Ronaldo Martins da Costa
Graduado em Análise de Sistemas pelo Centro Universitário Sagrado Coração (1996), Mestre em Ciência da Computação pela Universidade Federal de São Carlos (2002) e Doutor em Engenharia Elétrica pela Universidade de São Paulo (2010). Pós-doutorado no Stevens *Institute of Technology*, nos Estados Unidos (2015). Professor associado do Instituto de Informática da Universidade Federal de Goiás, com experiência em produção científica nas áreas de Inteligência Artificial, Visão Computacional e aplicações em saúde, agricultura e sistemas inteligentes.

Luiz Gustavo Santos Veríssimo
Graduado em Direito pela Universidade Federal de Goiás (2021) e em Sistemas de Informação pela Universidade Estadual de Goiás (2013). Especializado em Direito Digital e Proteção de Dados pelo Gran Centro Universitário (2024) e Especializado em Gestão Estratégica da Polícia Judiciária pela Escola Superior da Polícia Civil de Goiás (2021). Mestrando em Ciência da Computação pela Universidade Federal de Goiás.

ENVIESAMENTO ALGORÍTMICO EM MASSA: IMPACTOS, DESAFIOS ÉTICOS E PROPOSTAS DE MITIGAÇÃO NA ERA DA INTELIGÊNCIA ARTIFICIAL

RESUMO:

A quarta revolução industrial trouxe profundas transformações tecnológicas, impulsionadas pela integração da Inteligência Artificial (IA) em diferentes áreas da sociedade. Este artigo investiga como os vieses cognitivos humanos são transferidos para os sistemas de IA, gerando impactos éticos e sociais significativos. A pesquisa adota uma abordagem qualitativa baseada em revisão bibliográfica e análise de casos, abordando exemplos práticos nas mídias sociais e no Poder Judiciário. Além disso, discute estratégias para mitigar o enviesamento algorítmico, como a explicabilidade dos modelos e a diversidade nas equipes de desenvolvimento, bem como iniciativas regulatórias em contextos nacionais e internacionais. Conclui-se que o enfrentamento do enviesamento algorítmico requer esforços multidisciplinares para promover a transparência, a segurança e a confiança no uso da IA.

PALAVRAS-CHAVE:

Inteligência Artificial; Enviesamento Algorítmico; Vieses Cognitivos; Explicabilidade; Regulamentação de IA.

INTRODUÇÃO

As inovações tecnológicas trazidas pela quarta revolução industrial, também chamada de revolução digital ou revolução 4.0 (FLORES; SANTOS, 2021), especialmente pela digitalização das relações sociais e pelo advento da Inteligência Artificial (IA), geram desafios importantes do ponto de vista ético e social. Preocupações com o desemprego causado pela substituição do trabalho humano por máquinas, as decisões tomadas por robôs autônomos, a necessidade de transparência no desenvolvimento de sistemas inteligentes e a possibilidade de seu uso impactar os direitos das pessoas ocupam o palco deste complexo e disruptivo cenário.

Entre os maiores desafios no uso de IA, estão os vieses algorítmicos. Estes vieses advêm de heurísticas humanas e, ao serem transferidos para as máquinas através dos processos de IA, são por elas exacerbados (CASTRO; BOMFIM, 2020). Assim, questões de gênero, raça, etnia e ideologias são tratadas indevidamente pelos sistemas de IA gerando repercussões negativas sobre pessoas em todo o mundo. Isto ocorre desde as mídias sociais até por meio de mecanismos de busca, serviços de *streaming* de áudio e vídeo, decisões tomadas por órgãos do Poder Judiciário e pelos demais poderes públicos. Onde quer que haja IA atuando, seja de forma explícita ou latente, são reproduzidos os diversos vieses humanos.

Vale destacar que os vieses, em geral, são transferidos para as máquinas de forma não intencional, pois contidos nos conjuntos de dados usados nos treinamentos dos algoritmos de IA. Outrossim, originam-se no cérebro humano como uma estratégia de sobrevivência evolutiva, na forma de atalhos mentais utilizados para responder com a rapidez necessária aos estímulos complexos do ambiente (CIALDINI, 2020). Sendo assim, este trabalho visa perscrutar o conceito de vieses cognitivos, bem como de inteligência artificial e aprendizado de máquina, e investigar de que forma o enviesamento ocorre no campo da IA e quais as possíveis propostas para a eliminação ou redução destes vieses. Para tanto, são analisados casos reais de enviesamento algorítmico nas mídias sociais e no Poder Judiciário, dialogando com os achados na literatura sobre os temas em estudo.

METODOLOGIA

Trata-se de pesquisa qualitativa, do tipo exploratória, através de revisão bibliográfica e análise de casos. Para tal, realizou-se a busca pelas palavras-chave: "algoritmo", "alienação", "enviesamento", "inteligência artificial" e "vieses", bem como de seus sinônimos em língua inglesa: "*algorithm*", "*alienation*", "*bias*", "*artificial intelligence*", "*biases*" em bases de dados científicas reconhecidas, a saber: ACM Digital Library, Capes Periódicos, IEEE XPlore e Google Acadêmico, em um recorte de seis anos (entre 2019 e 2024). Assim, buscaram-se os trabalhos mais recentes e relevantes sobre o tema, os quais foram inteiramente analisados, pretendendo-se, através do método dedutivo, desenvolver um conhecimento atualizado e fidedigno a ser documentado. Por fim, complementou-se o estudo com livros importantes sobre o tema dos vieses cognitivos, bem como com legislações disponíveis *online*.

INTELIGÊNCIA ARTIFICIAL, MACHINE LEARNING E ALGORITMOS

Autores de diferentes áreas possuem definições diversas sobre inteligência artificial (IA). Em geral, a definem como a habilidade das máquinas de simular a capacidade humana de aprender e de resolver problemas complexos a partir do conhecimento obtido. Nesse sentido, Winston (1992) a conceitua como: "o estudo das computações que tornam possível perceber, raciocinar e agir."

O aprendizado de máquina – *machine learning* – é o campo da inteligência artificial que estuda a construção de modelos computacionais capazes de aprender com dados através de treinamento, que pode se dar com a intervenção direta do programador (aprendizado supervisionado) ou apenas de forma indireta (aprendizado não supervisionado) ou ainda de forma autônoma por parte da máquina, que com base em tentativa e erro reajusta os pesos de suas entradas até obter um grau aceitável de saídas corretas (aprendizado por reforço).

No âmbito do *machine learning*, destaca-se o *deep learning* ou aprendizado profundo. Trata-se de uma evolução dos sistemas de redes neurais artificiais, surgidos a partir da criação pelos cientistas de um modelo matemático que simula a estrutura do neurônio biológico. Os neurônios artificiais, chamados perceptrons, se agrupam formando redes neurais, as quais, agrupadas em múltiplas camadas, formam as redes neurais profundas, que funcionam com entradas e reentradas de dados em um processo denominado *backpropagation* ou retropropagação, simulando as sinapses cerebrais, produzindo o denominado aprendizado profundo. As redes neurais profundas são a tecnologia mais avançada e complexa no campo da inteligência artificial.

Em um nível mais elementar, tem-se que os algoritmos são a base do funcionamento dos processos de inteligência artificial. Podem ser conceituados como sequências de passos realizados pelo computador, a partir do recebimento de uma entrada – *input* – até a produção da saída esperada – *output*.

De forma mais precisa, CORMEN et al. (2012) definem algoritmo como: "qualquer procedimento computacional bem definido que toma algum valor ou conjunto de valores como entrada e produz algum valor ou conjunto de valores como saída". Ainda:

> Também podemos considerar um algoritmo como uma ferramenta para resolver um problema computacional bem especificado. O enunciado do problema especifica em termos gerais a relação desejada entre entrada e saída. O algoritmo descreve um procedimento computacional específico para se conseguir essa relação entre entrada e saída (CORMEN et al., 2012).

Feitas estas considerações, cabe introduzir o conceito de vieses e enviesamento e como ele ocorre no campo da Inteligência Artificial.

OS VIESES

Psicólogos evolucionistas afirmam que os vieses cognitivos são frutos de mecanismos cerebrais de simplificação da realidade percebida. Segundo CIALDINI (2022), o ser humano está o tempo todo submetido a uma enorme quantidade de informações, sendo que o cérebro não possui energia e tempo suficiente para processar todas elas, razão pela qual a mente cria atalhos perceptivos, geralmente baseados em estereótipos. A criação de vieses pelo cérebro humano é, portanto, inevitável.

Como exemplo, o autor cita o estereótipo de que "se é caro, então é bom" e "se é barato, então é ruim", frequentemente utilizado por compradores que, não possuindo informações suficientes para avaliar a qualidade de um produto, recorrem ao preço como atalho mental para decidir se irão comprá-lo ou não (CIALDINI, 2022, p. 21).

Nesse sentido, KAHNEMAN (2021) reforça que o pensamento humano é permeado por heurísticas e vieses. Segundo o autor, heurísticas são atalhos mentais que o cérebro utiliza para tomar decisões ou fazer julgamentos de maneira rápida e eficiente, especialmente em situações

de incerteza ou quando não se tem todas as informações disponíveis. Por sua vez, os vieses são erros sistemáticos de julgamento que resultam do uso de heurísticas ou de outros processos cognitivos limitados ou equivocados, refletindo em padrões de pensamento irracional.

Ainda, (HASELTON, 2016 apud AMBROS et al., 2019) ressalta o demonstrado por pesquisadores do comportamento humano, de que vieses cognitivos são uma adaptação biológica do cérebro humano para lidar com problemas específicos de forma ágil e eficiente em um ambiente informacional ambíguo e complexo. AMBROS et al. (2019) anotam a existência de sete tipos principais de vieses cognitivos, abordados por estudos de diferentes áreas: o da representatividade, o do *status quo*, o da ancoragem e ajustamento, o da confirmação, o da disponibilidade, o do espelhamento de imagem e o da atribuição.

De forma sintética, os autores trazem que o viés da representatividade é caracterizado por pré-julgamentos de pessoas e eventos com base em similaridades a um grupo ou evento particular conhecido; o do *status quo* é marcado pela preferência dos indivíduos a manterem seu estado atual, independentemente de um possível ganho com uma mudança de estado; o da ancoragem baseia-se na existência de uma suposição inicial tida como âncora, em que o indivíduo ajusta gradualmente as novas informações para serem compatíveis com a âncora.

O viés da confirmação, por sua vez, similar ao processo de ancoragem, induz o indivíduo a considerar apenas evidências que coadunem com suas hipóteses iniciais, ignorando qualquer evidência contrária; o da disponibilidade, trata-se do apego por informações que podem rapidamente ser trazidas à mente, em detrimento de outras evidências igualmente ou até mesmo mais válidas; o mecanismo do espelhamento de imagem funciona quando o indivíduo projeta seu modelo mental, esquema ou sistema de crenças em outra pessoa, assumindo que o outro se comportará como ele próprio se comportaria em determinado cenário.

Por fim, o viés de atribuição caracteriza-se pela tendência a supervalorizar os fatores internos e a subestimar o impacto de fatores externos, quando se tenta compreender e explicar o comportamento de outras pessoas, resultando em um mau julgamento do comportamento alheio.

Em resumo, a criação de vieses é uma atividade orgânica dos seres humanos, importante do ponto de vista da sobrevivência e evolução da espécie. Sendo assim e dialogando com o objeto de estudo, cabe investigar de que forma estes vieses são transferidos para as máquinas através da inteligência artificial e especificamente do *machine learning*, e quais as possíveis estratégias para mitigar esse fenômeno.

ENVIESAMENTO ALGORÍTMICO EM MASSA

Tendo em vista a 4ª Revolução Industrial, marcada por profundas transformações na sociedade, com a interação entre os domínios físicos, digitais e biológicos (CASTRO; BOMFIM, 2021), as tecnologias da informação e das comunicações passam a ser o elemento central de interligação entre países, economias, instituições, pessoas e culturas. Paralelamente, o enviesamento algorítmico tem se tornado cada vez mais presente em todos os espaços, visto que as relações sociais passaram a se dar através dos meios tecnológicos.

Para entender como se dá o enviesamento algorítmico em massa, é preciso compreender como os vieses cognitivos humanos são transferidos para as máquinas. Segundo VIEIRA (2019), essa transferência acontece por meio dos programadores e se dá basicamente em três momentos: a) durante o enquadramento do problema que se busca resolver com a programação; b) durante a coleta dos dados que servirão de base para o treinamento do sistema de IA; c) durante a preparação ou pré-processamento dos dados.

Com relação aos dados utilizados no treinamento dos modelos de IA, são essencialmente produzidos por seres humanos e, portanto, carregados com seus vieses. Tais vieses, que envolvem tendências ideológicas, de gênero e de raça, são transferidos para os algoritmos de forma não intencional, e uma vez que integram a estrutura cognitiva artificial da máquina, dificilmente poderão ser eliminados. Daí a necessidade de se filtrar os dados de entrada, antes do processo de treinamento.

Além dos dados usados no treinamento, CASTRO; BOMFIM (2021) ressaltam a existência de vieses oriundos dos próprios desenvolvedores do algoritmo. Observa-se que os programadores são, em sua maioria, brancos, de classe média e do sexo masculino, sendo que frequentemente transferem preconceitos ocultos nos padrões de linguagem para a máquina. Nesse ínterim, os autores anotam a necessidade de se ter equipes de desenvolvedores com maior diversidade racial, socioeconômica e de gênero, além da atuação de órgãos estatais com autoridade moral e de supervisão voltadas para o setor de desenvolvimento de software.

Ademais, entre os principais casos de enviesamento algorítmico abordados por diferentes autores, merecem destaque o enviesamento nas mídias sociais e no Poder Judiciário.

ENVIESAMENTO ALGORÍTMICO NAS MÍDIAS SOCIAIS

Segundo Castro e Bomfim (2021), o enviesamento nas mídias sociais ocorre principalmente através do sistema de recomendações. Os autores apontam que a internet agrupa pessoas que compartilham os mesmos *likes*, criando bolhas isoladas entre "aqueles que concordam com A" e "aqueles que concordam com B", existindo poucas conexões entre os dois grupos.

Outrossim, um estudo de JAIN et al. (2023) demonstrou que a mídia social baseada em vídeos YouTube reforça estereótipos de submissão e inferioridade aos homens, atribuídos às mulheres indianas, na medida em que recomenda vídeos de criadores de conteúdo daquele país contendo tais vieses. Isto se explica pelo fato de o algoritmo de inteligência artificial do Youtube utilizar o histórico de visualização do usuário como base para o sugestionamento de novos conteúdos, fazendo com que vídeos contendo vieses de gênero sejam recomendados.

Nesse sentido, é notório que as redes sociais digitais utilizam algoritmos que aprendem com a experiência passada do usuário e entregam novos conteúdos relacionados, reforçando vieses existentes e evitando a criação de conflitos, tudo em razão de um sistema de *marketing* que funciona como pano de fundo, gerando lucros estratosféricos para grandes corporações privadas, tanto anunciantes como proprietárias das redes.

BEZERRA; COSTA (2022) observam que a coleta massiva de dados pessoais (*Big Data*) por parte das grandes corporações do ramo da tecnologia (*Big Techs*) se dá no bojo do denominado capitalismo de dados, que é a forma de manifestação do sistema capitalista na qual os dados pessoais são a principal matéria prima e mercadoria, aquecendo uma economia baseada em vigilância e dominação, que se dá por meio das redes sociotécnicas.

Nesse ínterim, os autores anotam que existe uma grande assimetria entre a quantidade de dados que são coletados e o conhecimento acerca de como são utilizados pelos algoritmos atuantes sobre o *Big Data*, redundando em uma opacidade algorítmica, também chamada de caixa-preta algorítmica. De um lado, há a dificuldade de se explicar o funcionamento dos sistemas inteligentes, devido à complexidade de sua arquitetura lógica e matemática, que torna o processo decisório praticado pelo sistema de IA opaco até mesmo para *experts*. De outro lado, existem as leis de propriedade industrial e mesmo a ausência de interesse das corporações em conferir transparência e explicabilidade aos seus algoritmos.

Finalmente, Bezerra e Costa (2022) aduzem que, em sendo a sociedade desigual e eivada de preconceitos, principalmente raciais e étnicos, é lógico inferir que conjuntos de dados consubstanciados no *Big Data*, usados no treinamento dos modelos de IA, estarão igualmente permeados por tais vieses, pois são dados extraídos das próprias relações sociais e, mais precisamente, do comportamento dos cidadãos no ambiente digital. Assim, dá-se causa ao que os autores chamam de desigualdade datificada.

Em suma, uma vez que os algoritmos responsáveis pela classificação, agrupamento e entrega de conteúdo nas mídias sociais são treinados a partir de bases de dados enviesadas, é natural que o resultado de seu processamento seja igualmente enviesado, potencializando as desigualdades e outros vieses encontrados na sociedade (Puschel et. al 2022), reverberando em grupos de pessoas demasiadamente beneficiadas, em detrimento de outras, prejudicadas, como é o caso das mulheres indianas.

ENVIESAMENTO ALGORÍTMICO NO PODER JUDICIÁRIO

O uso de inteligência artificial no Poder Judiciário é crescente. No princípio, verificava-se a utilização de sistemas inteligentes para realizar tarefas repetitivas e sem cunho decisório, como para a movimentação de processos. Entretanto, em um segundo momento, verifica-se também a gradual utilização de algoritmos para apoiar as decisões judiciais ou mesmo para produzi-las por completo.

O caso mais emblemático é o do sistema COMPAS (*Correctional Offender Management Profiling For Alternative Sanctions*[1]), desenvolvido pela empresa Equivant (antiga Northpointe) e utilizado pelos tribunais dos Estados Unidos para subsidiar decisões acerca da aplicação de medidas alternativas à prisão, no âmbito da Justiça Criminal. O sistema é responsável por predizer a probabilidade de reincidência do condenado, a partir de dados obtidos com um questionário respondido por ele, contendo 137 perguntas de teor questionável, as quais devem receber uma pontuação de zero a dez. Conforme demonstrado por pesquisas posteriores, muitas destas perguntas são relacionadas à raça do indivíduo.

Nesse contexto, vale mencionar que a ONG ProPublica fez um estudo sobre o sistema COMPAS no ano 2016, do que restou comprovado que o sistema possui enviesamento racial. Conforme Nobre (2020), a entidade demonstrou que pessoas brancas receberam um escore mais baixo do que o real para a reincidência, enquanto pessoas negras receberam um escore mais alto do que o real. Além disso, embora o sistema devesse ser utilizado apenas para nortear o magistrado quando da decisão acerca de medidas alternativas à prisão, foi constatado que os juízes estavam utilizando os resultados do COMPAS nas sentenças condenatórias criminais, agravando a situação dos apenados.

Ainda segundo Nobre (2020), as perguntas feitas aos réus com o questionário do COMPAS envolvem sua situação socioeconômica, já que se referem ao bairro onde moram, entre outros fatores que resultam em estereótipos. Deste modo, ficou evidenciado que o COMPAS fere os princípios penais do estado de inocência[2], da culpabilidade[3] e da personalidade da pena[4], violando, por fim, a dignidade da pessoa humana. Como solução, a autora propõe a eliminação das questões envolvendo raça e condenações anteriores do réu, além da transparência acerca do código-fonte do algoritmo, possibilitando o direito ao contraditório e à ampla defesa.

1 Perfil de Gestão Correcional de Ofensores para Sanções Alternativas

2 O princípio segundo o qual ninguém será considerado culpado senão por meio de sentença penal condenatória definitiva.

3 O princípio segundo o qual ninguém pode ser responsabilizado penalmente por um crime sem que tenha agido com dolo (intenção) ou culpa (negligência, imprudência ou imperícia).

4 O princípio da personalidade da pena estabelece que a pena não pode ultrapassar a pessoa do condenado.

Um caso concreto de aplicação do COMPAS, ocorrido no ano de 2013, chamou a atenção da mídia. Eric Loomis foi preso e teve sua liberdade provisória negada pelo Poder Judiciário do estado de Wisconsin com base no alto escore dado pelo sistema. Sua defesa técnica pleiteou ao juízo de primeira instância que a empresa Equivant explicasse o funcionamento do algoritmo e como ele chegou ao escore atribuído ao réu, entretanto, o pedido foi negado. Recorrendo à Suprema Corte do estado de Wisconsin, a defesa teve outra negativa, fundamentada nas leis de proteção de propriedade intelectual (*software proprietário*).

Com novo recurso, desta vez direcionado à Suprema Corte do país, a defesa novamente se viu frustrada, pois a Corte afirmou que o caso não possuía repercussão geral no meio jurídico, e, portanto, não seria apreciado. Não obstante ocorrido no exterior, vale mencionar que no Brasil tal decisão seria, desde o início, marcada por nulidade insanável, já que viciada com relação à motivação, conforme inteligência do art. 93, inciso IX, da Constituição Federal: "IX todos os julgamentos dos órgãos do Poder Judiciário serão públicos, e fundamentadas todas as decisões, sob pena de nulidade (...)". Conforme VAZ et al. (2021), fundamentar significa "expor, lógica e coerentemente, as razões pelas quais determinada decisão foi proferida. Significa, pois, uma justificação". Não se trata apenas de requisito formal da decisão, mas substancial. Na seara da Justiça Criminal, o julgador deve expor seus fundamentos decisórios de modo a enfrentar cada um dos argumentos aduzidos pela defesa.

No caso em tela, o fato de o juízo ter fundamentado decisão cerceadora de liberdade de pessoa até então considerada inocente, tendo em vista o princípio do estado de inocência, com base em um output de um sistema de IA cujo modo de funcionamento e variáveis decisórias são desconhecidos, implica ainda em lesão ao direito ao contraditório e à ampla defesa do réu, importando em clara nulidade.

Noutro giro, dados do Painel de Projetos de IA no Poder Judiciário (CONSELHO NACIONAL DE JUSTIÇA, 2024) apontam que o uso de inteligência artificial no Poder Judiciário brasileiro ocorre com cautela, com funcionalidades que em geral apoiam as atividades das áreas meio, contando com pelo menos 140 projetos de IA em 62 tribunais. Entretanto, há um movimento crescente na busca pela automatização das decisões.

Nesta senda, merece destaque o advento do sistema Galileu, desenvolvido pelo Tribunal Regional Federal da 4ª Região (TRF-4), o qual possui a capacidade de pesquisar jurisprudências e redigir minutas de decisões judiciais automaticamente (SUPREMO TRIBUNAL FEDERAL, 2024). Tal sistema foi objeto de termo de cooperação técnica entre o TRF-4 e o Supremo Tribunal Federal, em agosto de 2024, a fim de que seja também implementado no Supremo. Destes eventos, é lógico inferir que a automatização da escrita de decisões judiciais por meio de IA em todos os tribunais é uma realidade próxima e inevitável.

IA E A NECESSIDADE DE TRANSPARÊNCIA

Para que o enviesamento da Inteligência Artificial seja combatido, é necessário que a tecnologia seja transparente. Para que haja transparência, é necessário ter explicabilidade. Neste vértice, ALVES; ANDRADE (2021) afirmam que é possível transformar a caixa-preta da IA em uma caixa de vidro. Com a metáfora, os autores se referem a conferir explicabilidade para os algoritmos de IA, de modo que os tornem transparentes para quem quer que os investigue. Para tanto, ressaltam a existência de duas principais técnicas de XAI (*Explainable Artificial Intelligence*) com potencial de resolver a questão da opacidade da maior parte dos modelos de IA: SHAPs (*Shapley Additive Explanations*) e LIME (*Local Interpretable Model-agnostic Explanations*).

As SHAPs funcionam indicando o peso da influência de cada variável envolvida no

processo decisório do algoritmo de IA. Através das *SHAPs*, o interessado tem um relatório de como o sistema decidiu e sob que percentual cada variável decisória atuou no processo. A técnica *LIME*, por sua vez, é aplicada a modelos de classificação de imagem. Consiste em gerar perturbações na imagem de entrada a fim de encontrar *superpixels*, que são áreas da imagem que correspondem ao que está presente na base de dados de classificação, assim apontando o porquê determinada imagem foi classificada de acordo com uma dada categoria.

Além de reduzir a opacidade, os autores afirmam que as técnicas de *XAI* podem revelar falhas algorítmicas e mitigar o enviesamento, garantindo ainda o direito à explicação. Ressaltam que os algoritmos frequentemente oferecem respostas aparentemente corretas, mas por premissas incorretas ou indesejáveis, o que pode vir à tona através do uso de *XAI*, possibilitando aos programadores a melhoria dos modelos.

Sob outro prisma, Alves e Andrade (2021) argumentam que nem todo sistema de IA precisa ser explicável, mas apenas aqueles que tratam informações pessoais sensíveis e geram um impacto importante nos direitos das pessoas. O necessário é que se busque um caminho para que o desenvolvimento e uso de IA seja fundamentado nos pilares da transparência e da segurança jurídica. É preciso que, sempre que a aplicação da IA gere repercussões sérias para alguém, tal indivíduo possua direito de explicação de como o algoritmo funciona, e de que forma ele decide.

De mais a mais, é inviável pensar em IA transparente apenas com a apresentação do código-fonte, principalmente quando se trata de *deep learning*, em que o funcionamento dos algoritmos é dinâmico e altamente complexo. O que se espera, com o amadurecimento do discurso sobre a transparência de IA, é que os futuros modelos sejam capazes de apontar probabilisticamente a influência de cada variável utilizada em seu processo decisório, tornando o seu funcionamento o mais claro possível e de acordo com o destinatário da explicação. Ou seja, a IA deve ser capaz de se autoexplicar de uma forma para um interessado leigo e de outra forma para um interessado com expertise no assunto, tendo como objetivo a compreensibilidade.

IA E REGULAMENTAÇÃO

A regulamentação de IA no Brasil ainda é incipiente. O art. 20 da Lei Geral de Proteção de Dados (LGPD) – Lei nº 13.709/2018 é o dispositivo legal com maior peso sobre a questão. O caput institui o direito de revisão de decisões tomadas por IA que utiliza dados pessoais:

> Art. 20. O titular dos dados tem direito a solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses, incluídas as decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade (BRASIL, 2018).

Ademais, os parágrafos primeiro e segundo do dispositivo instituem o direito à explicação sobre o sistema de IA que toma decisão automatizada, sempre que solicitado. Entende-se que a solicitação deve se dar por pessoa interessada, isto é, que tenha algum direito seu afetado pela decisão automática:

> § 1º O controlador deverá fornecer, sempre que solicitadas, informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada, observados os segredos comercial e industrial.
>
> § 2º Em caso de não oferecimento de informações de que trata o § 1º deste artigo baseado na observância de segredo comercial e industrial, a autoridade nacional poderá realizar auditoria para verificação de aspectos discriminatórios em tratamento automatizado de dados pessoais (BRASIL, 2018).

Também é digno de realce o art. 30 da LINDB (Lei de Introdução às Normas do Direito Brasileiro – Decreto-Lei nº 4.657/1942), que confere às autoridades públicas o poder regulatório para tratar de assuntos de sua competência. Vale mencionar que a Autoridade Nacional de Proteção de Dados (ANPD) possui competência administrativa para editar regulamentos acerca do uso de dados por inteligência artificial no país:

> Art. 30. As autoridades públicas devem atuar para aumentar a segurança jurídica na aplicação das normas, inclusive por meio de regulamentos, súmulas administrativas e respostas a consultas.
>
> Parágrafo único. Os instrumentos previstos no caput deste artigo terão caráter vinculante em relação ao órgão ou entidade a que se destinam, até ulterior revisão (BRASIL, 1942).

Ainda, possuem importante status em termos de regulamentação de IA no Poder Judiciário as resoluções nº 331/2020 e nº 332/2020 do Conselho Nacional de Justiça (CNJ), que tratam, respectivamente, da base nacional de dados do Poder Judiciário (Datajud) e da ética, a transparência e a governança na produção e no uso de inteligência artificial no Poder Judiciário. O art. 7º, parágrafo 3º da Resolução nº 332/2020 traz que, no caso do uso de IA para suporte às decisões judiciais, "a impossibilidade de eliminação do viés discriminatório do modelo de Inteligência Artificial implicará na descontinuidade de sua utilização, com o consequente registro de seu projeto e as razões que levaram a tal decisão".

Sob a ótica do direito comparado, tem-se que o primeiro marco legal significativo em regulamentação de IA no mundo é o IA Act da União Europeia (EUROPEAN COMMISSION, 2024). A legislação, que entrou em vigor em 1º de agosto de 2024, cria diversos mecanismos de supervisão do desenvolvimento de IA na Europa, com vistas a reduzir os riscos gerados pelo uso de IA, além de aumentar a transparência e a confiabilidade da tecnologia. Nesse sentido, a lei cria quatro níveis de risco para sistemas de IA: risco inaceitável, alto risco, risco limitado e risco mínimo.

Os sistemas de risco inaceitável passam a ser proibidos, sendo aqueles que importam em clara ameaça à segurança, aos meios de subsistência e aos direitos das pessoas. Como exemplo, tem-se a pontuação social por governos e brinquedos que usam assistência por voz que incentiva comportamentos perigosos.

Os sistemas de alto risco, por outro lado, estão sujeitos a obrigações rigorosas antes de serem colocados no mercado, que incluem a avaliação e mitigação de riscos, a alta qualidade dos conjuntos de dados que alimentam o sistema, visando eliminar riscos e resultados discriminatórios, a rastreabilidade das atividades, a documentação detalhada sobre o sistema e sua finalidade, medidas de supervisão humana adequadas para minimizar os riscos, além de alto nível de robustez, segurança e precisão (EUROPEAN COMISSION, 2024). Como exemplos de sistemas de IA de alto risco, estão aqueles voltados a infraestruturas críticas, como transportes, saúde, educação, aplicação da lei, migração e administração da justiça.

Os sistemas de risco limitado e de risco mínimo, por sua vez, são aqueles cujo uso possui baixa probabilidade de acarretar danos aos direitos individuais e coletivos. Para estes, a lei estabelece basicamente obrigações de transparência, visando promover a confiança. Para sistemas de risco limitado, como *chatbots*, é essencial informar os usuários de que estão interagindo com máquinas, garantindo decisões informadas, além de identificar conteúdos gerados por IA, especialmente aqueles destinados a informar o público ou que sejam falsificações profundas – *deepfake*. Já os sistemas de risco mínimo, como jogos de vídeo com IA ou filtros de spam, têm uso livre e abrangem a maioria das aplicações atuais na União Europeia.

Em que pese o Brasil ter tido iniciativas legislativas no sentido de se instituir uma Política Nacional de Inteligência Artificial, tal como o projeto de lei nº 5.691/2019 que restou prejudicado e arquivado, está claro que os debates sobre a área ainda carecem de muito amadurecimento, até quiçá se aproximar da experiência europeia.

CONSIDERAÇÕES FINAIS

A quarta revolução industrial trouxe consigo uma integração sem precedentes entre os domínios físicos, digitais e biológicos, destacando a Inteligência Artificial como um dos pilares dessa transformação. Contudo, a ascensão dessa tecnologia também revelou desafios significativos, como o enviesamento algorítmico, que reflete e amplia os preconceitos já presentes nas sociedades.

Este estudo abordou como os vieses cognitivos humanos, inevitáveis no cotidiano, são transferidos para os sistemas de IA durante o enquadramento dos problemas, a coleta e o processamento de dados. A análise de casos práticos, como o enviesamento no sistema COMPAS e nas redes sociais, demonstra os impactos sociais e éticos desse fenômeno. Observou-se que a ausência de diversidade humana nas equipes de desenvolvimento e a opacidade dos algoritmos agravam ainda mais o problema.

A necessidade de medidas mitigadoras é evidente. Estratégias como a explicabilidade algorítmica (*XAI*), por meio de técnicas como *SHAP* e *LIME*, mostram-se promissoras para tornar os sistemas de IA mais transparentes e confiáveis. Além disso, iniciativas regulatórias como o AI Act da União Europeia apontam para a importância de marcos legais robustos que assegurem a justiça, a segurança e a confiança no uso dessa tecnologia.

Assim, conclui-se que o enfrentamento do enviesamento algorítmico demanda uma abordagem multidisciplinar, que envolva não apenas avanços técnicos, mas também esforços legislativos e éticos, a serem empreendidos por governos, empresas e sociedade civil. No cenário brasileiro, faz-se necessário um amadurecimento dos debates na seara legislativa, com a inclusão de especialistas do campo da ética em ciência da informação, além de profissionais do ramo de IA, não se limitando aos juristas. Apenas por meio da conjugação de tais iniciativas será possível construir uma IA que promova a inclusão e a equidade, minimizando os riscos de danos aos direitos individuais e coletivos e de perpetuação de desigualdades históricas.

REFERÊNCIAS

ALVES, M. A. S.; ANDRADE, O. M. de. Da "caixa-preta" à "caixa de vidro": o uso da explainable artificial intelligence (XAI) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Direito Público**, [S. l.], v. 18, n. 100, 2022. DOI: 10.11117/rdp.v18i100.5973. Disponível em: <https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5973>. Acesso em: 23 dez. 2024.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, [2018]. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 23 dez. 2024.

BRASIL. **Decreto-Lei nº 4.657, de 4 de setembro de 1942**: Lei de Introdução às normas do Direito Brasileiro. Rio de Janeiro. Disponível em: <https://www.planalto.gov.br/ccivil_03/decreto-lei/del4657compilado.htm>. Acesso em: 23 dez. 2024.

CASTRO, B. F. de; BOMFIM, G. **A INTELIGÊNCIA ARTIFICIAL, O DIREITO E OS VIESES**. Revista Ilustração, [S. l.], v. 1, n. 3, p. 31–45, 2021. DOI: 10.46550/ilustracao.v1i3.23. Disponível em: <https://journal.editorailustracao.com.br/index.php/ilustracao/article/view/23>. Acesso em: 23 dez. 2024.

CIALDINI, Robert B. **As Armas da Persuasão 2.0**: edição revista e ampliada. Rio de Janeiro: Harper Collins, 2022.

CONSELHO NACIONAL DE JUSTIÇA. **Painel da Pesquisa sobre Inteligência Artificial 2023**: resultados da pesquisa IA no Poder Judiciário. Resultados da Pesquisa IA no Poder Judiciário. 2024. Disponível em: <https://paineisanalytics.cnj.jus.br/single/?appid=43bd4f8a-3c8f-49e7-931f-52b789b933c4&sheet=e4072450-982c-48ff-9e2d-361658b99233&theme=horizon&lang=pt-BR&opt=ctxmenu,currsel&select=Ramo%20da%20Justi%C3%A7a,&select=Tribunal,&select=Seu%20Tribunal/%20Conselho%20possui%20Projeto%20de%20IA?,>. Acesso em: 27 dez. 2024.

CORMEN, Thomas H. et al. **Algoritmos**: teoria e prática. 3. ed. São Paulo: Elsevier, 2012.

European Comission. **AI Act**. 2024. Disponível em: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Acesso em: 02 jan. 2024>.

FLORES, N. C. DA S.; SANTOS, R. DE S. A. . DIREITO E INTELIGÊNCIA ARTIFICIAL: METAMORFOSE, VIESES ALGORÍTMICOS E DECISIONISMO TECNOLÓGICO NO BRASIL. **Revista Direito e Justiça: Reflexões Sociojurídicas**, v. 21, n. 40, p. 99-113, 24 maio 2021.

JAIN, S.; KAUR, K. Exploring Gender Stereotyping in Indian Social Media Content: A NVivo-Based Content Analysis. In: INTERNATIONAL CONFERENCE ON COMPUTING COMMUNICATION AND NETWORKING TECHNOLOGIES (ICCCNT), 14., 2023, Delhi, India. Proceedings [...]. Delhi: IEEE, 2023. p. 1-6. DOI: 10.1109/ICCCNT56998.2023.10306799.

KAHNEMAN, Daniel. **Rápido e Devagar**: duas formas de pensar. Rio de Janeiro: Objetiva, 2021.

NOBRE, Daniela Kojiio. Processo Penal e Direitos Humanos: notas iniciais sobre sistemas que utilizam inteligência artificial em julgamentos. Dimensões Jurídicas dos Direitos Humanos, Rio de Janeiro, v. 4, p. 83-92, 2020. Disponível em: <https://www.caedjus.com/wp-content/uploads/2020/11/dimensoes_juridicas_dos_direitos_humanos_vol4.pdf#page=83>. Acesso em: 15 dez. 2024.

Supremo Tribunal Federal. **STF e TRT-4 firmam acordo para compartilhar desenvolvimento de sistema de inteligência artificial**: termo de cooperação visa automatização de tarefas burocráticas e da pesquisa de jurisprudência. Termo de cooperação visa automatização de tarefas burocráticas e da pesquisa de jurisprudência.. 2024. Disponível em: https://noticias.stf.jus.br/postsnoticias/stf-e-trt-4-firmam-acordo-para-compartilhar-desenvolvimento-de-sistema-de-inteligencia-artificial/#:~:text=Na%20assinatura%2C%20o%20ministro%20Barroso,impressionado%20com%20o%20sistema%20Galileu.. Acesso em: 27 dez. 2024.

VAZ, Andréa Arruda; GOMES, Eduardo Biacchi; DIAS, Sandra Mara de Oliveira. Limites Éticos para o Uso da Inteligência Artificial no Sistema de Justiça Brasileiro, de Acordo com a Lei 13.709 de 2018 (LGPD) e Resoluções 331 e 332 do Conselho Nacional de Justiça. **Revista Internacional Consinter de Direito**, [S.L.], p. 107-124, 21 dez. 2021. CONSINTER. http://dx.doi.org/10.19135/revista.consinter.00013.04.

VIEIRA, Leonardo Marques. **A PROBLEMÁTICA DA INTELIGÊNCIA ARTIFICIAL E DOS VIESES ALGORÍTMICOS: CASO COMPAS**. In: Brazilian Technology Symposium, Campinas, SP: Mackenzie, 2019.

WINSTON, P. H. Artifi cial Intelligence. 3. ed. [S. l.]: Addison-Wesley, 1992.