ENVIESAMENTO EM SISTEMAS DE INTELIGÊNCIA ARTIFICIAL E SEUS REFLEXOS NO DIREITO

Bias in Artificial Intelligence Systems and Its Repercussions on Law

Cinthia Obladen de Almendra Freitas[1]

RESUMO

O artigo apresenta conceitos e fundamentos necessários a compreensão do que é enviesamento algorítmico, especificando fontes e formas de viés (*bias*), bem como exemplificando problemas e pontos de atenção. O objetivo é mostrar que o enviesamento em sistemas de Inteligência Artificial é tema complexo e exige um olhar crítico tanto do ponto de vista jurídico como tecnológico, sem esquecer aspectos éticos, apontando para a explicabilidade da Inteligência Artificial. O trabalho adotou método dedutivo de pesquisa e tem por premissa que a Inteligência Artificial é composta por algoritmos que tratam dados, assumindo a IA como não-coisa e discutindo sobre os reflexos no Direito decorrentes do enviesamento em sistemas de IA. O artigo apresenta diversas considerações que relacionam desde a prestação de contas, auditabilidade e *accountability* até elementos do Direito Constitucional, Direitos Humanos e Direitos Fundamentais, mencionando ainda a governança e a proteção de dados.

Palavras-chaves: Direito e Tecnologia; Inteligência Artificial, Enviesamento.

ABSTRACT

The article presents concepts and fundamentals necessary to understand what algorithmic bias is, specifying sources and forms of bias, as well as exemplifying problems and points of attention. The objective is to show that bias in Artificial Intelligence systems is a complex issue and requires a critical analysis from both a legal and technological point of view, without forgetting ethical aspects, pointing to the explainability of Artificial Intelligence. The work adopted a deductive research method and is based on the premise that Artificial Intelligence is composed of algorithms that process data, assuming AI as a non-thing and discussing the repercussions on Law resulting from bias in AI systems. The article presents several considerations that relate from accountability, auditability and accountability to elements of Constitutional Law, Human Rights and Fundamental Rights, also mentioning governance and data protection.

Keywords: *Law and Technology; Artificial Intelligence, Bias.*

[1] Professora Permanente do Programa de Pós-Graduação em Direito (PPGD) da PUCPR. Doutora em Informática Aplicada pela PUCPR. Mestre em Engenharia Elétrica e Informática Industrial pela UTFPR. Engenheira Civil pela UFPR. Membro Consultivo da Comissão de Direito Digital e Proteção de Dados da OAB/PR. Membro Consultivo do Instituto Nacional de Proteção de Dados (INPD). E-mail: cinthia.freitas@pucpr.br

SUMÁRIO

INTRODUÇÃO

O tema viés ou enviesamento (do inglês *bias*) sempre desperta interesse e, normalmente, já vem carregado de pré conceitos e mitos. Freitas e Barddal (2019, p. 111) afirmam que "Em oposição aos seres humanos, os computadores não têm preferências nem atitudes. Se um modelo preditivo for corretamente projetado, ele será imparcial e não conterá vieses. Na prática, o modelo realizará cálculos e fornecerá respostas objetivas, neutras e confiáveis às consultas realizadas.". E isso sempre causa um certo espanto.

Explica-se por partes. Em oposição aos seres humanos, os computadores não têm preferências nem atitudes: Sim, computadores processam dados por meio de algoritmos, portanto, não-coisas (Freitas, 2024) e, assim, os computadores são excelentes realizadores de cálculos e não de gostos e preferências. Fato é, que por meio de dados e algoritmos pode-se extrair padrões comportamentais que representam gostos ou preferências. Se um modelo preditivo for corretamente projetado, ele será imparcial e não conterá vieses: Sim, reforça-se a parte do "corretamente projetado", isto significa que tanto a base de dados quanto os algoritmos não devem conter vieses. Na prática, o modelo realizará cálculos e fornecerá respostas objetivas, neutras e confiáveis às consultas realizadas: Sim, o modelo (matemático, estatístico ou probabilístico) realiza os cálculos necessários para apontar um resultado numérico, sendo esse resultado significativo para a aplicação em questão ou para a solução do problema ou para realizar uma tarefa em específico.

O que é viés e enviesamento? Por quais motivos pode-se ter uma decisão enviesada? Inicialmente, Freitas e Barddal (2019, p. 120) explicam que as técnicas de Aprendizagem de Máquina (*Machine Learning*) funcio-

nam "com a premissa de que se "aprende" como se comportar baseado na experiência passada. Mais importante, deve-se ter em mente que as decisões passadas são majoritariamente proferidas por seres humanos, que são potencialmente tendenciosos.". E esta experiência passada está inclusa nas bases de dados usadas para treinar os modelos.

Deste modo, algoritmos mal projetados, desenvolvidos ou testados podem produzir resultados potencialmente discriminatórios ou prejudiciais aos indivíduos. Um exemplo a ser citado é a seleção de candidatos/as à entrevista para uma vaga de emprego, em que o algoritmos pode ser desenvolvido com viés para não selecionar mulheres ou pessoas de determinada região ou bairro. E, se o sistema for uma "caixa-preta" (*black-box*), o que dificulta ainda mais uma IA Explicável, pode ser muito difícil compreender por que motivos determinadas candidatas foram rejeitadas dificultando a identificação e o tratamento de vieses. Há que se esclarecer que sistemas "caixa-preta" não são sempre discriminatórios ou possuem vieses, mas a falta de transparência em si pode dificultar a capacidade daqueles/as que são afetados/as por decisões automatizadas, de modo a não entender a lógica subjacente e seu impacto potencial. Outro exemplo é a caracterização de perfil (*profiling*) por meio da aplicação do *credit score* (avaliação de crédito) a partir de modelos de IA usados para aprovação ou não de crédito, sendo que os clientes bancários podem não ter um entendimento completo e correto sobre as decisões automatizadas que afetam suas vidas financeiras.

Para realização da pesquisa foi aplicado o método dedutivo. O artigo visa esclarecer o tema de enviesamento em sistemas de Inteligência Artificial (IA) trazendo conceitos e fundamentos necessários a compreensão do tema. Partindo-se da premissa de que a IA é composta por algoritmos que atuam sobre dados, assume-se a IA como não-coisa e discute-se sobre os reflexos no Direito decorrentes do enviesamento em sistemas de IA, trabalhando especialmente as fontes e formas de enviesamento desde a base de dados até o desenvolvimento de algoritmos. O artigo possui uma concepção tecno-jurídica a partir de revisão bibliográfica especializada. Há que se compreender aspectos tecnológicos e éticos para reduzir ou eliminar os impactos advindos do enviesamento em sistemas de IA, vi-

sando o benefício da sociedade contemporânea permeada por algoritmos na Era das Não-Coisas (Han, 2022). O artigo é resultado de projeto de pesquisa aprovado na Chamada Universal 10/2023 do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

1 ENVIESAMENTO EM SISTEMAS DE INTELIGÊNCIA ARTIFICIAL

Ross (2020, p. 1) afirmou que: "*If you are human, you are biased.*"[2]. E, mais: "*While we have generally thought about bias in relationship to people and prejudice, we have biases in all aspects of our lives. We are biased toward particular kinds of televiosion shows or movies, certain foods or kinds of foods, as well as certain kinds of books or stories*."[3] (ROSS, 2020, p. 4). Isso pode parecer uma afirmação dura e insensível, mas Ross ainda aponta que "*Virtually any preference we have is likely to have some bias associated with us*."[4].

Para Ferrara (2023, p. 2) "*Bias is defined as a systematic error in decision-making processes that results in unfair outcomes*."[5] Baer (2019, p. 3) define: "*Inclination or prejudice for or against person ou group, especially in a way considered to be unfair*."[6] E, complementa: "*Biases are double-edged swords*."[7] Como comportamento humano, Baer explica que "*Biases typically are not a character flaw or rare aberration but rather the necessary cost of enabling the human mind to make thousands of decisions every day im a seemongly effortless, ultra-fast*

---

2 Tradução livre: Se você é humano, você é tendencioso.

3 Tradução livre: Embora geralmente pensemos sobre vieses em relação às pessoas e preconceitos, temos preconceitos em todos os aspectos de nossas vidas. Somos tendenciosos em relação a determinados tipos de programas de televisão ou filmes, certos alimentos ou tipos de alimentos, bem como certos tipos de livros ou histórias.

4 Tradução livre: Praticamente qualquer preferência que tenhamos provavelmente terá algum preconceito associado a nós.

5 Tradução livre: Viés é definido como um erro sistemático em processos de tomada de decisão que resulta em resultados injustos.

6 Tradução livre: Inclinação ou preconceito a favor ou contra uma pessoa ou grupo, especialmente de uma forma considerada injusta.

7 Tradução livre: Os vieses são facas de dois gumes.

manner."[8] Isto acontece por que a mente "salta" diretamente para uma conclusão sem considerar todos os detalhes e fatos, ganhando velocidade na tomada de decisão. Não se está dizendo aqui que isto é certo ou errado, mas é a maneira como neurocientistas e psicólogos explicam o comportamento do cérebro humano. Baer (2019, p. 9) explica que esse funcionamento do cérebro é resultado da necessidade de equilíbrio entre 03 (três) objetivos simultâneos: precisão (*acurracy*), velocidade (*speed*) e eficiência (*energy*).

Ross (2020, p. 9) apresenta de maneira muito interessante o espaço de representação (domínios) de viés, expressando as maneiras como os vieses podem ser aplicados: construtivo ou destrutivo; e, ainda, se o viés é: a favor ou contra um determinado indivíduo ou grupo de indivíduos, conforme mostrado na Figura 1.

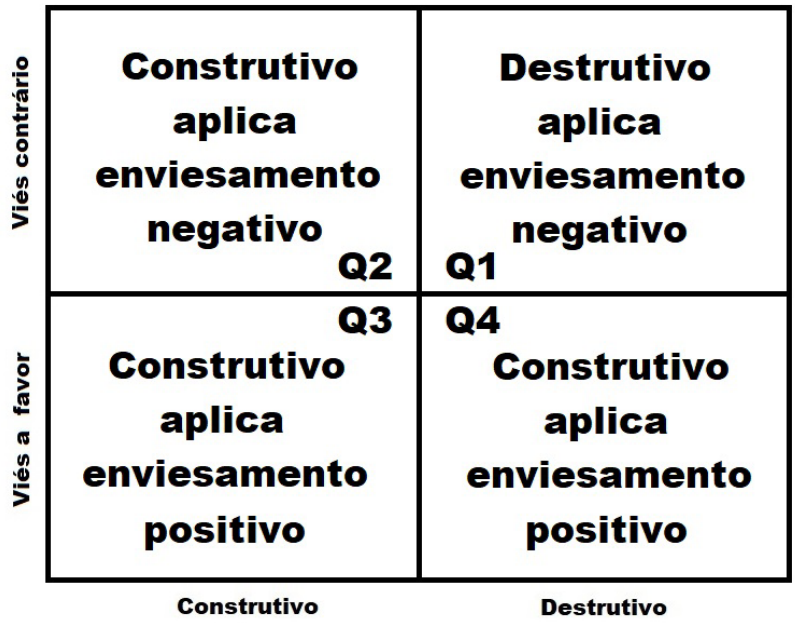

Figura 1 – Domínios de Viés. Fonte: adaptado de (Ross, 2020, p. 9).

[8] Tradução livre: Os vieses geralmente não são uma falha de caráter ou uma aberração rara, mas sim o custo necessário para permitir que a mente humana tome milhares de decisões todos os dias de uma maneira aparentemente fácil e ultrarrápida.

Explica-se Q1 a Q4 para facilitar o entendimento:

- Q1: é quando algo ou alguém destrutivo aplica viés contra um determinado grupo de indivíduos, sendo uma das maiores preocupações em termos jurídicos, visto que é sobre este tipo de enviesamento que as leis procuram ter por princípio a não discriminação;
- Q2: é quando algo ou alguém construtivo aplica viés contrariamente a um determinado grupo, por isso o enviesamento é negativo, visto não favorecer este grupo em específico;
- Q3: é quando algo ou alguém construtivo aplica viés a favor de um determinado grupo, de modo a favorecer tal grupo, sendo um risco estabelecer a exceção à regra;
- Q4: é quando algo ou alguém destrutivo aplica viés a favor de um determinado grupo, estabelecendo expectativas não realistas.

No contexto da IA, o viés da IA, também chamado de viés do Aprendizado de Máquina ou viés do algoritmo ou viés algorítmico ou viés estatístico, pode ser oriundo de diferentes fontes e refere-se à ocorrência de resultados tendenciosos devido à vieses humanos que distorcem dados de treinamento originais ou ao algoritmo da IA, gerando resultados distorcidos e potencialmente prejudiciais aos indivíduos afetados pela tomada de decisão tendenciosa. O viés algorítmico é definido como a presença de disparidades sistemáticas e injustas nos resultados de um sistema de IA, enraizadas em dados de treinamento tendenciosos e escolhas de projeto de desenvolvimento de *software* (Dwork, et al., 2012) (Min, 2023). Baer (2019, p. 5) mostra o outro lado da moeda explicando que o enviesamento está presente se a média de todas as previsões desviar sistematicamente da resposta correta e, portanto, considerando o seguinte exemplo, se um algoritmo de uma determinada instituição bancária atribuir uma probabilidade de inadimplência de 5% a 10.000 clientes diferentes, seria de se esperar que 500 dos 10.000 clientes entrassem em inadimplência, mas ao se analisar a situação descobre-se que, na realidade, 10% dos clientes entram em inadimplência. Porém toda vez que um inadimplente tem um

passaporte de um país específico, o algoritmo corta a estimativa real pela metade tornando o algoritmo tendencioso - neste caso, a favor desses cidadãos.

Vieses são entendidos também como atalhos cognitivos que podem resultar em julgamentos que levam à práticas discriminatórias. Para Smith e Rustagi (2020, p. 20) *bias* é "*tendency, inclination, or prejudice toward or against something or someone*".[9] Os autores explicam que os seres humanos vivenciam vieses o tempo todo e que os cérebros humanos são programados para serem tendenciosos, visto que, de um modo geral, o cérebro humano opera com 02 (dois) tipos de modelos: a) modelo que se refere ao pensamento automático e rápido que opera com pouco ou nenhum controle voluntário, gerando as impressões e intuições e, também, informando os instintos "viscerais" ou emocionais e b) modelo que necessita de esforço deliberado e está vinculado à especializações e escolhas. Por exemplo, selecionar frutas e verduras no supermercado é uma atividade que ativa o modelo tipo (a). Já o reconhecimento facial de pessoas ativa o modelo tipo (b). O modelo tipo (a) auxilia os seres humanos a organizar e gerenciar todos os estímulos que são enfrentados constantemente. Por outro lado, é aqui que os vieses cognitivos aparecem, visto que o modelo (a) depende de associações e categorias para discernir padrões e fazer julgamentos de forma rápida e eficiente. E, os seres humanos aprendem a fazer associações e categorizar com base em suas experiências pessoais, educação, criação e comunidades, portanto, também nos estereótipos e normas (sociais, culturais, legais, entre outras) que os acompanham.

Smith e Rustagi (2020, p. 20) apontam ainda que o viés estatístico ou viés de algoritmo afeta a precisão e a confiabilidade da previsão de um modelo em um sistema de IA. Por isso, os autores, definem "*biased AI*" (IA tendenciosa) como sistemas de IA que resultam em previsões e saídas imprecisas e/ou discriminatórias para certos subconjuntos da população. Importante perceber que ao se referir a certos subconjuntos da população, os autores estão apontando para erros específicos e não gene-

[9] Tradução livre: tendência, inclinação ou preconceito em direção ou contra algo ou alguém.

ralizados para toda a população (estatística). Um exemplo de viés de algoritmo que afeta a confiabilidade é mostrado no exemplo de Baer (2019, p. 5), no qual uma determinada instituição bancária adquire um modelo para análise de *credit score* a partir da taxa média de utilização do cartão de crédito como um dos preditores de inadimplência dos clientes. Neste sentido, o algoritmo tem por premissa que clientes com baixa utilização do limite concedido são mais confiáveis que clientes com altas taxas de utilização do limite do cartão de crédito. Assim, para os clientes considerados seguros, o algoritmo aumenta o limite, o que cria uma referência circular, ou seja, no momento que o limite é aumentado, a taxa de utilização diminui, fazendo com que o algoritmo aumente ainda mais o limite e isso pode acontecer até que o limite de crédito atinja níveis estratosféricos que estão totalmente além das possibilidades de pagamento por parte do cliente ao banco. O exemplo é ilustrativo e, é claro, que o algoritmo pode ter um regra de parada.

Min (2023, p. 1) entende que dados de treinamento tendenciosos contribuem significativamente para resultados tendenciosos em sistemas de IA, exigindo técnicas de pré-processamento de dados para mitigação. E, ainda, explica que se as escolhas de *design* algorítmico não forem cuidadosamente consideradas, podem perpetuar os vieses, necessitando-se de algoritmos com imparcialidade.

Entende-se, portanto, que o viés pode surgir de várias fontes, incluindo a coleta de dados, *design* de algoritmo e interpretação humana. Por exemplo, modelos de Aprendizagem de Máquina podem aprender e replicar padrões de vieses presentes nos dados usados para treiná-los, alcançando resultados injustos ou discriminatórios. É importante identificar e abordar o viés na IA para garantir que esses sistemas sejam justos e equitativos para todos os usuários.

2 FONTES E FORMAS DE ENVIESAMENTO EM SISTEMAS DE IA

Collett e Dillon (2019) apresentaram um estudo relacionado à *Biased Datasets*, ou seja, bases de dados enviesados. E, descreveram três fontes

de viés quando se trata de sistemas de IA, a saber: dados, algoritmos e ausência de transparência. As autoras explicam que os conjuntos de dados não são representativos, especialmente quando se trata de grupos minoritários, de maneira que em alguns casos, isso é causado pelo fato de que alguns não têm acesso à tecnologia e, portanto, não estão gerando dados. Isso significa que esses indivíduos não são representados nos dados e isso propaga vieses e exclusões. Desta explicação pode-se compreender que não gerar dados amplia a invisibilidade de determinados grupos. Já os algoritmos necessitam de desenvolvedores, programadores, engenheiros/as de *software* que por sua vez não exibem diversidade. Assim, considerando que a natureza humana trabalha dentro da visão individualizada e pessoal de mundo, isso leva à imposição de visões e valores em sistemas algorítmicos, o que, por sua vez, reforça os vieses sociais. Desta explicação pode-se compreender que os algoritmos são desenvolvidos por indivíduos que carregam suas visões de mundo para dentro dos sistemas de IA ou outro sistemas computacionais. E, a ausência de transparência é devida aos sistemas de IA não fornecerem explicações para suas decisões.

Baer (2019, p. 11-18) apresenta 05 categorias de vieses: (i) *action-oriented biases*: reflete a visão da natureza em que agir com maior velocidade muitas vezes é necessário e obrigatório, ou seja, o que lhe faz pensar que pode sobreviver na natureza com maior eficiência: entender de IA ou saber lidar com uma cobra venenosa?; (ii) *stability biases*: os vieses de estabilidade são uma forma da natureza ser eficiente e um bom exemplo é o preço âncora em um cardápio de restaurante, o qual levará o cliente a escolher pratos mais caros (Dooley, 2012) (Freitas; Batista, 2015); (iii) *pattern-recognition biases*: essa categoria de viés lida com um problema muito incômodo para a tarefa de reconhecimento por humanos: grande parte da percepção sensorial humana é incompleta, e há muito ruído no que é percebido por cada indivíduo. Esses vieses são particularmente relevantes porque o reconhecimento de padrões é uma tarefa de alto interesse para ser realizada por algoritmos. E, para resolver o problema de dar sentido a dados ruidosos e incompletos, o cérebro humano precisa desenvolver regras. Erros sistemáticos (ou seja, vieses) ocorrem se tais regras estiverem erradas ou se uma regra for incorretamente aplicada, levando tanto ao

viés de confirmação (*confirmation bias*) quanto ao viés de estereótipos (*stereotyping bias*), explicados a seguir; (iv) *interest biases*: esse enviesamento vai além de meros atalhos baseados em heurísticas, visto que a partir dos objetivos de ação, estabilidade e reconhecimento de padrões visa-se tomar a decisão "correta" da forma mais precisa, rápida e eficiente possível, de modo que o viés de interesse considera explicitamente a questão "o que eu quero?". Caso cada indivíduo analise cuidadosamente seus pensamentos durante a tomada de decisão, perceberá que o subconsciente influencia a tomada de decisão (Dooley, 2012); e (v) *social biases*: são uma subcategoria de viés de interesse, mas sua importância justifica alguns esclarecimentos, uma vez que seres humanos são sociais e temem estar fora de um grupo ou isolados. Por isso, na tomada de decisão o cérebro sopesa os benefícios e qualquer ação possível contra o risco de ser condenado ao ostracismo ou, atualmente, ao cancelamento nas redes sociais. O enviesamenteo social pode ser decorrente, portanto, das seguintes formas de viés: *sunflower management* – é o viés de concordar com o chefe e *groupthink* – é o viés de concordar com o consenso do grupo.

Como formas e fontes de enviesamento, pode-se ainda citar (Ferrara, 2023) (Min, 2023) (Smith; Rustagi, 2020):

- Viés de algoritmo: quando o projeto (*design*) do algoritmo contém especificações ou particularidades que são tendenciosas ou quando o resultado do modelo não é corretamente compreendido e gera interpretações enviesadas;
- Viés cognitivo: como mencionado anteriormente, os seres humanos são tendenciosos e o viés pessoal pode se infiltrar em uma coleta de dados ou desenvolvimento de algoritmo sem ser percebido de maneira antecipada, impactando o conjunto de dados ou o comportamento do modelo;
- Viés de confirmação: intimamente relacionado ao viés cognitivo, isso acontece quando a IA depende muito de crenças ou tendências preexistentes nos dados, ampliando os vieses já existentes, de modo a não identificar novos padrões ou tendências;

- Viés de exclusão: ocorre quando dados importantes são deixados de lado, por exemplo, durante o treinamento, visto que o desenvolvedor não os considerou como elementos importantes;
- Viés de medição: causado por dados incompletos, geralmente por descuido ou falta de preparação da base de dados, sendo esta não representativa e/ou significativa em termos estatísticos, resultando em uma amostragem incompleta ou deficiente;
- Viés de homogeneidade de exogrupo: há uma tendência de as pessoas terem uma melhor compreensão dos membros do endogrupo — o grupo ao qual pertencem — e pensarem que são mais diversos do que os membros do exogrupo. Por isso, esse tipo de viés é resultado de não se saber o que não se sabe, ou seja, o grupo minoritário não integra a base de dados de treinamento, por exemplo, por critério racial ou de idade, e, assim, os algoritmos são menos capazes de distinguir entre indivíduos que não fazem parte do grupo majoritário nos dados de treinamento, levando a vieses, classificação incorreta e respostas incorretas;
- Viés de preconceito: ocorre quando estereótipos e suposições sociais falhas encontram seu caminho no conjunto de dados do algoritmo, o que inevitavelmente leva a resultados tendenciosos. Por exemplo, a IA pode retornar resultados mostrando que apenas homens são juízes;
- Viés de recordação: decorre do procedimento de rotulação de dados, sendo os rótulos aplicados de forma inconsistente por observações subjetivas, Por exemplo, pode-se rotular uma imagem de impressão palmar (das mãos) como se fosse plantar (dos pés);
- Viés de amostra ou seleção: a base de dados apresenta quantidade insuficiente de exemplos, os exemplos não são representativos ou são incompletos. Por exemplo, ao se treinar um modelo chatGPT para médicos com base em documentação gerada

por médicos iniciantes, médicos especialistas e experientes irão considerar o sistema muito aquém do esperado;

- Viés de estereótipos: pode ocorrer quando um sistema de IA reforça estereótipos prejudiciais. Por exemplo, ao se traduzir "*the taxi driver is waiting for the passenger*" para o português resulta "o taxista está esperando o passageiro", no gênero masculino.

Especificamente sobre vieses introduzidos pelas bases de dados, Baer (2019, p. 69-77) explica que, incialmente, considerando o procedimento de coleta de dados, os vieses podem ser advindos de duas situações: a) dados qualitativos subjetivos: criados por humanos, como por exemplo as avaliações de restaurantes que são naturalmente tendenciosas. O problema está em saber como tais dados foram gerados; b) dados aparentemente quantitativos: envolvem números que são gerados por um processo semelhante aos dados subjetivos e, portanto, podem ser afetados pelos mesmos problemas, embora pareçam enganosamente objetivos. Considera ainda, dois tipos de situações em que a fonte de dados transmite um viés: a) dados espelham comportamento tendencioso: por exemplo, o pagamento de bônus anual aos colaboradores de uma empresa pode dar a impressão de que advém de uma métrica objetiva, mas pode refletir vieses de gênero na maneira como a organização avalia homens e mulheres; b) eventos traumáticos: são eventos únicos que não são preditivos em resultados futuros, mas, no entanto, criam um viés descomunal do algoritmo.

E, relacionado ao procedimento de desenvolvimento de modelos, o autor também apresenta duas maneiras pelas quais os cientistas de dados podem introduzir vieses: a) vieses conceituais: gerados por causa de decisões específicas de *design* de modelo que criam uma representação distorcida da realidade embutida na base de dados; b) uso de dados inadequados: cria vieses por meio de falhas na maneira como os dados são pré-processados, agregados ou transformados.

De maneira interessante e inovadora, Baer (2019, p.87-94) aborda um ponto bastante polêmico relativo aos vieses introduzidos pelo próprio

algoritmo o que leva ao entendimento sobre erros, tipos de erros e as consequências dos erros de algoritmos. É importante perceber que Baer (2019, p. 88) afirma que: "*Humans are used to making binary predictions, such as "It will rain today!" or "Don't hire her: I don't think she's up to the task." or "Let's buy that one: it looks yummy!"*[10]". E, contrariamente, os algoritmos trabalham com probabilidades (valores entre 0% e 100%). E os algoritmos apontam: "*83% chance of rain*."[11] ou "*98% chance of you liking this cake* ."[12]. Deste modo, toda estimativa de um algoritmo está acompanhada de uma probabilidade de erro (incerteza), havendo risco do algoritmo estar completamente errado em sua estimativa o que não atende ao critério de estar, em média, correto. E, isto pode acontecer por se tratar de uma base de dados não significativa e representativa do problema ou o modelo estar extrapolando resultados a partir de uma base de dados pequena (em quantidade de exemplares e em representatividade) ou o modelo foi treinado em excesso (*overfitting*). Há que se analisar o tamanho da amostra em termos de diversidade de situações reais a serem representadas na base de dados.

3 REFLEXOS NO DIREITO

Os modelos de IA refletem os vieses presentes em seus dados de treinamento, o que pode levar a resultados que perpetuam estereótipos ou discriminação, mesmo que involuntariamente. Isso é contrário aos preceitos de diversidade, equidade e inclusão. É necessário que os desenvolvedores de *software* conduzam revisões amplas para garantir que os resultados não sejam apenas precisos, mas também livres de vieses não intencionais.

Há que se verificar e avaliar as informações de origem que fundamentam os resultados de sistemas de IA. Necessita-se também que tanto

---

10 Tradução livre: Os humanos estão acostumados a fazer previsões binárias, como "Vai chover hoje!" ou "Não a contrate: não acho que ela esteja à altura da tarefa." ou "Vamos comprar aquele: parece uma delícia!".

11 Tradução livre: 83% de probabilidade de chuva.

12 Tradução livre: 98% de probabilidade de você gostar deste bolo.

desenvolvedores quanto usuários fiquem atentos à ocorrência de vieses nos resultados, principalmente quando as aplicações forem utilizadas para tomada de decisão ou análise de dados (*credit score*[13] [14] [15], *consumer score*[16], *health score*[17], *social score*[18], entre outros).

Ao enfrentar os problemas decorrentes do enviesamento em sistemas de IA, as empresas podem: mitigar riscos, manter reputação da marca, ter uma proposta que agregue valor, estar em conformidade ou à frente de legislações e regulamentos voltados à IA e aumentar a competitividade (Smith; Rustagi, 2020, p. 4). No outro lado da balança estão os perigos, sendo que Smith e Rustagi (2020, p. 6-7) apresenta o "Mapa do Enviesamento em IA" (*Bias in AI Map*), o qual parte dos dados usados para treinar os algoritmos, explicando que esse é um ponto crítico do Mapa. E assim é, devido a coleta massiva de dados a partir de inúmeros pontos distintos de coleta, por exemplo, as atividades diárias dos indivíduos – comportamento do consumidor – em estruturas multiplataformas tecnológicas ou não. Esses dados estarão repletos de preconceitos raciais, econômicos e de gênero, por exemplo, visto que a influência humana não pode ser eliminada dos dados. São os seres humanos que decidem o que, onde e como os dados são coletados e categorizados, bem como a parametrização dos dados. Além disso, os dados são rotulados, o que é uma atividade humana, por muitas vezes e, portanto, carregada da subjetividade. Desse modo, pode-se considerar que os vieses adentram os algoritmos de diferentes maneiras, desde a definição da funcionalidade de um modelo de IA até as restrições ou limitações sob as quais o modelo irá operar, podendo também ser proveniente do procedimento de seleção dos *inputs* (que definem as variáveis) que o algoritmo deve considerar para,

---

13 Disponível em: https://www.usa.gov/credit-score

14 Disponível em: https://www.canada.ca/en/financial-consumer-agency/services/credit-reports-score/order-credit-report.html

15 Disponível em: https://www.servicopublico.pt/credit-score/

16 Disponível em: https://skeepers.io/en/blog/customer-scoring-definition-benefits/

17 Disponível em: https://play.google.com/store/apps/details?id=com.csoft.healthscore&hl=en_US

18 Disponível em: https://www.technologyreview.com/2022/11/22/1063605/china-announced-a-new-social-credit-law-what-does-it-mean/

posteriormente, encontrar padrões e apontar decisões. E, os perigos advindos de sistemas de IA enviesados podem ser oriundos de duas fontes: a) jurídicos/legais: alocação injusta de recursos, serviço abaixo da média, riscos à saúde, violação de liberdades civis, depreciativo e reforço aos preconceitos e b) financeiros/econômicos: riscos, marca/reputação, custos de recursos alocados e multas.

Vários trabalhos abordam o problema de garantir que os algoritmos sejam justos, ou seja, que não apresentem viés em relação a grupos étnicos, de gênero ou outros grupos. Para tanto, Smith e Rustagi (2020, p. 11) apontam que para estabelecer políticas e práticas que permitam o desenvolvimento responsável de algoritmos pode-se considerar algumas perguntas, a exemplo de: a) O processo de desenvolvimento é padronizado com ferramentas para identificar, documentar e mitigar as deficiências e riscos do modelo de IA?; b) A equipe considera onde e como integrar processos humanos durante o ciclo de vida do sistema de IA?; O sistema de IA é auditado – incluindo auditorias internas e externa?; d) Existem mecanismos de *feedback* robustos incorporados aos sistemas de IA para que os usuários possam facilmente relatar problemas de desempenho que encontrarem e (se não houver como cancelar), ter um processo de apelação para solicitar revisão humana?. Essas quatro perguntas estão relacionadas, não somente com o desenvolvimento de sistemas de IA, mas com a IA Explicável (Freitas, 2024).

Baer (2019, p. 53-57) trabalha em seu livro com o objetivo de explicar como os algoritmos podem distorcer a tomada de decisão, destacando como os algoritmos funcionam, de modo a possibilitar também o entendimento do ramo de Aprendizagem de Máquina (*Machine Learning*). O autor busca explicações sobre como os preconceitos do mundo real são refletidos pelos algoritmos, sendo tais explicações diversas e complexas, mas podendo ser sumarizadas partindo-se da premissa que se os vieses no mundo real criam sua própria realidade, as técnicas estatísticas perderão seu poder de remover tais vieses por si mesmas, de modo que em tais situações, o algoritmo, sem dúvida, se tornará cúmplice em perpetuar vieses e, assim, ter ramificações cada vez mais profundas na realidade. No entanto, mesmo um algoritmo tendencioso pode ser um mal menor

quando comparado ao julgamento humano que implica vieses ainda piores. Lembre-se do que Ross (2020) afirmou e representou por meio da Figura 1. E, a longo prazo, os algoritmos podem se tornar a solução para vieses do mundo real.

No que se refere aos vieses a partir dos cientistas de dados, desenvolvedores e programadores, Baer (2019, p. 59-68) afirma que o viés de confirmação afeta tanto o projeto (*design*) do modelo quanto a amostragem de dados, sendo que ao ocorrer no projeto do modelo pode comprometer a escolha de variáveis dependentes e independentes. Já na amostragem, pode causar a escolha e organização de um conjunto incompleto de dados que carece de observações que desafiariam a hipótese do cientista de dados. Isto torna o treinamento do modelo inadequado e com consequências adversas futuras. Outro ponto relevante para Baer (2019) é o esgotamento do ego (*ego depletion*), ou seja, um cansaço mental que pode ser causado pelo fato do cientista de dados ter que tomar um número excessivo de microdecisões e, gradualmente, acaba por introduzir vieses como uma forma de minimizar o esforço cognitivo, por isso, a motivação de vieses prejudiciais tem mais probabilidade de afetar o trabalho do cientista de dados em um estado de fadiga mental. Por outro lado, o excesso de confiança faz com que os cientistas de dados rejeitem sinais de que o modelo pode ser tendencioso, mesmo na ausência de esgotamento do ego.

É importante destacar que vieses reduzem a precisão dos modelos em sistemas de IA e, portanto, reduzem o seu potencial, seja de aplicabilidade, confiança, transparência ou responsabilidade. E, os riscos em sistemas enviesados advêm, portanto, seja da base de dados ou dos algoritmos, por não serem representativos da sociedade, podendo ser super ou sub-representativo de certas identidades de maneira a não refletir a sociedade. Soma-se a lista de riscos, outro que pode ser combinado com o anterior e refere-se ao fato de se ter sistemas precisos, mas representativo de uma sociedade injusta, por exemplo, preconceitos contra certos grupos que refletem preconceitos/discriminação já existentes (Smith; Rustagi, 2020, p. 23). Os vieses estarão escondidos em sistemas precisos. Por isso, Juarez Freitas e Thomas Bellini Freitas (2020, p. 94) apontam que o "resgate da liberdade em relação aos vieses é crucial para a tutela dos direitos

fundamentais em situação de pronunciado risco, tendo de mobilizar a colaboração útil da própria IA.", ou seja, busca-se por explicabilidade.

Ao avaliar os riscos advindos de algoritmos enviesados, Baer (2019, p. 118) aponta que os danos causados devem ser analisados sob uma perspectiva ética, mas também quantificados economicamente a partir de 03 fontes: riscos legais, riscos reputacionais e riscos de negócio.

E de maneira contundente D'Ignazio (2020) afirma que o risco associado aos dados está em que "*Data is never this raw, truthful input and never neutral. It is information that has been collected in certain ways by certain actors and institutions for certain reasons*."[19]. Em complemento, Filimowicz (2022, p. i) aduz que "*The output of computational models is directly tied not only to their inputs but to the relationship and assumptions embedded in their model design, many which are of a social and cultural, rather than physical and mathematical, nature*."[20]. Portanto, não somente os dados, mas os relacionamentos entre sub-conjuntos de dados, bem como de algoritmos. E compreender tudo isto em um cenário de não-coisas é ainda mais complexo. Por isso resolver a complexidade é a chave para a explicabilidade, a qual, por sua vez destrancará o problema do enviesamento. Não só eliminar o enviesamento, o que pode ser uma tarefa hercúlea, mas estar atento a sua provável existência. Como mencionado anteriormente, os vieses não podem estar escondidos em sistemas precisos.

Svensson (2020, p. 21-39) afirma e questiona: "*AI is just statistics, just really advanced mathematics. But must magic and mathematics be in opposition to each other*?"[21] O autor discute sobre valores e vieses na cultura tecnológica, explicando que existem dois polos: os desenvolvedores de *software* que estão por detrás das telas/monitores e os usuários que estão

---

[19] Tradução livre: dados nunca são uma entrada crua e verdadeira e nunca são neutros. São informações que foram coletadas de certas maneiras por certos atores e instituições por certas razões.

[20] Tradução livre: A saída (*outputs*) dos modelos computacionais está diretamente ligada não apenas às suas entradas (*inputs*), mas também ao relacionamento e às suposições incorporadas no *design* do modelo, muitas das quais são de natureza social e cultural, em vez de física e matemática.

[21] Tradução livre: IA é apenas estatística, apenas matemática realmente avançada. Mas mágica e matemática devem estar em oposição uma à outra?

em frente às telas/monitores. E a mágica? O autor responde que "*The magic of tech consists of programming languages: binary machines code of ones and zeroes that search and find hidden patterns in so-called big data and perform tricks such as explaning and predicting future behavoir.*"[22]. O autor conceitua a existência da *mathemagics*, visto que a mágica da tecnologia está na mágica matemática, possibilitando a criatividade, o pensar fora da caixa, mesmo que racionalmente e com base na lógica numérica. Em verdade a mágica e a matemática não estão em oposição, mas em composição. É a *mathemagics* que está no centro da cultura tecnológica incluindo seu fascínio, vieses e contradições. E a combinação da matemática com a mágica se faz cada vez mais presente nas empresas de tecnologia. É a combinação de interesses privados na promoção de certos sites e o status de monopólio de um número relativamente pequeno de empresas que ofertam serviços, a exemplo dos mecanismos de busca na Internet, que leva a algoritmos de busca tendenciosos que privilegiam a branquitude e discriminam pessoas que se autodeclaram pretos ou pardos (para o IBGE[23] a população negra é o somatório de quem se autodeclara como preto ou pardo). De fato, como a tecnologia é projetada e desenvolvida por humanos, afirma Svensson, ela opera com preconceitos, assim como o resto da sociedade.

Svensson (2020, p. 27) exemplifica a mágica da tecnologia por meio de unicórnios ou as *startups* que alcançam o valor de 1 bilhão de dólares ou mais como símbolos de organização de tecnologia que tentam tornar o mundo um lugar melhor. Tais símbolos fazem parte do espaço representativo da mágica da tecnologia. Para o autor, outros exemplos são a Realidade Virtual (RV) e a Realidade Aumentada (RA) que parecem tornar real o fascínio do ser humano pela tecnologia e pela curiosidade sobre o funcionamento do cérebro humano: "*detach yourself from humanity and some of the shortcomings of being human*."[24]. O autor conclui que a *mathemagics* está na possibilidade do impos-

[22] Tradução livre: A magia da tecnologia está na linguagens de programação: máquinas binárias de código de uns e zeros que buscam e encontram padrões ocultos no *Big Data* e realizam truques como explicar e prever comportamentos futuros.

[23] Disponível em: https://biblioteca.ibge.gov.br/visualizacao/periodicos/3105/cd_2022_etnico_racial.pdf

[24] Tradução livre: desapegue-se da humanidade e de algumas das deficiências de ser humano.

sível, na possibilidade de resolver todos os tipos de problemas (encontrando soluções tecnológicas e computacionais) e na necessidade de enfrentar o mundo real com seus enviesamentos, os quais se replicam na tecnologia. Não se pode elevar a tecnologia ao fascinante, caso contrário os desenvolvedores serão os mágicos. Em verdade, a mágica deve estar na natureza humana, na imprevisibilidade e na unicidade de ser um humano.

Então, Bolukbasi et al. (2016) perguntam: "*Man is to Computer Programmer as Woman is to Homemaker*?"[25]. A pergunta é provocativa e os autores discutem o enviesamento partindo da premissa que a aplicação cega de técnicas de Aprendizagem de Máquina apresenta o risco de amplificar vieses presentes nas bases de dados. Por meio de protocolo experimental, os autores demonstram, utilizando artigos do Google Notícias, que o treinamento de modelos a partir de palavras por meio de técnicas de Processamento de Linguagem Natural exibe estereótipos de gênero feminino/masculino em uma extensão perturbadora. O objetivo dos autores é mostrar que existem técnicas para apresentar aos algoritmos de treinamento exemplos com linguagem neutra em relação ao gênero e, assim, remover estereótipos de gênero, como a associação entre as palavras recepcionista e feminino, enquanto são mantidas associações desejadas, por exemplo entre as palavras rainha e feminino.

Outro estudo relacionado com a discriminação de gênero foi realizado por Buolamwini e Gebru (2018), de modo a demonstram que algoritmos de Aprendizagem de Máquina podem discriminar com base em classes como raça e gênero. Os autores apresentam uma abordagem para avaliar o viés presente em algoritmos e conjuntos de dados de análise facial automatizada com relação a subgrupos fenotípicos. Usando o sistema de classificação Fitzpatrick Skin Type[26] aprovado por dermatologistas, os autores caracterizaram a distribuição de gênero e tipo de pele em duas bases de dados de referência para análise facial: IJB-A[27] (com 11.754 ima-

---

[25] Tradução livre: O homem é para o programador de computador o que a mulher é para a dona de casa?

[26] Disponível em: https://www.arpansa.gov.au/sites/default/files/legacy/pubs/RadiationProtection/FitzpatrickSkinType.pdf e https://www.healthline.com/health/beauty-skin-care/fitzpatrick-skin-types#Alternatives

[27] Disponível em: https://paperswithcode.com/dataset/ijb-b

gens, 55.025 frames e 7.011 videos) e Adience[28] (com 26.580 imagens). Os resultados apontam que os dados são compostos predominantemente por indivíduos de pele mais clara: 79,6% para IJB-A e 86,2% para Adience. Deste modo, os autores introduziram um novo conjunto de dados de análise facial equilibrado por gênero e tipo de pele. E, procederam a análise de 03 (três) sistemas comerciais de classificação de gênero usando o conjunto de dados reequilibrado/ Concluíram que mulheres de pele mais escura são o grupo mais mal classificado, com taxas de erro de até 34,7%. A taxa máxima de erro para homens de pele mais clara é de 0,8%. As disparidades substanciais observadas demonstram que sistemas de classificação de gênero exigem urgentemente atenção se empresas comerciais desejam desenvolver algoritmos de análise facial genuinamente justos, transparentes e responsáveis.

West et al. (2019, p. 3-5) apresentam em relatório que o setor de IA precisa de uma mudança profunda na forma como aborda a atual crise de diversidade, uma vez que a indústria de sistemas de IA precisa reconhecer a gravidade do problema de diversidade e admitir que os métodos existentes falharam em lidar com a distribuição desigual de poder e os meios pelos quais a IA pode reforçar tal desigualdade. Além disso, as autoras mostraram que os vieses em sistemas de IA refletem padrões históricos de discriminação, portanto, a não diversidade e os vieses são manifestações de um mesmo problema sócio-cultural e devem ser abordadas em conjunto. As autoras reconhecem que corrigir o viés em sistemas de IA é quase impossível quando esses sistemas sofrem de obscuridade ou opacidade. E, ponderam que a transparência é essencial e começa com o rastreamento e a divulgação de onde os sistemas de IA são usados e para qual propósito. Propõem que testes rigorosos devem ser exigidos em todo o ciclo de vida dos sistemas de IA aplicados em domínios sensíveis. Aqui entende-se, tal qual o artigo 20 da LGPD, sistemas que envolvam aspectos pessoal, profissional, de consumo e de crédito ou aspectos de personalidade (BRASIL, 2018). Há também que se realizar testes de pré-lançamento, auditoria independente e monitoramento contínuo para constatar viés,

[28] Disponível em: https://paperswithcode.com/dataset/adience

discriminação e outros problemas. E as pesquisas sobre viés e justiça (*fairness*) precisam ir além da eliminação de viés técnico para incluir uma análise social mais ampla de como a IA é aplicada em cada contexto, exigindo conhecimentos transdisciplinares. E, finalmente, as autoras recomendam que métodos para abordar viés e discriminação em sistemas de IA precisam incluir avaliações sobre se certos sistemas devem ser efetivamente projetados, com base em uma avaliação de risco completa. Eis aqui a necessidade de avaliação de riscos.

Baer (2019, p. 129-171) trata sobre estratégias gerenciais para corrigir viés algorítmico observando três problemas e apontando soluções: (i) como detectar vieses em algoritmos, (ii) estratégias gerenciais para corrigir viés algorítmico e (iii) como gerar uma base de dados sem enviesamentos. A detecção de vieses exige um constante monitoramento do funcionamento do algoritmo com a finalidade de alertar sobre problemas, de modo a realizar três etapas: a) definir a métrica que se deseja analisar estabelecendo as faixas (range) entre o que é "normal" para aquela situação; b) construir o procedimento ou rotina que será executada periodicamente para calcular a métrica estabelecida; c) avaliar se o resultado da métrica está fora dos limites de normalidade e decidir o que precisa ser feito. Para tal, o autor recomenda uma adequada análise de distribuição, em termos estatísticos, aplicando, por exemplo, o teste de significância estatística (Spiegel, 1978, p. 299-368).

Em relação às estratégias gerenciais, Baer (2019, p. 165-166) recomenda ajustar a arquitetura de decisão para que tendenciosos possam ser controlados e os resultados das decisões restringidos, por meio, por exemplo, de substituições, de partes do algoritmo, destinadas a reduzir ou eliminar o viés. Para remover vieses em bases de dados, o autor recomenda tornar explícito o viés introduzindo indicadores para as fontes de enviesamento, por exemplo, variáveis explicativas, as quais são instanciadas com um mesmo valor para que assim permaneçam durante a execução do código-fonte implementado a partir do algoritmo em questão. Caso, essas variáveis sofram alguma alteração de valor, explicita-se o problema e sabe-se em que parte do algoritmo isso está ocorrendo. Além

disto, simulações são essenciais para oportunizar a necessidade de ajustes nas regras de decisão.

E, finalmente, Baer (2019, p. 171) apresenta reflexões sobre como gerar uma base de dados sem enviesamentos, iniciando por executar um processo cuidadoso de *design* e *redesign* de modelo, preferencialmente realizando um projeto piloto executado a partir de uma base de dados padrão, sem enviesamentos. E, para uma base de dados não conter vieses, é necessário, como já mencionado anteriormente, que a base contenha todos os exemplos necessários à representatividade do problema em questão ou pode-se trabalhar com uma base de dados aleatórios, ou seja, os dados não devem descrever um padrão determinístico, mas uma distribuição de probabilidade, por exemplo, uma distribuição normal (Kazmier, 1982, p. 115-118).

Cabe destacar que Baer (2019, p. 171) indica que "*If humans are involved in the generation or collection of the data, executing such na unbiased data collection is a challenging effort because it require complete compliance by the front--line*."[29] E, são muitas as conformidades a serem atendidas nos dias atuais, sendo tal prática importantíssima: "*And only organizations that embed the regular generation of fresh, unbiased data into their "business as usual" operations can ensure their mine od data gold never is polluted by biases*."[30] Ross (2020, p. 125-142) defende a ideia de se criar organizações mais conscientes, uma vez que qualquer viés individual também pode ser algo coletivo em uma organização, sendo facilmente justificado pela frase: "*The way we do things around here*."[31], o que, claramente, não é aceitável e nem pode justificar um passado construído a partir de enviesamentos humanos e não poderá ser a justificativa do futuro para os algoritmos e a Inteligência Artificial.

Portanto, são muitos os reflexos no Direito advindo de enviesamentos em sistemas de IA. Esta última seção explicou e mencionou alguns destes reflexos a saber: ausência de diversidade, equidade e inclusão, de

---

[29] Tradução livre: Se houver humanos envolvidos na geração ou coleta de dados, executar uma coleta de dados imparcial é um esforço desafiador porque exige total conformidade de quem executa essa tarefa.

[30] Tradução livre: E somente organizações que incorporam a geração regular de dados novos e imparciais em suas operações "normais" podem garantir que sua mina de ouro de dados nunca seja poluída por enviesamentos.

[31] Tradução livre: A maneira como fazemos as coisas por aqui.

modo a gerar discriminação, especialmente em grupos minoritários ou vulneráveis. Além é claro de alocação injusta de recursos, riscos à saúde, violação de liberdades civis, e não respeito aos direitos fundamentais a partir de sistemas depreciativo e reforço a preconceitos.

Há que se ter em mente que vieses reduzem a precisão dos modelos em sistemas de IA e, portanto, reduzem o seu potencial, seja de aplicabilidade, confiança, transparência ou responsabilidade. E, os riscos em sistemas de IA enviesados advêm, seja da base de dados ou dos algoritmos, não-coisas, por não serem representativos da sociedade, podendo ser super ou sub-representativo de certas identidades de maneira a não refletir a sociedade. Assim não há como escapar da explicabilidade em sistemas de IA (IA *Explicável ou AI Explainable ou XAI*) (Freitas, 2024) associada a avaliação de riscos (Freitas, 2023).

CONSIDERAÇÕES FINAIS

Decorrente do que foi apresentado e discutido, conclui-se que:

- Os sistemas de IA precisam contar com mecanismos para garantir a prestação de contas, auditabilidade e responsabilização (*accountability*);
- Os sistemas de IA devem ser embasados em critérios de justiça e equidade (*fairness*) que precisam ser aplicados aos conjuntos de dados coletados e mantidos de forma responsável pelo ente privado ou público que faz uso de tais sistemas;
- Direito Constitucional: há que se refletir sobre os efeitos dos sistemas de IA na democracia e no Estado de Direito;
- Direitos Humanos: os sistemas de IA devem atender à noções gerais normativas e éticas de justiça, equidade e justiça social, tomando por base leis relacionadas aos Direitos Humanos, de modo a garantir igualdade, não discriminação e inclusão;
- Direitos Fundamentais: os cidadãos devem ter controle total sobre seus dados pessoais, ao passo que seus dados pessoais (incluindo-se os dados biométricos e genéticos) não devem ser

usados para causar danos ou discriminação, prevalecendo a privacidade e a governança de dados (*compliance*);

- A governança e proteção de dados é e será cada vez mais uma responsabilidade crucial aos entes estatais e privados, visto que sempre houve o entendimento que os aspectos jurídicos não acompanham os aspectos tecnológicos. Todo o ecossistema de inovação de IA precisa entender o que significa usar tais tecnologias a partir de diretrizes ético-jurídicas, visto que a IA não tem boas ou más intenções, nem mesmo conta com a capacidade de fazer julgamentos. A responsabilidade sobre essas habilidades são exclusivas de quem adota e faz uso de sistemas de IA;
- Ao se considerar que existe um ciclo de vida do projeto de sistemas de IA, esse ciclo deve influenciar as estruturas ou instituições econômicas, legais, culturais e políticas, para que, por sua vez, as normas, procedimentos e políticas públicas sejam também influenciadas por ações e decisões em todo o ecossistema de inovação de IA.

Finalmente, o ecossistema de inovação de IA afetará Estado, empresas, pessoas e sociedade, portanto os aspectos jurídicos não podem deixar de lado os aspectos tecnológicos, esquecendo-se normas técnicas (ISO/IEC ou ABNT) ou modelos de boas práticas com base na Segurança da Informação ou riscos.

REFERÊNCIAS

BAER, Tobias. Understand, manage and prevent algorithmic bias. A guide for businnes users and data scientists. Germany: Apress Media, 2019.

BOLUKBASI, Tolga; CHANG, Kai-Wei; ZOU, James; SALIGRAMA, Venkatesh; KALAI, Adam. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Proc. of 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 2016. Disponível em: https://papers.nips.cc/paper/6228-man-is-tocomputer-programmer-as-woman-is-tohomemaker-debiasing-word-embeddings.pdf Acesso em: 14 abr. 2025.

BRASIL. **Lei 13.709, de 14 de agosto de 2018**, Lei Geral de Proteção de Dados - LGPD, 2018.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proc. of Machine Learning Research, Conference on Fairness, Accountability, and Transparency, v. 81, 2018. p. 1-15. Disponível em: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf Acesso em: 14 abr. 2025.

COLLETT, Clementine; DILLON, Sarah. AI and Gender - Four Proposals for Future Research. Cambridge: The Leverhulme Centre for the Future of Intelligence, 2019. Disponível em: https://api.repository.cam.ac.uk/server/api/core/bitstreams/04e69f8b-ccf1-4c3a-affd-8c28ad273873/content Acesso em: 14 abr. 2025.

DOOLEY, Roger. **Como Influenciar a Mente do Consumidor**: 100 maneiras de convencer os consumidores com técnicas de neuromarketing. Trad. Luciene Scalzo. São Paulo: Elsevier, 2012.

DWORK, Cynthia, HARDT, Moritz; PITASSI, Toniann; REINGOLD, Omer; ZEMEL, Rich. Fairness through awareness. *In:* **Proceedings of the 3rd Innovations in Theoretical Computer Science Conference** (ITCS12), 2012. p. 214-226. Disponível em: https://doi.org/10.1145/2090236.2090255 Acesso em: 14 abr. 2025.

D'IGNAZIO, Catherine. Data is never a raw truthful input and it is never neutral. Entrevista para Zoe Corbyn, **The Guardian**, 2020. Disponível em: https://www.theguardian.com/technology/2020/mar/21/catherine-dignazio-data-is-never-a-raw-truthful-input-and-it-is-never-neutral Acesso em: 14 abr. 2025.

FERRARA, Emilio. **Fairness and Bias in Artificial Intelligence**: a brief survey of sources, impacts, and mitigation strategies. 2023. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4615421 Acesso em: 14 abr. 2025.

FILIMOWICZ, Michael. Systemic Bias – algorithm and Society. Routledge Taylor & Francis Group: London, 2022.

FREITAS, Cinthia Obladen de Almendra. O Direito e a Inteligência Artificial como N**ão-**Coisa. **CONPEDI Law Review**, XIII Encontro Internacional do CONPEDI Uruguai – Montevidéu, v. 10, n. 1, pp. 88 – 109, jul.-dez., 2024.

FREITAS, Cinthia Obladen de Almendra. Riscos e Proteção de Dados Pessoais. **RRDDIS - Revista Rede de Direito Digital, Intelectual & Sociedade**, v. 2, p. 225-247, 2023.

FREITAS, Cinthia Obladen de Almendra; BATISTA, Osvaldo Henrique dos Santos. Neuromarketing e as Novas Modalidades de Comércio Eletrônico (m-s-t-f--commerce) frente ao Código de Defesa do Consumidor. **Derecho y Cambio Social,** v. 42, 2015, p. 1-22.

FREITAS, Cinthia Obladen de Almendra; BARDDAL, Jean Paul. Análise preditiva e decisões judiciais: controvérsia ou realidade?. **DEMOCRACIA DIGITAL E GOVERNO ELETRÔNICO**, v. 1, p. 107-126, 2019.

HAN, Byung-Chul. **Não-coisas**: transformações no mundo em que vivemos. Trad. Ana Falcão Bastos. Lisboa: Relógio D'Água Editores, 2022.

MIN, Alfonso. **Artificial Intelligence and Bias**: challenges, implications, and remedies. 2023. Disponível em: https://www.researchgate.net/publication/374232878_ARTIFICIAL_INTELLIGENCE_AND_BIAS_CHALLENGES_IMPLICATIONS_AND_REMEDIES Acesso em: 14 abr. 2025.

ROSS, Howard J. Everyday bias: identifying and navigating unconscious judgments in our daily lives. London: The Rowman & Littlefield Publishing Group, Inc., 2020.

SMITH, Genevieve; RUSTAGI, Ishita. Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook. **Berkeley Haas Center for Equity**, Gender and Leadership, July, 2020. Disponível em: https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf Acesso em: 14 abr. 2025.

SPIEGEL, Murray R. **Probabilidade e estatística**. Coleção Schaum. Trad. Alfredo Alves de Farias. São Paulo: McGraw-Hill do Brasil, 1978.

SVENSSON, Jakob. Modern Mathemagics: values and biases in tech culture. In: **Systemic Bias – algorithm and Society**. Series Editor: Michael Filimowicz, Routledge Taylor & Francis Group: London, 2022.

WEST, Sarah Myers; WHITTAKER, Meredith; CRAWFORD, Kate. Discriminating Systems: Gender, Race, and Power in AI. **AI Now Institute**, 2019. Disponível em: https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2 Acesso em: 14 abr. 2025.

KAZMIER, Leonard J. **Estatística Aplica à Economia e Administração**. Trad. Carlos Augusto Crusius; Revisão Técnica Jandyra M. Fachel. São Paulo: Pearson Makron Book, 1982.

Recebido em 17 de maio de 2025
Aprovado em 25 de junho de 2025