o tempo de resposta a ocorrências" (Instituto Igarapé, 2023).

Não há, no entanto, regulamentação que exija e fiscalize os protocolos dos softwares utilizados, que estabeleça os cuidados com relação aos dados para treinamento dos modelos, nível de acurácia, possibilidade de auditagem externa, transparência com relação ao uso, nem como será a proteção dos dados coletados. Isso porque o artigo 4º da Lei Geral de Proteção de Dados Pessoais (Presidência da República do Brasil, 2018) estabeleceu expressamente a necessidade de aprovação de lei específica para segurança pública e investigação criminal, o que não foi feito até o presente momento.

[4] Como a organização Algorithmic Justice League, o Big Brother Watch, a Liberty Human Rights, a campanha Ban Facial Recognition e a Internet Freedom Foundation.

Segundo aduz Vega Iracelay, "o avanço da inteligência artificial vem ocorrendo, até agora, em meio a um vácuo regulatório geral" (Vega Iracelay, 2018). Essa ausência de regulamentação, responsabilidade, auditagem e transparência na utilização da IA vem angariando consequências nefastas do ponto de vista social, pois amplia e perpetua discriminações e a seletividade que permeia a política criminal, sob a ilusória impressão de que se baseia em modelos matemáticos imparciais, precisos e objetivos. Em verdade, os resultados estão carregados de subjetividade e vieses discriminatórios tendentes a catalisar e perpetuar a marginalização e o racismo. Conforme bem apontado na exposição de motivos do Anteprojeto de Lei de Proteção de Dados para segurança pública e persecução penal:

> [...] Apesar do crescimento vertiginoso de novas técnicas de vigilância e de investigação, a ausência de regulamentação sobre o tema gera uma assimetria de poder muito grande entre os atores envolvidos (Estado e cidadão). Nesse contexto, o titular dos dados é deixado sem garantias normativas mínimas e mecanismos institucionais aplicáveis para resguardar seus direitos de personalidade, suas liberdades individuais e até a observância do devido processo legal. (Presidência da Câmara dos Deputados, 2019)

Há que se voltar uma atenção rigorosa sobre essas tecnologias, para que não venham a contribuir com a manutenção e intensificação do racismo estrutural que permeia o sistema penal, violando garantias fundamentais. Uma vez que os sistemas são comprovadamente inefetivos, intensificarão prisões ilegais e arbitrárias de pessoas negras mediante o falso reconhecimento positivo. Modelos como esses, que têm pontos cegos que os tornam discriminatórios, servirão apenas para reforçar padrões racistas socialmente difundidos, potencializados e perpetuados através do viés de automação.[5]

A proibição da utilização de reconhecimento facial para fins de segurança pública é primordial no estado atual da arte. Os falhos softwares de reconhecimento facial constituem tecnologias sobrepostas em alicerces discriminatórios, capazes de dinamizar a seletividade do sistema penal. Teremos muitos retrocessos em conquistas civilizatórias, assim como violação a direitos fundamentais, se difundirmos a utilização de reconhecimento facial na segurança pública. Também os modelos preditivos, utilizados para auxiliar o policiamento ou sentenças penais, a exemplo do Compas e do PredPol, não podem ser permitidos, mesmo que dotados de supervisão humana (Nunes, 2021). Modelos que forem treinados e alimentados a partir de estatísticas originalmente racistas servirão para retroalimentar a seletividade do sistema penal, escondidos sob um manto de "cientificidade matemática".

[5] "O viés de automação se apresenta como uma das espécies dos vieses cognitivos humanos que ocorre pela propensão de favorecer sugestões de sistemas automatizados de tomada de decisão. Isso ocorre quando o humano sobrevaloriza a resposta da máquina e passa a não refletir acerca da correção de seus resultados. Tal viés conduz as pessoas a não reconhecerem quando os sistemas automatizados erram e a seguirem seus resultados quando apresentadas informações contraditórias" (Nunes, 2021).

Com relação aos demais modelos algorítmicos, é fundamental ter uma preocupação especial com os dados que serão utilizados para treinamento dos modelos, assim como as variáveis que serão levadas em conta. Afinal, o algoritmo não cria o preconceito, apenas repete os padrões contemplados nos dados de treinamento e validação dos modelos (matemáticos ou probabilísticos). Portanto, um dos grandes desafios está em controlar esses dados e fiscalizar as correlações realizadas pelo aprendizado de máquina. É imprescindível uma conscientização e engajamento para o desenvolvimento de modelos que não perpetuem os preconceitos inseridos nos dados sociais, e que contaminam todo o sistema de justiça criminal. Urge uma regulamentação dos sistemas que se utilizam de inteligência artificial, nomeadamente aqueles que interferem em esferas jurídicas dos cidadãos, contemplando ações de mitigação, proteção e fiscalização adequadas e suficientes para proteção de liberdades individuais e dos valores democráticos.

3. Alternativas hábeis para atenuar a possibilidade de enviesamento dos modelos

Os dados constituem a matéria-prima para o desenvolvimento dos algoritmos de aprendizado de máquina. Se queremos melhorar o resultado final, livrá-lo o máximo possível de enviesamentos humanos, precisamos fiscalizar e controlar os dados utilizados para treinamento e validação dos modelos.

Para tanto, um importante critério a ser implementado na busca de uma maior neutralidade do algoritmo é a remoção de todos os atributos sensíveis presentes nos dados. Atributos como raça, gênero, posição social, religião, bem como indicativos que forneçam esses atributos, devem ser extirpados do treinamento do modelo. Logo, as perguntas realizadas devem ser cuidadosamente pensadas para que não abordem, tampouco indiquem sinais, de atributos sensíveis. A eliminação de informações sensíveis dos dados de treinamento conduziria, a princípio, a uma maior neutralidade algorítmica e resultaria em um modelo livre de discriminação.

Também é preciso ter atenção com os dados e variáveis que serão selecionados para treinar o algoritmo. A escolha equivocada pode resultar em um padrão enviesado. Tomemos como exemplo o que aconteceu com o PredPol. Lum e Isaac, da organização sem fins lucrativos H*uman* R*ights* D*ata* A*nalysis* G*roup*,[6] demonstraram que, para sugerir prováveis locais de crimes, o algoritmo se valia de dados sobre prisões anteriores, o que impulsionou o policiamento excessivo em bairros de maioria negra. No entanto, quando os pesquisadores alteraram os dados de treinamento, inserindo dados relativos ao uso geral de drogas da cidade, com base nas estatísticas nacionais, o policiamento foi distribuído de maneira muito mais uniforme (Lum & Isaac, 2016).

[6] O Grupo de Análise de Dados de Direitos Humanos é uma organização sem fins lucrativos e apartidária que aplica ciência rigorosa à análise de violações de direitos humanos em todo o mundo. Foi fundada em 1991 por Patrick Ball (https://hrdag.org/)

Serhii Pospielov, engenheiro de software, explica que esse tipo de viés de amostragem "ocorre quando os dados usados para treinamento não são grandes ou representativos o suficiente, levando a uma representação incorreta da população real" e salienta que uma alternativa para mitigar a ocorrência desse viés é a utilização de uma amostragem aleatória na seleção de dados. Para o engenheiro a amostragem aleatória simples é um método muito efetivo para garantir que todos na população tenham as mesmas chances de serem selecionados no conjunto de dados de treinamento, o que minimiza o viés de amostragem (Pospielov, 2022).

Outro critério que deve ser observado é a promoção da máxima transparência do modelo. Essa transparência deve abranger os dados que foram utilizados, os resultados obtidos, a equipe desenvolvedora, o nível de acurácia etc. O princípio da transparência, segundo Vega Iracelay, implica saber se o cidadão está submetido a uma interação com um sistema de IA, "se existem mecanismos para prevenir uma indevida discriminação e quais são os meios para conhecer o processo de decisão adotado por um mecanismo automatizado" (Vega Iracelay, 2018). A transparência dessas informações permite a auditagem do algoritmo e a adoção de medidas corretivas.

A transparência, e a consequente explicabilidade, foram erigidas a um dos cinco princípios elencados pela Organização para a Cooperação e Desenvolvimento Econômico em sua Recomendação do Conselho sobre Inteligência Artificial — a OECD/LEGAL/0449, a primeira norma intergovernamental sobre IA. Segundo a Recomendação, os envolvidos com sistemas de IA devem comprometer-se com a transparência e a divulgação responsável em relação aos sistemas, devendo fornecer informações apropriadas e consistentes com o estado atual da arte. Também deve ser garantida a conscientização das partes interessadas sobre suas interações com os sistemas de IA, inclusive no local de trabalho. Ademais, a transparência deve permitir que as pessoas afetadas por um sistema de IA compreendam o resultado e possibilitar que desafiem esse resultado com base em informações de fácil compreensão sobre os fatores e a lógica que serviu de base para a decisão.

É certo que muitos algoritmos são opacos, ou seja, ainda que exista uma regra de aprendizagem, a maneira como se dá o aprendizado não é controlada, nem mesmo pelos desenvolvedores — o que se denomina *black box*. Então, não há como obter 100% de transparência. Por outro lado, a transparência relativa aos dados utilizados no treinamento do modelo, aos resultados obtidos, ao nível de acurácia e à equipe desenvolvedora já permitiria a constatação de dados sensíveis, de vieses, de falhas relacionadas com a composição da equipe, e permitiria correções.

Ainda, é imperioso que os sistemas sejam projetados, contem com *designs*, que permitam a auditoria e a prestação de contas. A opacidade de alguns modelos não afasta a possibilidade de auditagem e a prestação de contas relativas aos dados, ao resultado final, a acurácia etc.

Outra questão que também merece especial atenção refere-se à diversidade nas equipes desenvolvedoras. É preciso uma equipe multidisciplinar e heterogênea. Tal diversidade deve envolver profissionais de distintas áreas, mas também a diversidade racial e de gênero, para ampliar, assim, as perspectivas e contextos representados e possibilitar a construção de modelos mais amplos e democráticos.

Vega Iracelay inclui essa preocupação e salienta a importância de garantir a participação das mulheres e das minorias para que o desenvolvimento de modelos de IA "*sean inclusivos, y representen la pluralidad de intereses, culturas, intereses y perspectivas*. E*n especial es imprescindible asegurar la igualdad de género en el desarrollo de* IA, *siendo este uno de los grandes desafíos en la* U*topía del* I*nternet"* (Vega Iracelay, 2018).

Também é preciso ampla publicidade para que os usuários estejam inequivocadamente cientes de que estão diante de um sistema de IA. A publicidade também permite maior controle por parte de agentes externos. Ademais, as vítimas de preconceitos, ou qualquer outro tipo de violações cometidas por sistemas IA, somente poderão ter ciência, contestar e provar a violação se tiverem acesso aos dados (Vega Iracelay, 2018).

Convém advertir que atribuir à supervisão humana a solução dos problemas de enviesamento algorítmico é indesejado e pueril. O viés algorítmico acontece, exatamente, porque os algoritmos reproduzem os vieses cognitivos humanos, os valores e os preconceitos inerentes aos indivíduos. Logo, não é suficiente uma rigorosa supervisão humana em decisões sensíveis, que interferem de maneira contundente em direitos fundamentais dos cidadãos. Há que se ter prudência e evitar o desenvolvimento de modelos que possam aprofundar as já abissais mazelas sociais do preconceito e da segregação (Nunes, 2021).

Por fim, insta salientar que, em que pese a existência desta contundente problemática de natureza ética, não devemos deixar de utilizar os algoritmos em inúmeras funções de auxílio. Até porque, por outro lado, se os algoritmos forem bem utilizados, eles constituirão grandes aliados no aprimoramento do ambiente institucional judicial brasileiro, trazendo a esperança de uma justiça mais célere, efetiva e capaz de transmitir maior segurança jurídica e estabilidade. Eles podem, inclusive, eliminar alguns dos vieses cognitivos humanos por não estarem sujeitos, por exemplo, à fadiga, ao humor, à fome — elementos externos à atividade de julgar, mas que sabidamente interferem na tomada de decisão judicial.

Nesse sentido cabe aventar a pesquisa realizada por Danziger et al. Os pesquisadores analisaram 1.112 decisões, de oito juízes israelenses, incumbidos de julgar pedidos de concessão de liberdade condicional. A conclusão foi que quando os juízes fazem decisões sequenciais repetidas, eles tendem a decidir mais a favor do *status quo* ao longo do tempo, mas eles podem superar essa tendência fazendo uma pausa

para refeição. O estudo mostrou que cerca de 65% das decisões foram a favor do autor no início de cada sessão (pela manhã, após o café da manhã e após a pausa para o almoço), e eles diminuíram gradualmente para 0-10% ao final de cada sessão. Os autores concluíram que fatores como o cansaço, o esgotamento mental e a fome influenciam diretamente na tomada de decisão judicial (Danziger et al., 2011). Em regra, a tomada de decisão mediante cumprimento de critérios objetivos e com o uso de modelos algorítmicos estaria desprovida desse tipo de viés cognitivo humano.

Por fim, como Vega Iracelay pontua muito bem, a ameaça promovida por sistemas de IA é real, já ocorre e tende a se aprofundar a passos largos, motivo pelo qual a regulamentação é inevitável, necessária e urgente. Contudo, uma vez que essa tecnologia transcende as fronteiras e tem uma natureza global, avaliar o seu real impacto, as melhores estratégias e a regulamentação deve ser feito de maneira global, e não isoladamente por país. "A cooperação entre governos e intervenientes privados é necessária para garantir que aproveitamos as oportunidades que a IA proporciona, mitigando os riscos que ela pode acarretar" (Vega Iracelay, 2018).

Ainda que constituam ferramentas extremamente promissoras e profícuas, a inteligência artificial tem limitações e, em muitos casos, a sua utilização impacta de maneira drástica os direitos fundamentais. Por isso, é imprescindível voltarmos uma atenção rigorosa sobre a utilização de inteligência artificial em situações que interferem na esfera jurídica dos cidadãos, em modelos algorítmicos que possam ferir garantias fundamentais. É premente que haja regulamentação, fiscalização e transparência para que a utilização de inteligência artificial não venha solapar conquistas civilizatórias e contribuir para a massificação e perenização da discriminação, do sexismo e do racismo.

Conclusões

O presente estudo se dedicou, inicialmente, a tecer uma breve introdução sobre os conceitos de inteligência artificial, algoritmos e explicar o uso problemático dos algoritmos e seus possíveis vieses. Essa introdução foi necessária para passar à segunda parte do estudo, a qual se ocupou de demonstrar alguns dos perniciosos impactos que modelos algorítmicos enviesados vêm produzindo, afrontando garantias fundamentais e intensificando o racismo e a segregação que já são presentes em nossa sociedade e no sistema de justiça.

Foi possível apresentar alguns modelos que vêm sendo utilizados e que estão chegando ao Brasil, e demonstrar de que maneira contribuem para a manutenção do racismo, do sexismo e da segregação.

Restou claro, assim, que é fundamental voltarmos uma atenção rigorosa sobre a utilização de inteligência artificial em situações que interferem na esfera jurídica dos cidadãos, em modelos algorítmicos que possam ferir garantias fundamentais.

Continuaremos a construir e utilizaremos modelos enviesados e potencialmente lesivos se não focalizarmos na base de dados usada para treinamento e validação dos modelos e se não nos preocuparmos com diversidade nas equipes desenvolvedoras, publicidade, transparência, explicabilidade e *accountability*.

No que diz respeito à utilização de reconhecimento facial na segurança pública, restou evidente que se trata de uma tecnologia que deve ser banida. São inúmeras as limitações atuais da tecnologia e a manifesta possibilidade de afronta a direitos fundamentais, como a privacidade, a liberdade de expressão, de reunião e de associação, além da dignidade e do devido processo legal. Não há maturidade tecnológica, efetividade, tampouco transparência e ciência por parte da população. O reconhecimento facial constitui uma ameaça real aos direitos fundamentais, carreando consequências perversas no âmbito social, o que amplia e perpetua o racismo e a seletividade presentes no sistema de justiça criminal.

Ao final, a terceira parte do estudo procurou elencar algumas medidas necessárias e hábeis para atenuar a possibilidade de enviesamento dos modelos, como o carecimento de regulamentação e fiscalização. É essencial que as leis acompanhem a tecnologia e regulamentem a utilização de modelos de IA, mormente com a finalidade na segurança pública, exigindo transparência, prestação de contas, diversidade dentro das equipes e auditagem externa. Estas são premissas imprescindíveis para a construção de modelos mais amplos e democráticos.

Não obstante, é certo que a utilização de inteligência artificial integra massivamente as nossas vidas e esse caminho apenas se expandirá e generalizará. Os desafios de natureza ética impendem vigilância, fiscalização, regulamentação, mas não afastam a circunstância de que a automação e a inteligência artificial podem ser utilizadas em inúmeras funções de auxílio, nomeadamente no âmbito judicial, proporcionando uma grande evolução na prestação da tutela jurisdicional.

A absorção de inovações tecnológicas nos tribunais brasileiros perfaz um poderoso aliado a fim de melhorar o ambiente institucional judicial brasileiro, trazendo celeridade, efetividade e segurança jurídica. A utilização de modelos algorítmicos pode, inclusive, eliminar alguns dos vieses cognitivos humanos, por não estarem sujeitos, por exemplo, à fadiga, ao humor, nem à fome — elementos externos à atividade de julgar, mas que sabidamente interferem de maneira determinante na tomada de decisão judicial.

Convém, por fim, enaltecer que as plataformas de IA desenvolvidas para o Judiciário brasileiro restringem-se, em sua maioria, ao apoio a atividades de automação. Com relação ao auxílio na tomada de decisão judicial, os algoritmos, hoje, operam no modo sugestivo e não decisório. A supervisão humana é imprescindível, não só pelo nível de IA que está operando no Judiciário, mas também por ser um dos requisitos exigidos pelo Conselho Nacional de Justiça para o funcionamento dos

modelos de IA — de acordo com a Resolução 332 (Conselho Nacional de Justiça, 2020), que dispõe sobre a ética no uso de IA no Poder Judiciário.

Na seara penal, que constitui o ponto nevrálgico da temática, principalmente em relação aos direitos fundamentais, o artigo 23º da Resolução 332 (Conselho Nacional de Justiça, 2020) prevê expressamente essa preocupação. O referido dispositivo estipula que a utilização de modelos de IA em matéria penal não deve ser estimulada, sobretudo com relação à sugestão de modelos de decisões preditivas.

O que precisamos é estar conscientes, atentos e engajados. Da mesma forma, é preciso regulamentação e fiscalização. É fundamental que haja transparência, auditagem e prestação de contas de cada modelo que interfere na esfera jurídica dos cidadãos. Isso permite ciclos positivos de *feedbacks* e *accountability*, e o aprimoramento dos modelos com vistas a contornar possíveis falhas de funcionamento e melhorar o seu resultado geral, reduzindo a incidência de nefastos vieses algorítmicos.

A inserção e o aumento da utilização de IA, em todos os aspectos de nossas vidas, é uma tendência inexorável. Em vez de resistir à mudança inevitável, devemos descobrir a melhor maneira de trilhar essa senda, tirando-lhe o melhor em benefício da justiça e da sociedade. As tecnologias disruptivas constituem ferramentas úteis ao aprimoramento da atividade jurisdicional. Todavia, além de olharmos para o futuro — e para as potencialidades que os recursos de IA podem ainda proporcionar — precisamos, necessariamente, preocuparmo-nos com um problema imediato e urgente, que diz respeito aos meios de evitarmos que os algoritmos reproduzam e ampliem nossas mazelas sociais do presente e solapem inegociáveis conquistas civilizatórias.

Referências

Biddle S. (2019, 26 de novembro). A*mazon's* R*ing Planned* N*eighborhood* "W*atch* L*ists*" B*uilt on* F*acial* R*ecognition*. The Intercept. https://theintercept.com/2019/11/26/amazon-ring-home-security-facial-recognition/

Breland, A. (2017, 4 de dezembro). H*ow white engineers built racist code – and why it's dangerous for black people*. The Guardian. https://www.theguardian.com/technology/2017/dec/04/racist-facial-recognition-white-coders-black-people-police

Buranyi, S. (2017, 8 de agosto). R*ise of the racist robots – how* AI *is learning all our worst impulses*. The Guardian. https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses

Conselho Nacional de Justiça [CNJ]. (2020). J*ustiça em números. 2020*. CNJ. https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/

Conselho Nacional de Justiça [CNJ]. (2020, 21 de agosto). Resolução n.º 332 de 21/08/2020. Dispõe sobre a ética, a transparência e a governança na produção e no uso de Inteligência Artificial no Poder Judiciário e dá outras providências. Diário da Justiça eletrônico (DJe) n.º 274. https://atos.cnj.jus.br/atos/detalhar/3429

Damasceno, V. & Fernandes, S. (2021, 9 de julho). S*ob críticas por viés racial, reconhecimento facial chega a 20 estados*. Folha de S. Paulo. https://www1.folha.uol.com.br/cotidiano/2021/07/sob-criticas-por-vies-racial-reconhecimento-facial-chega-a-20-estados.shtml

Danziger, S., Levav, J. & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. P*roceedings of the National Academy of Sciences of the United States of America*, 108(17), 6889-6892. https://doi.org/10.1073/pnas.1018033108

Da Rosa, A. & Guasque, B. (2021). O avanço da disrupção nos tribunais brasileiros. In D. Nunes, P. H. Lucon e E. Navarro Wolkart (Orgs.), I*nteligência artificial e direito processual*. O*s impactos da virada tecnológica no Direito Processual* (pp. 93-121). JusPODVIM.

Dastin, J. (2018, 10 de outubro). A*mazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Freitas, C. O. A. (2020, 4 de agosto). A *obscuridade dos algoritmos e a LGPD*. Estadão/Blog do Fausto Macedo. https://politica.estadao.com.br/blogs/fausto-macedo/a-obscuridade-dos-algoritmos-e-a-lgpd/

Fundação Getulio Vargas [FGV]. (2020). Introdução. In L. F. Salomão (Coord.), I*nteligência Artificial: tecnologia aplicada à gestão dos conflitos no âmbito do Poder Judiciário Brasileiro* (pp. 10-24). https://ciapj.fgv.br/sites/ciapj.fgv.br/files/estudos_e_pesquisas_ia_1afase.pdf

Haskins, C. (2019, 7 de fevereiro). A*mazon's Home Security Company Is Turning Everyone into Cops*. VICE. https://www.vice.com/en/article/qvyvzd/amazons-home-security-company-is-turning-everyone-into-cops

Hill, K. (2020, 29 de dezembro). A*nother Arrest, and Jail Time, Due to a Bad Facial Recognition Match*. The New York Times. https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

Instituto Igarapé. (2023). C*rime Radar*. https://igarape.org.br/tech/crimeradar/

Kelley, J. & Guariglia, M. (2020, 10 de junho). A*mazon Ring Must End Its Dangerous Partnerships with Police*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2020/06/amazon-ring-must-end-its-dangerous-partnerships-police

Larson, J., Mattu, S., Kirchner, L. & Angwin, J. (2016, 23 de maio). H*ow We Analyzed the COMPAS Recidivism Algorithm*. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lum, K. & Isaac, W. (2016). To predict and serve? S*ignificance, 13*(5), 14-19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Ministério da Justiça e Segurança Pública. (2019, 24 de outubro). Portaria n.º 793, de 24 de outubro de 2019. Regulamenta o incentivo financeiro das ações do Eixo Enfrentamento à Criminalidade Violenta, no âmbito da Política Nacional de Segurança Pública e Defesa Social e do Sistema Único de Segurança Pública, com os recursos do Fundo Nacional de Segurança Pública, previstos no inciso I do art. 7º da Lei nº 13.756, de 12 de dezembro de 2018. Diário Oficial da União de 25/10/2019. https://www.in.gov.br/en/web/dou/-/portaria-n-793-de-24-de-outubro-de-2019-223853575

Navas Navarro, S. (2017). Derecho e inteligencia artificial desde el diseño. Aproximaciones. In S. N. Navarro (Coord.), I*nteligencia artificial: tecnología, derecho* (pp. 23-72). Tirant Lo Blanch.

Nieva Fenoll, J. N. (2018). I*nteligencia Artificial y Proceso Judicial*. Marcial Pons.

National Institute of Standards and Technology [NIST]. (2019, 19 de dezembro). *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software*. https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software

Nunes, D. (2021, 25 de junho). A *supervisão humana das decisões de inteligência artificial reduz os riscos*? Consultor Jurídico. https://www.conjur.com.br/2021-jun-25/nunes-supervisao-humana-decisoes-ia-reduz-riscos

O'Neil, C. (2019, 22 de janeiro). CCBLAB. *The Authority of the Inscrutable: An Interview with Cathy O'Neil*. https://lab.cccb.org/en/the-authority-of-the-inscrutable-an-interview-with-cathy-oneil/

O'Neil, C. (2020). Algoritmos de destruição em massa: como o Big Data aumenta a desigualdade e ameaça a democracia (R. Abraham, Trad.). Editora Rua do Sabão.

Organização para a Cooperação e Desenvolvimento Econômico [OECD]. (2023, 7 de novembro). *Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449*. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Panch, T., Mattie, H. & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. *Journal of Global Health*, *9*(2). https://doi.org/10.7189%2Fjogh.09.020318

Paul, K. (2019, 17 de abril). '*Disastrous' lack of diversity in* AI *industry perpetuates bias, study finds*. The Guardian. https://www.theguardian.com/technology/2019/apr/16/artificial-intelligence-lack-diversity-new-york-university-study

Paul, K. (2020, 11 de junho). *Amazon to ban police use of facial recognition software for a year*. The Guardian. https://www.theguardian.com/technology/2020/jun/10/amazon-rekognition-software-police-black-lives-matter

Pospielov, S. (2022, 20 de junho). *How To Reduce Bias in Machine Learning*. Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/guest-article/how-to-reduce-bias-in-machine-learning/

Presidência da Câmara dos Deputados – Brasil. (2019, 26 de novembro). *Anteprojeto de Lei de Proteção de Dados para segurança pública e persecução penal*. https://static.poder360.com.br/2020/11/DADOS-Anteprojeto-comissao-protecao-dados-seguranca-persecucao-FINAL.pdf

Presidência da República do Brasil. (2018, 14 de agosto). *Lei n.º 13.709 de 14 de agosto de* 2018. *Lei Geral de Proteção de Dados Pessoais* (LGPD). Diário Oficial da União de 15/08/2018. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

Ramos, S. (Coord.). (2019). *Retratos da Violência. Cinco meses de monitoramento, análises e descobertas. Junho a outubro* – 2019. Rede de Observatórios da Segurança; Centro de Estudos de Segurança e Cidadania. https://cesecseguranca.com.br/textodownload/retratos-da-violencia-cinco-meses-de-monitoramento-analises-e-descobertas/

Rodríguez, P. (2018). Inteligencia Artificial. Cómo cambiará el mundo (y tu vida). Ediciones Deusto.

Sadowski, J. (2020). Too Smart: How Digital Capitalism is Extracting Data, Controlling Our Lives, and Taking Over the World. The MIT Press.

Snow, J. (2018, 26 de julho). *Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots*. ACLU. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

U.S. Equal Employment Opportunity Commission. (2015). *Diversity in High Tech. Executive Summary*. https://www.eeoc.gov/special-report/diversity-high-tech

Vega Iracelay, J. J. (2018). Inteligencia artificial y derecho: principios y propuestas para una gobernanza eficaz. I*nformática y* D*erecho*: R*evista* I*beroamericana de* D*erecho* I*nformático*, (5), 13-48. https://dialnet.unirioja.es/servlet/articulo?codigo=6845781

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J., Schwarts, O., Campolo, A. & Krueger, G. (2018, 6 de dezembro). AI N*ow* 2018 R*eport*. AI Now Institute. https://ainowinstitute.org/publication/ai-now-2018-report-2