PRECONCEITO CODIFICADO: O VIÉS ALGORÍTMICO E SEUS IMPACTOS NA DESIGUALDADE DE GÊNERO

CODED BIAS: ALGORITHMIC DISCRIMINATION AND ITS IMPACTS ON GENDER INEQUALITY

Silvia Carlos da Silva Pimentel[1]

Raissa Amarins Marcandeli[2]

Resumo: Este estudo analisa como os algoritmos de inteligência artificial (IA) reproduzem padrões discriminatórios de gênero e seus impactos sociais. A crescente adoção da IA em setores como mercado de trabalho, saúde e segurança pública tem levantado preocupações sobre a perpetuação de desigualdades estruturais, especialmente contra as mulheres. A pesquisa, baseada em revisão de literatura e análise de casos, identifica setores mais suscetíveis à discriminação algorítmica e avalia práticas internacionais de mitigação. Os resultados indicam que o viés algorítmico decorre de dados de treinamento desbalanceados, falta de diversidade nas equipes de desenvolvimento e ausência de regulação adequada.

Palavras-chave: Inteligência Artificial; Viés Algorítmico; Desigualdade de Gênero; Governança Algorítmica. Discriminação Algorítmica.

Abstract: This study examines how artificial intelligence (AI) algorithms reproduce gender-based discriminatory patterns and their social impacts. The growing adoption of AI in sectors such as labor markets, healthcare, and public security raises concerns about the perpetuation of structural inequalities, particularly against women. Based on a literature review and case studies, the research identifies sectors most susceptible to algorithmic discrimination and assesses international mitigation practices. The findings indicate that algorithmic bias stems from unbalanced training data, lack of diversity in development teams, and the absence of adequate regulation.

Keywords: Artificial Intelligence; Algorithmic Bias; Gender Inequality; Algorithmic Governance. Algorithmic Discrimination.

1 Doutora em Direito - Pontifícia Universidade Católica de São Paulo; Professora Doutora - Pontifícia Universidade Católica de São Paulo; E-mail: spimentel@pucsp.br; ORCID: https://orcid.org/0000-0002-7171-5869.
2 Mestre em Sistema Constitucional de Garantia de Direitos - Centro Universitário de Bauru (ITE/Bauru); Professora Universitária – Faculdade Nove de Julho; E-mail: raissaamarins@hotmail.com; ORCID: https://orcid.org/0000-0002-0710-8677.

Introdução

A crescente integração da inteligência artificial em diversas áreas da sociedade tem levantado questões fundamentais sobre seus impactos sociais, especialmente no que se refere à perpetuação da desigualdade de gênero. Dessa maneira, este estudo tem como objetivo investigar de que forma os algoritmos de inteligência artificial (IA) reproduzem padrões discriminatórios contra as mulheres e suas consequências sociais, considerando que atualmente apenas 22% dos profissionais do setor são mulheres, segundo dados do Fórum Econômico Mundial (2024), o que pode contribuir significativamente para vieses nos sistemas desenvolvidos.

A relevância deste tema é evidenciada pelo uso exponencial de algoritmos em setores críticos da sociedade. No mercado de trabalho, sistemas de recrutamento baseados em IA podem reproduzir e amplificar preconceitos existentes no processo de seleção. Na segurança pública e redes sociais, decisões automatizadas têm demonstrado impactar diferentes grupos demográficos de maneira desproporcional, com evidências documentadas de discriminação em análises de crédito e sistemas de reconhecimento facial.

A metodologia adotada combina análise extensiva de literatura especializada com estudos de caso representativos, permitindo identificar padrões de discriminação algorítmica e avaliar estratégias internacionais bem-sucedidas para sua mitigação. Os objetivos específicos deste trabalho compreendem: examinar como dados de treinamento influenciam o viés algorítmico, com foco particular em conjuntos de dados históricos que refletem discriminação preexistente; identificar setores críticos afetados pela discriminação algorítmica de gênero; analisar a responsabilidade corporativa e técnica no desenvolvimento de sistemas de IA; e investigar soluções práticas, como auditorias algorítmicas e diversificação de bases de dados.

Este artigo está organizado em três seções principais. A primeira seção estabelece os fundamentos da IA e sua relação com processos decisórios automatizados. A segunda examina o viés algorítmico e sua influência específica na discriminação contra mulheres. A terceira aborda mecanismos regulatórios e de governança algorítmica essenciais para promover equidade de gênero. A conclusão, por sua vez, sintetiza as descobertas principais e apresenta recomendações práticas para reduzir o impacto negativo da IA na desigualdade de gênero.

1 Inteligência Artificial e a tomada de decisão algorítmica

Segundo McKinsey & Company (2024), o termo "inteligência artificial" foi cunhado pelo cientista da computação John McCarthy em 1956, durante um *workshop* em Dartmouth. No entanto, o conceito já havia sido discutido anteriormente por Alan Turing, que introduziu, em 1950, a ideia do "jogo da imitação", hoje conhecido como Teste de Turing, utilizado para avaliar a capacidade de uma máquina em demonstrar comportamento inteligente. Turing defendia que os pesquisadores deveriam concentrar-se em áreas que não exigissem grandes capacidades de percepção e ação, como jogos e tradução de idiomas.

Desde então, diversas abordagens surgiram no campo da IA, o que engloba a inteligência artificial simbólica, predominante até o final dos anos 1980, que utilizava símbolos e lógica para resolver problemas, mas carecia da capacidade de aprendizado autônomo. Com o avanço das redes neurais, inspiradas no funcionamento do cérebro humano, a IA passou a processar grandes volumes de dados e identificar padrões de forma iterativa, permitindo o desenvolvimento de modelos mais sofisticados, como aqueles que impulsionam a inteligência artificial generativa (McKinsey & Company, 2024).

A inteligência artificial e a tomada de decisão algorítmica, por sua vez, representam avanço significativo na capacidade dos sistemas computacionais de processar informações e realizar escolhas racionais com base em grandes volumes de dados. A abordagem da IA para a tomada de decisão baseia-se, em grande parte, na teoria da decisão e em modelos probabilísticos, permitindo que agentes inteligentes avaliem múltiplas possibilidades e escolham ações que maximizem um determinado objetivo (Russell; Norvig, 2010).

A utilização de redes bayesianas[3], por exemplo, possibilita um raciocínio importante sobre a incerteza, combinando elementos da IA simbólica e das redes neurais para melhorar a precisão das inferências algorítmicas. Além disso, o conceito de agentes inteligentes vem sendo amplamente estudado, pois engloba sistemas capazes de operar autonomamente, interagindo com o ambiente e aprendendo com novas experiências.

3 "Redes Bayesianas (RB) são mecanismos eficientes para análise de dados que apresentam relacionamentos de precedência temporal. Uma RB tem dois componentes: uma estrutura gráfica e os parâmetros numéricos. Tanto a estrutura gráfica quanto os parâmetros numéricos podem ser aprendidos automaticamente de uma base de dados. Este trabalho dá a base teórica e propõe um conjunto de algoritmos e estrutura de dados que permitem manipular grandes volumes de dados no processo de aprendizagem de uma RB. Enfoca a limitação teórica dos processos de aprendizagem e detalha o principal algoritmo de aprendizagem Bayesiana que explora o método de independência condicional" (Plentz, 2003, p. 26).

O uso de inteligência artificial na tomada de decisão algorítmica está presente em diversas áreas, como finanças, saúde e segurança, onde decisões automatizadas precisam ser eficientes e confiáveis. No entanto, questões éticas e desafios técnicos permanecem, especialmente no que diz respeito à transparência dos algoritmos e à possibilidade de viés nas decisões computacionais.

A evolução dos sistemas de IA tem sido acompanhada por debates sobre sua aplicabilidade e confiabilidade, sendo necessário um desenvolvimento contínuo de metodologias que garantam uma tomada de decisão algorítmica justa e ética. Russell e Norvig (2010, p. 02) apresentam uma tabela com quatro categorias de definições de inteligência artificial, organizadas em dois eixos: pensamento *versus* ação e humano *versus* racional:

Thinking Humanly	**Thinking Rationally**
"The exciting new effort to make computers think . . . *machines with minds*, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly	**Acting Rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole *et al.*, 1998) "AI . . . is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Figure 1.1 Some definitions of artificial intelligence, organized into four categories.

Na primeira parte, "pensando como humanos", Haugeland (1985) descreve o avanço significativo na tentativa de desenvolver computadores capazes de pensar, considerando a possibilidade de que as máquinas possuam mentes no sentido pleno e literal. De forma semelhante, Bellman (1978) destaca que a automação busca reproduzir atividades tradicionalmente associadas ao pensamento humano, tais como a tomada de decisão, a resolução de problemas e o aprendizado.

Na segunda, "agindo como humanos", Kurzweil (1990) define a inteligência artificial como a criação de máquinas capazes de executar funções que,

quando desempenhadas por seres humanos, exigem inteligência. De modo semelhante, Rich e Knight (1991) conceituam a área como o estudo voltado para o desenvolvimento de computadores que possam realizar tarefas em que, atualmente, os seres humanos ainda apresentam desempenho superior.

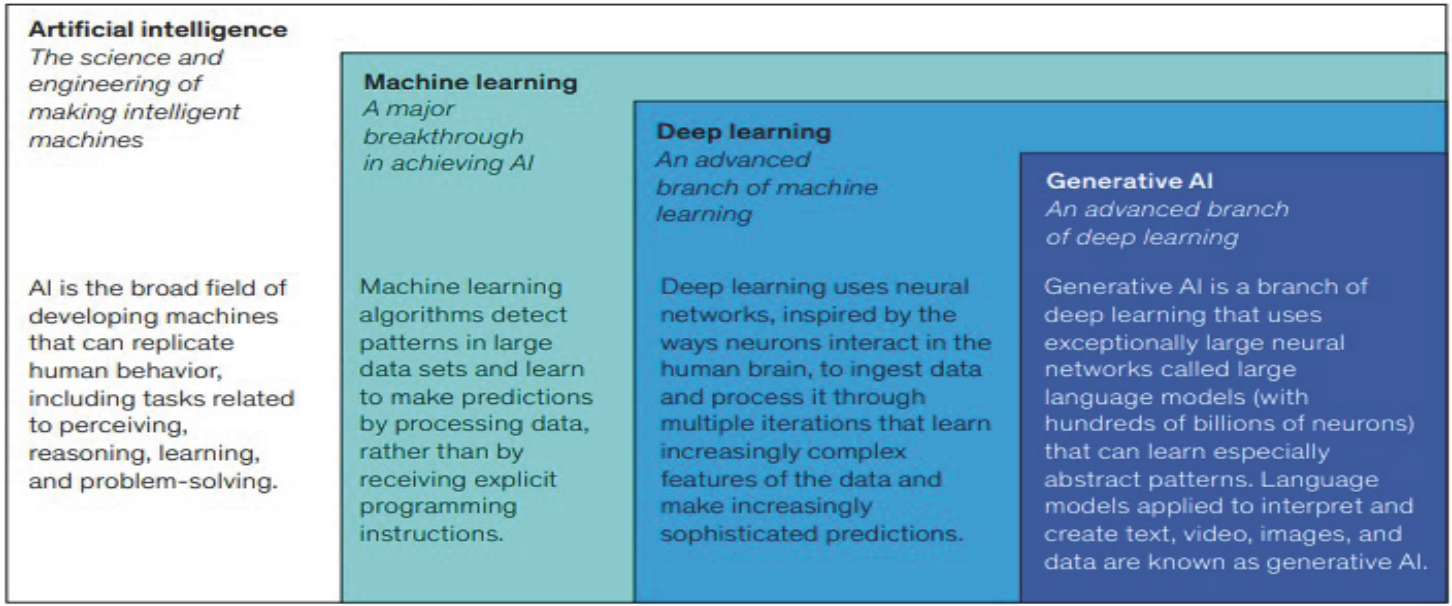
Na terceira, "pensando racionalmente", Charniak e McDermott (1985) definem a inteligência artificial como a investigação das faculdades mentais humanas por meio da utilização de modelos computacionais. De forma complementar, Winston (1992) caracteriza essa área de estudo como a análise dos processos computacionais que possibilitam a percepção, o raciocínio e a ação. Na última, "agindo racionalmente", Poole *et al.* (1998) conceituam a Inteligência Computacional como a área dedicada ao estudo e desenvolvimento de agentes inteligentes. Nessa mesma linha, Nilsson (1998) destaca que a inteligência artificial tem como foco a análise e implementação de comportamentos inteligentes em artefatos tecnológicos.

A inteligência artificial, em suas aplicações contemporâneas, é estruturada a partir de modelos que buscam simular a racionalidade humana em contextos de incerteza. De acordo com Russell e Norvig (2010), a IA pode ser compreendida como a construção de agentes que tomam decisões racionais, ou seja, capazes de escolher ações que maximizem determinados objetivos com base nas informações disponíveis.

A racionalidade algorítmica é orientada por dois pilares teóricos principais: a teoria da probabilidade, que estrutura as crenças de um agente sobre o ambiente, e a teoria da utilidade, que expressa suas preferências. A combinação dessas dimensões permite que algoritmos operem em contextos de incerteza, utilizando modelos matemáticos para avaliar riscos e benefícios associados a diferentes escolhas.

Sheikh, Prins e Schrijvers (2023) ressaltam a dificuldade em definir a inteligência artificial de forma precisa, uma vez que ela busca imitar ou simular um fenômeno ainda não completamente compreendido: a inteligência humana. Dessa forma, enquanto essa elucidação não for plenamente alcançada, torna-se inviável estabelecer com precisão os parâmetros para a replicação artificial da inteligência humana. McKinsey & Company (2024, p. 03) demonstra, por meio da imagem abaixo elencada, a evolução da inteligência artificial:

The evolution of artificial intelligence



A imagem acima apresentada expõe que a inteligência artificial é um campo abrangente da ciência e engenharia dedicado ao desenvolvimento de máquinas capazes de replicar comportamentos humanos, incluindo percepção, raciocínio, aprendizado e resolução de problemas. Ao longo dos anos, a IA evoluiu significativamente, impulsionada por avanços em aprendizado de máquina, aprendizado profundo e, mais recentemente, inteligência artificial generativa (McKinsey & Company, 2024).

Para além disso, a figura mostra que o aprendizado de máquina (*machine learning*) representa um marco essencial nessa evolução. Trata-se de um conjunto de algoritmos que identificam padrões em grandes volumes de dados e aprimoram suas previsões por meio do processamento dessas informações, sem a necessidade de instruções explícitas de programação (McKinsey & Company, 2024).

Dentro do aprendizado de máquina, o quadro destaca o aprendizado profundo (*deep learning*), como uma abordagem avançada baseada em redes neurais inspiradas no funcionamento do cérebro humano. Essas redes processam os dados por meio de múltiplas iterações, permitindo a identificação de padrões cada vez mais complexos e a realização de previsões altamente sofisticadas (McKinsey & Company, 2024).

No entendimento de McKinsey & Company (2024), uma das vertentes mais recentes e revolucionárias do aprendizado profundo é a inteligência artificial generativa (*generative AI*), que utiliza redes neurais extremamente avançadas, conhecidas como modelos de linguagem de grande escala (*LLMs – Large Language Models*), que podem conter centenas de bilhões de neurônios.

Afirma-se ainda que a inteligência artificial continua a evoluir, impulsionando inovações que transformam diversos setores da sociedade e redefinem a forma como interagimos com a tecnologia. Nesse sentido, em atenção ao conteúdo já abordado no presente estudo, há de se salientar que o aprendizado de máquina pode ser subdividido em duas categorias: supervisionado e não supervisionado. No primeiro, os modelos são treinados com base em correlações pré-definidas por especialistas humanos, enquanto, no segundo, os algoritmos identificam autonomamente padrões e estruturas nos dados sem intervenção prévia (Pinto, 2020).

Um avanço relevante nesse campo é o *deep learning*, que utiliza redes neurais artificiais para processar informações de maneira não linear e em múltiplas camadas, assemelhando-se ao funcionamento do cérebro humano. Essa tecnologia tem sido essencial para aplicações como reconhecimento de voz, identificação facial, tradução automática e reconhecimento de objetos,

dependendo de vastos conjuntos de dados para alcançar um desempenho eficiente (Pinto, 2020).

Sob essa perspectiva, importante compreender o conceito de algoritmos, que são conjuntos de fórmulas matemáticas compostas por uma série de instruções representadas por símbolos, os quais são processados por microprocessadores. O processamento ocorre em um ciclo contínuo de entrada (*input*) e saída (*output*) de dados, no qual as informações recebidas são analisadas e transformadas, gerando novas fórmulas que alimentam o sistema de maneira dinâmica e recorrente (Siegwart, 2004).

Segundo Lee (2020), os algoritmos de inteligência artificial dependem de três elementos fundamentais: uma grande base de dados (*big data*), alto poder de computação e a expertise de engenheiros especializados no desenvolvimento de algoritmos de IA. O aprendizado profundo (*deep learning*), ao ser aplicado na resolução de novos problemas, requer a combinação desses três fatores, sendo que, atualmente, os dados representam o aspecto central desse processo.

Em sua essência, algoritmo consiste em uma sequência de instruções estruturadas matematicamente, incluindo equações e fórmulas, com o objetivo de alcançar um resultado específico. Nesse contexto, os algoritmos são projetados para buscar métodos computacionais que possibilitem a simulação cognitiva (*cognitive simulation*), visando aproximar-se do comportamento humano na resolução de problemas e na geração de soluções (Flansinski, 2016).

Nessa perspectiva, tem-se que "A sequência de tomada de decisão através de um tratamento de dados por meio de um algoritmo depende do fornecimento de dados (*training data*, ou dados de treinamento). O sistema realizará um monitoramento de dados (*data mining* ou mineração de dados) e realizará um aprendizado para criar padrões e modelos de respostas" (Arantes, 2022, p. 83).

Portanto, as ferramentas de inteligência artificial não são neutras, pois são treinadas com base em dados que refletem padrões sociais, culturais e históricos, os quais podem conter vieses. O treinamento dos algoritmos ocorre por meio da coleta e análise de grandes volumes de dados, frequentemente oriundos de interações humanas, registros históricos e informações extraídas de diversas fontes. Dessa forma, qualquer distorção ou desigualdade presente nesses dados pode ser incorporada e replicada pela IA, o que impacta significativamente a tomada de decisões automatizadas.

A evolução da inteligência artificial e sua capacidade de tomada de decisão automatizada trouxeram avanços significativos em diversos setores, desde a economia até a segurança pública. No entanto, ao mesmo tempo em que as tecnologias prometem eficiência e imparcialidade, cresce a preocupação sobre os impactos sociais de seus modelos preditivos.

Uma das questões mais críticas nesse cenário é a reprodução e amplificação de desigualdades históricas, especialmente no que se refere à discriminação de gênero. O cenário existe porque os algoritmos de IA são treinados com dados históricos que refletem padrões sociais existentes, muitas vezes carregados de viés. Assim, para compreender como a inteligência artificial pode perpetuar ou até intensificar disparidades entre homens e mulheres, é essencial examinar as origens desses vieses algorítmicos e suas implicações diretas.

2 Viés Algorítmico com foco na discriminação contra as mulheres: origens e implicações

De acordo com a literatura, os algoritmos têm como um de seus principais objetivos a realização de previsões baseadas em probabilidades. Embora não sejam capazes de fornecer respostas absolutamente precisas para todas as questões, esses sistemas conseguem analisar os dados inseridos (*inputs*) e apresentar estimativas consistentes. Ressalta-se que a precisão desses resultados está diretamente relacionada à quantidade e à qualidade das informações fornecidas, aumentando, assim, a probabilidade de que o resultado se aproxime da realidade (Schertel; Mattiuzzo, 2019).

Nesse sentido, o aumento dos incentivos para o uso de processamento de dados por meio de algoritmos nas tomadas de decisão, aliado à acessibilidade e ao baixo custo das tecnologias que viabilizam esse processo, torna ainda mais urgente o debate sobre as implicações desses procedimentos para os indivíduos, bem como os riscos a eles associados (Schertel; Mattiuzzo, 2019), como a discriminação algorítmica. Os retromencionados autores elencam que (2019, p. 51-53):

> **Discriminação por erro estatístico** – todo e qualquer erro que seja genuinamente estatístico, abrangendo desde dados incorretamente coletados, até problemas no código do algoritmo, de forma que ele falhe em contabilizar parte dos dados disponíveis, contabilize-os de forma incorreta, etc. Basicamente, é o tipo de discriminação que decorre de um erro cometido pelos engenheiros ou cientistas de dados responsáveis pelo desenho do algoritmo; **Discriminação por generalização** – embora o modelo funcione bem e seja estatisticamente

> correto, leva a uma situação na qual algumas pessoas são equivocadamente classificadas em certos grupos. Por exemplo, se uma pessoa mora em uma vizinhança comumente associada à pobreza e o modelo não possui nenhuma outra informação além de seu endereço para decidir se ela é ou não uma boa candidata para um empréstimo, ele a classificará como pertencente a um grupo do qual ela talvez não seja parte, caso ela se apresente como um caso atípico. Isso poderia ocorrer na hipótese de essa pessoa ter uma renda superior ou inferior às pessoas de sua vizinhança, por exemplo. Desse modo, embora o algoritmo esteja correto e as informações também, ainda assim o resultado será uma generalização incorreta, na medida em que mesmo um resultado estatisticamente relevante apresentará um percentual de pessoas que não se encaixam perfeitamente naquela média. Isso se dá pela própria natureza de qualquer exercício probabilístico; **Discriminação pelo uso de informações sensíveis** – a razão pela qual consideramos esta categoria como discriminatória, embora muitas vezes seja estatisticamente correta, é porque ela se baseia em dados ou proxies legalmente protegidos. [...] **Discriminação limitadora do exercício de direitos** – novamente, aqui falamos de uma categoria que pode apresentar resultados estaticamente corretos e relevantes, mas que ainda assim consideramos discriminatória. Ao contrário da categoria anterior, o problema advém não do tipo de dado utilizado, mas da relação entre a informação utilizada pelo algoritmo e a realização de um direito. Se há uma conexão estrita entre ambos e se o direito em questão é demasiadamente afetado, provável que o uso seja discriminatório. *(Grifos nossos).*

Sob essa ótica, é possível concluir que os vieses presentes na inteligência artificial podem surgir de diferentes maneiras. Em primeiro lugar, os vieses nos dados de treinamento são um fator determinante na perpetuação de desigualdades. Se os dados utilizados para treinar os modelos forem desbalanceados ou refletirem discriminações estruturais, a IA pode reproduzir essas tendências. Um exemplo ilustrativo é o de sistemas de recrutamento automatizados que, ao serem treinados com um histórico predominantemente masculino de contratações, tendem a favorecer homens em detrimento de mulheres, reforçando desigualdades já existentes no mercado de trabalho (Borges; Filó, 2021).

A *Amazon* decidiu selecionar os melhores candidatos para vagas de engenharia de *software*. Para isso, utilizou algoritmo de aprendizado de máquina que buscava padrões nos dados históricos dos candidatos. No passado, havia mais candidatos do sexo masculino bem-sucedidos na área de engenharia de software, portanto, algumas das correlações entre as características dos candidatos e a probabilidade de sucesso estavam ligadas ao sexo, e não à

aptidão (Adams-Prassl, 2023). O sistema da empresa aprendeu rapidamente a penalizar candidaturas oriundas de duas faculdades exclusivamente femininas. Esse tipo de filtragem reforça estereótipos de gênero e reproduz desigualdades estruturais, excluindo candidatas qualificadas e perpetuando um ciclo de exclusão no ambiente corporativo. Assim, o viés algorítmico, ao invés de neutralizar a discriminação, a automatiza em escala ampliada e invisível.

Nessa acepção, considerando padrões históricos de inscrição em cursos, um *chatbot* pode acabar reproduzindo vieses existentes ao recomendar cursos. Por exemplo, se historicamente alunas demonstraram menor propensão a se inscrever em cursos de Ciência da Computação, o sistema pode, com base nesses dados, sugerir preferencialmente cursos de Humanidades para estudantes do sexo feminino, o que evidencia como algoritmos podem perpetuar desigualdades de gênero, ao basear suas recomendações em tendências passadas sem considerar a individualidade e o potencial de cada estudante (Williams, 2024).

Além disso, há vieses na curadoria dos dados, decorrentes das escolhas feitas pelos desenvolvedores no momento da coleta e organização das informações. A ausência de representatividade em determinados conjuntos de dados pode levar à marginalização de grupos sociais minoritários, tornando a IA incapaz de reconhecer e atender adequadamente essas populações.

Nesse sentido, a título de exemplo, Davis, Williams e Yang (2021) destacam que a força de trabalho em tecnologia, incluindo os trabalhadores da empresa analisada, era composta desproporcionalmente por homens. Como consequência, o sistema de contratação, ao ser alimentado com esses dados históricos, aprendeu a considerar os homens como candidatos preferenciais, reproduzindo o viés de gênero existente no setor.

Da mesma forma, os vieses algorítmicos podem emergir mesmo quando os dados são aparentemente equilibrados, uma vez que os modelos estatísticos utilizados na tomada de decisão podem amplificar padrões de maneira desproporcional, favorecendo certos grupos em detrimento de outros. Outro aspecto relevante refere-se aos vieses de interpretação, que dizem respeito ao modo como as decisões geradas pela inteligência artificial são aplicadas no contexto social. Por conseguinte, ferramentas utilizadas para avaliação de risco em concessões de crédito, decisões judiciais ou monitoramento de segurança podem ser interpretadas e implementadas de maneira a reforçar desigualdades estruturais preexistentes.

Andrés, Gimeno e Cabo (2021) analisam as diferenças no acesso ao crédito bancário entre empreendedores homens e mulheres na Espanha.

Utilizando uma amostra de mais de 80.000 empresas iniciadas entre 2004 e 2014, os autores identificam que as mulheres empreendedoras solicitam menos crédito e, quando o fazem, enfrentam maior dificuldade para obtê-lo no primeiro ano de atividade da empresa. Além disso, os resultados indicam que empresas lideradas por mulheres apresentam menor taxa de inadimplência nos primeiros anos, o que sugere a presença de um viés inconsciente na concessão de crédito, caracterizado por padrões duplos de avaliação.

O padrão sugere a presença de viés nos sistemas de avaliação de risco, que, embora aparentemente neutros, operam com base em dados históricos e parâmetros que refletem estereótipos de gênero. Algoritmos utilizados por bancos e instituições financeiras podem estar reproduzindo decisões enviesadas ao associar, de forma implícita, maior confiabilidade e competência à figura masculina. Com isso, perpetua-se um ciclo de exclusão econômica, em que mulheres recebem menos financiamento, têm sua capacidade de expansão empresarial limitada e, assim, permanecem sub-representadas no setor, alimentando os próprios dados que sustentam o viés discriminatório.

Outrossim, pesquisas demonstram que modelos de incorporação de palavras, como o *Google* Tradutor, apresentam vieses de gênero em sua programação. Um estudo realizado pelo laboratório Fast, do Instituto de Dados da Universidade de São Francisco (EUA), exemplifica essa questão ao traduzir frases como "Ela é médica. Ele é enfermeiro" do inglês para o turco — idioma que possui um pronome singular neutro de gênero — e, em seguida, de volta para o inglês. O resultado revelou uma inversão nos papéis de gênero, com as frases retornando como "Ele é médico. Ela é uma enfermeira", evidenciando o viés algorítmico presente nesse tipo de tecnologia (Penchikala, 2018).

Programas como Word2Vec e GloVe, que são bibliotecas de recursos incorporados de palavras, também apresentam vieses de gênero. Pesquisadores identificaram uma série de preconceitos de gênero entre grupos inteiros de palavras, resultando em analogias como "pai é médico como mãe é enfermeira" e "homem é programador de computador como mulher é dona de casa", o que evidencia a reprodução de estereótipos de gênero por essas tecnologias (Penchikala, 2018).

De acordo com um estudo da UNESCO (2024), os modelos de linguagem de grande escala (*LLMs*), como GPT-3.5, GPT-2 da OpenAI e Llama 2 da Meta, apresentam evidências significativas de vieses de gênero, frequentemente associando mulheres a papéis domésticos e termos relacionados à família, enquanto vinculam homens a funções executivas e

termos ligados a carreira e negócios. O estudo também revelou a presença de conteúdos homofóbicos e estereótipos raciais, especialmente em modelos de código aberto, como o Llama 2, que geraram narrativas negativas sobre pessoas LGBTQIA+ e atribuíram ocupações estigmatizadas a grupos étnicos específicos. A UNESCO (2024) destacou a necessidade urgente de implementação de marcos regulatórios claros e monitoramento contínuo pelas empresas, a fim de mitigar esses vieses, conforme estabelecido na Recomendação sobre a Ética da Inteligência Artificial, adotada pela organização em 2021.

A ausência de uma revolução legislativa e social paralela pode levar a IA a perpetuar e até agravar as disparidades de gênero no ambiente de trabalho e na sociedade. A automação e a IA têm remodelado o mercado de trabalho, afetando especialmente mulheres sem diploma universitário, já que setores como administração, varejo e finanças, que tradicionalmente empregam muitas mulheres, estão sendo substituídos por sistemas automatizados (*Capitol Technology University,* 2024).

Além disso, a IA pode exacerbar desigualdades de gênero ao utilizar dados que refletem os vieses da sociedade, perpetuando práticas discriminatórias de contratação e ampliando a disparidade salarial, uma vez que modelos de previsão de salários baseados em dados históricos tendem a reproduzir essas diferenças. O receio de aprender a utilizar *softwares* de IA pode também limitar o potencial de desenvolvimento profissional das mulheres (*Capitol Technology University*, 2024).

Outro aspecto crítico está relacionado à segurança pessoal, com o uso de *deepfakes* – mídias manipuladas por IA – que podem colocar mulheres em situações falsas e comprometedoras, gerando violações de privacidade, danos à reputação e manipulações políticas. As mulheres de grupos marginalizados, como mulheres negras, LGBTQIA+, com deficiência e de baixa renda, enfrentam riscos ainda maiores devido a fatores demográficos e sociais interseccionais[4] (*Capitol Technology University*, 2024). Por exemplo, descobriu-se que os modelos de inteligência artificial usados na aprovação de contratações e empréstimos prejudicam desproporcionalmente as minorias raciais (Aninze, 2024).

Na área da saúde, sistemas baseados em IA podem perpetuar desinformação sobre questões de saúde feminina, conduzir a diagnósticos

4 Interseccionalidade, termo cunhado por Kimberlé Crenshaw, refere-se às maneiras pelas quais várias formas de discriminação (por exemplo, racismo, sexismo, classismo) se cruzam e resultam em experiências únicas de opressão para indivíduos com múltiplas identidades marginalizadas (Hampton, 2021). No contexto da IA, a interseccionalidade destaca como preconceitos baseados em raça, gênero e outros atributos podem interagir e resultar em discriminação agravada contra indivíduos como mulheres negras.

incorretos e minimizar sintomas devido a dados enviesados, afetando principalmente mulheres de comunidades marginalizadas, que enfrentam barreiras econômicas, digitais e linguísticas. Além disso, o uso da IA em diagnósticos de saúde reprodutiva e tratamentos de fertilidade levanta preocupações éticas relacionadas à privacidade de dados sensíveis, especialmente em contextos em que os direitos reprodutivos estão sendo limitados (*Capitol Technology University*, 2024).

De acordo com Frazão (2021), os sistemas algorítmicos têm desempenhado um papel central na formulação de julgamentos e predições sobre indivíduos, abrangendo suas características, méritos, perfis, preferências, inclinações e probabilidades em diversos contextos, desde o consumo de produtos até a reincidência criminal, bem como suas capacidades, aptidões e vulnerabilidades. Tais práticas tornam-se especialmente preocupantes quando se observa que algoritmos já demonstraram maior rigor na avaliação de indivíduos pertencentes a grupos racializados ou a mulheres, perpetuando dinâmicas discriminatórias de maneira automatizada e, muitas vezes, invisível.

Os algoritmos são frequentemente revestidos por uma aparência de neutralidade, ao passo que sua pessoalidade originária se dissipa em razão de distintos fatores tecnológicos. A dificuldade de acesso às linhas de código, seja pela exigência de conhecimentos técnicos específicos ou pelas restrições impostas por direitos de propriedade intelectual, associada às múltiplas camadas de lógicas matemáticas envolvidas nos processos de automação e análise, bem como à utilização de dados considerados "brutos", contribui para a construção de uma percepção equivocada de neutralidade e objetividade dessas tecnologias (Joyce *et al.*, 2021).

A partir da virada do milênio, as ciências humanas passaram a compreender os algoritmos não como meras ferramentas lógicas inatas e intrinsicamente objetivas, mas como mediadores culturais. O contexto se deve ao fato de que os dados inseridos nesses sistemas refletem recortes específicos da realidade social e, por meio de sequências programadas, os algoritmos processam essas informações, modificando-as e influenciando a forma como são interpretadas e utilizadas.

A abordagem possibilita demonstrar que tanto os algoritmos quanto os dados utilizados em sua configuração não são neutros. Os algoritmos, enquanto instrumentos tecnológicos, incorporam em seu design premissas e objetivos que refletem as intenções e escolhas dos agentes responsáveis por sua criação (Beer, 2009).

De modo semelhante, os processos de classificação e organização promovidos por esses sistemas estão imersos em julgamentos morais

implícitos, os quais geram impactos na distinção e na estratificação social (Fourcade; Healy, 2013). Nessa mesma perspectiva, os dados também possuem um caráter politizado e são produtos sociais, uma vez que derivam tanto da estrutura social na qual foram coletados, refletindo, por exemplo, desigualdades estruturais relacionadas a raça e gênero, quanto dos métodos e critérios adotados na coleta, os quais estão sujeitos a disputas em torno de premissas e finalidades (Fonseca, 2024).

Sob essa visão, considerar a inevitável não neutralidade dos algoritmos é um exercício essencial na contemporaneidade, visto que um número crescente de atividades humanas é mediado ou realizado por esses softwares. O reconhecimento da falácia da imparcialidade algorítmica constitui uma etapa fundamental para o uso, produção e análise responsáveis dessas tecnologias, permitindo o controle e a prevenção dos riscos sociais que emergem em uma sociedade cada vez mais moldada pela digitalização (Fonseca, 2024).

Abreu, Furtado e Santos (2022) argumentam que, embora os sistemas de inteligência artificial sejam avançados, eles não possuem a capacidade de julgamento efetivo. Assim, a IA não se torna, em si, preconceituosa, nem desenvolve consciência sobre preconceitos, uma vez que a discriminação é uma característica inerente ao comportamento humano. Conforme Coutinho (2021), os algoritmos sempre refletem os valores humanos de seus programadores, de modo que eventuais resultados indesejáveis estão relacionados aos vieses e discriminações presentes na própria sociedade.

A transferência da capacidade decisória dos seres humanos para as máquinas poderia sugerir que a neutralidade algorítmica eliminaria o viés humano oriundo de uma sociedade patriarcal. Contudo, os preconceitos de gênero enraizados na sociedade foram não apenas replicados, mas também ampliados pelas decisões algorítmicas (Costa, 2020).

Os vieses de gênero, presentes nos algoritmos, resultam em tratamentos diferenciados entre homens e mulheres, configurando discriminação algorítmica quando violam o princípio da isonomia. Tal prática é repudiada pelo ordenamento jurídico brasileiro, que adota os princípios da igualdade e da não discriminação como fundamentos do Estado de Direito, aplicáveis inclusive ao tratamento de dados pessoais, conforme disposto no art. 6º, IX, da Lei Geral de Proteção de Dados (LGPD) (Costa, 2020).

Ademais, o ordenamento jurídico brasileiro prevê a proteção contra práticas discriminatórias em desfavor das mulheres, abrangendo todas as formas de discriminação, inclusive aquelas perpetradas por meio das novas tecnologias. A proteção encontra respaldo na Constituição Federal, especialmente no art. 5º, caput e inciso I, e no art. 7º, incisos XX e XXX,

bem como em normas internacionais das quais o Brasil é signatário, como a Convenção da ONU para a Eliminação de Todas as Formas de Discriminação contra a Mulher (1969), as Convenções 100 e 111 da OIT e a Convenção Interamericana para Prevenir, Punir e Erradicar a Violência contra a Mulher (1994) (Costa, 2020).

Apesar do importante arcabouço normativo, há registros de casos concretos de discriminação de gênero por algoritmos, evidenciando tratamentos diferenciados em *softwares* de reconhecimento facial, análise de imagens e processamento de linguagem, que acabam por reforçar estereótipos de gênero associados às mulheres (Costa, 2020). Tal fenômeno não pode ser dissociado do contexto sociocultural brasileiro, que possui raízes profundamente enraizadas no patriarcado. A estrutura social vigente historicamente perpetua a discriminação e a violência contra as mulheres, refletindo-se em diversas esferas, inclusive na tecnologia. Nesse sentido, a inteligência artificial, ao invés de atuar como um vetor de mudança e promoção da igualdade, corre o risco de ampliar e automatizar preconceitos já existentes, caso não sejam adotadas medidas eficazes para mitigar esses vieses em sua programação e aplicação.

Diante desse cenário, o debate é particularmente relevante quando se analisam as interseções entre inteligência artificial, racismo e misoginia, pois evidencia como a tecnologia pode ser tanto um instrumento de progresso quanto um vetor de perpetuação de discriminações estruturais, caso não seja adequadamente regulada e aprimorada. Nesse contexto, a regulação e a governança algorítmica emergem como ferramentas fundamentais para promover a transparência, responsabilidade e equidade nos processos decisórios baseados em algoritmos. A adoção de políticas públicas e *frameworks* regulatórios adequados pode não apenas identificar e corrigir distorções discriminatórias, mas também estabelecer parâmetros éticos que garantam a igualdade de gênero no desenvolvimento e na aplicação dessas tecnologias.

Assim, o próximo tópico abordará como a regulação e a governança algorítmica podem ser estruturadas para enfrentar os desafios apresentados, destacando práticas e diretrizes que visam a construção de um ambiente digital mais justo e inclusivo para todos os gêneros.

3 O papel da Regulação e da Governança Algorítmica na promoção da Igualdade de Gênero

West, Whittaker e Crawford (2019) apontam que o setor de inteligência artificial enfrenta crise de diversidade relacionada a gênero e raça. De acordo com a pesquisa, apenas 18% dos autores em conferências de destaque na área são mulheres, enquanto mais de 80% dos professores de IA são homens. A desigualdade é ainda mais evidente na indústria, onde as mulheres representam 15% da equipe de pesquisa em IA do *Facebook* e apenas 10% no *Google*.

Dados sobre trabalhadores trans e outras minorias de gênero são inexistentes, e o cenário para pessoas negras é ainda mais alarmante, com apenas 2,5% da força de trabalho do *Google* composta por negros, número que chega a 4% no *Facebook* e na *Microsoft*. Apesar de décadas de esforços para mitigar o desequilíbrio, a situação permanece crítica.

West, Whittaker e Crawford (2019) defendem que a indústria de inteligência artificial deve reconhecer a gravidade da crise de diversidade, admitindo que as abordagens atuais não conseguiram enfrentar adequadamente a distribuição desigual de poder e os mecanismos pelos quais a IA pode reforçar tais desigualdades. Eles destacam ainda que o viés nos sistemas inteligentes reflete padrões históricos de discriminação, sendo ambos aspectos de um mesmo problema que deve ser tratado de forma integrada.

Os autores também criticam o foco restrito na inclusão de "mulheres na tecnologia", que frequentemente beneficia mulheres brancas e negligência outras interseccionalidades. Os estudos em IA, segundo os autores, tendem a assumir o gênero de forma binária, baseando-se em aparências físicas e estereótipos, o que apaga diversas formas de identidade de gênero (West, Whittaker e Crawford, 2019).

Além disso, West, Whittaker e Crawford (2019) argumentam que o simples ajuste no "pipeline" de candidatos diversos, do ambiente educacional para o mercado, não resolverá os problemas de diversidade. O foco nessa estratégia não aborda questões estruturais mais profundas, como a cultura do ambiente de trabalho, assimetrias de poder, práticas excludentes de contratação e remuneração, assédio e tokenismo, que contribuem para a evasão ou exclusão de grupos marginalizados no setor de IA.

Ainda, West, Whittaker e Crawford (2019) ressaltam a necessidade urgente de reavaliar o uso de sistemas de IA para classificação, detecção e previsão de raça e gênero. Lembram que a história da "ciência racial"

evidencia os perigos da classificação baseada em aparência física, prática que é cientificamente falha e passível de abusos. Ferramentas de IA que afirmam detectar sexualidade, prever criminalidade ou avaliar competências por meio de traços faciais e microexpressões são profundamente problemáticas, pois replicam padrões de preconceito racial e de gênero, perpetuando e justificando desigualdades históricas. O uso comercial dessas tecnologias, segundo os autores, é motivo de grande preocupação.

Nadeem, Abedin e Marjanovic (2020) destacam que os principais fatores que contribuem para o viés de gênero na inteligência artificial incluem a falta de diversidade tanto nos conjuntos de dados quanto entre os desenvolvedores, o viés estrutural presente na sociedade e o preconceito, consciente ou inconsciente, incorporado pelos programadores nos dados utilizados.

Para Nadeem, Abedin e Marjanovic (2020) é essencial que a sociedade continue os esforços para minimizar o viés de gênero sistêmico, além de adotar medidas específicas para reduzir distorções nos resultados da IA, o que pode ser feito por meio de práticas aprimoradas de design e implementação, orientadas por novas diretrizes éticas. Defendem ainda que enfrentar o viés de gênero na IA requer abordagens integradas, como a resolução de estereótipos enraizados na sociedade, a garantia de diversidade nas equipes de desenvolvimento, a redução de preconceitos na criação de algoritmos e a implementação ética e justa das aplicações de inteligência artificial.

Nishant, Schneckenberg Ravishankar (2023) asseveram a necessidade de uma abordagem mais refinada para o desenvolvimento da IA. Para garantir o uso responsável das tecnologias baseadas em IA e a evolução de sociedades equitativas impulsionadas pelo progresso tecnológico, defendem especificações mais precisas sobre os limites da agência de julgamento e da autoridade decisória da IA.

Lütz (2022) argumenta que abordar o problema da discriminação algorítmica baseada em gênero a partir da análise dos impactos diretos e indiretos possibilita que pesquisadores e formuladores de políticas compreendam melhor a complexidade do problema e identifiquem a necessidade de ações legais e políticas. A perspectiva facilita a compreensão dos mecanismos que sustentam o funcionamento dos algoritmos, evidenciando a importância dos efeitos indiretos de gênero, que, embora possam parecer irrelevantes à primeira vista, são fundamentais para entender e solucionar a discriminação algorítmica. Além disso, o papel dos algoritmos e motores de busca na formação e perpetuação de estereótipos e vieses de gênero deve ser considerado na formulação de políticas públicas.

Borgesius (2020) defende que a legislação antidiscriminatória e a lei de proteção de dados constituem os principais instrumentos legais no combate à discriminação ilegal promovida por sistemas algorítmicos, sendo eficazes na proteção das pessoas quando devidamente aplicadas. O autor propõe formas de aprimorar a aplicação das normas antidiscriminatórias, embora reconheça que certos tipos de decisões algorítmicas não são abrangidos pela legislação vigente, mesmo quando resultam em diferenciação ou discriminação injusta.

Tal cenário ocorre porque muitas leis antidiscriminatórias se restringem a critérios protegidos, como a origem étnica, deixando de fora diferenciações baseadas em categorias recém-criadas que não se correlacionam com esses critérios, mas que ainda podem ser injustas por perpetuarem desigualdades sociais. O autor aponta a necessidade de regulamentações adicionais para assegurar a equidade e os direitos humanos na tomada de decisões algorítmica, mas ressalta que a adoção de regras gerais não seria eficaz. Assim como a revolução industrial exigiu legislações específicas para diferentes áreas, como segurança no trabalho e proteção ambiental, a regulação da tomada de decisões algorítmica deve ser setorial, considerando os riscos e valores particulares de cada contexto (Borgesius, 2020).

Para mitigar esses impactos, é fundamental fomentar uma força de trabalho mais inclusiva na área de IA incentivando a presença feminina nas ciências, tecnologia, engenharia e matemática (STEM), com destaque para modelos femininos de sucesso e programas de mentoria. O fortalecimento de legislações contra o assédio *on-line* e a criminalização de *deepfakes* são medidas importantes para proteger a privacidade das mulheres (*Capitol Technology University*, 2024).

Além disso, políticas corporativas que considerem as responsabilidades de cuidado familiar, como horários flexíveis e opções de trabalho remoto, podem contribuir para um ambiente de trabalho mais inclusivo. A avaliação contínua da equidade salarial e das oportunidades de promoção para mulheres é essencial para reduzir as disparidades de gênero e minimizar os efeitos negativos da IA (*Capitol Technology University*, 2024).

Nessa senda, tem-se que a ascensão das tecnologias baseadas em inteligência artificial e algoritmos tem redesenhado significativamente os processos decisórios em diversas esferas da sociedade, desde o mercado de trabalho até o acesso a serviços públicos. No entanto, o potencial dessas tecnologias para perpetuar ou mesmo agravar desigualdades de gênero é uma preocupação crescente, especialmente quando se consideram os vieses presentes nos dados utilizados para treinar algoritmos.

Estudos como o de Buolamwini e Gebru (2018) evidenciam a disparidade de desempenho em sistemas de reconhecimento facial com base no gênero e na tonalidade da pele, reforçando a necessidade de abordar essas questões. Nessa esteira, é essencial ressaltar que se a equipe de desenvolvimento dos sistemas inteligentes for predominantemente branca e masculina, ela pode não abordar ou mesmo não entender plenamente as experiências únicas das mulheres negras, levando a sistemas de IA a serem tendenciosos (Ulnicane, 2024).

Por exemplo, sistemas de recrutamento baseados em IA podem perpetuar a discriminação ao utilizar dados históricos de contratação que refletem a desigualdade de gênero no mercado de trabalho. Além disso, assistentes virtuais com vozes e personas femininas podem reforçar estereótipos de subserviência e objetificação das mulheres. Nesse cenário, a regulação e a governança algorítmica emergem como instrumentos fundamentais para assegurar que as novas tecnologias contribuam para a promoção da igualdade de gênero, e não para a sua erosão. No contexto da IA, a necessidade de estabelecer marcos normativos que garantam a transparência, a responsabilização e a equidade nos processos decisórios é imperativa.

A Lei Geral de Proteção de Dados Pessoais no Brasil representa um passo importante nesse sentido ao prever o direito à revisão de decisões automatizadas (art. 20) e estabelecer o princípio da não discriminação no tratamento de dados pessoais (art. 6º, inciso IX) (Brasil, 2018). A governança algorítmica, por sua vez, envolve a criação de mecanismos de supervisão e controle sobre o desenvolvimento e a implementação de algoritmos, assegurando que os sistemas estejam alinhados com princípios éticos e direitos fundamentais. O conceito transcende a mera conformidade legal, abrangendo a adoção de práticas corporativas transparentes, auditorias independentes e a participação de grupos diversos no processo de desenvolvimento tecnológico.

A inclusão de mulheres e minorias de gênero nas equipes de desenvolvimento é um exemplo concreto de como a governança pode atuar na promoção da igualdade. West, Whittaker e Crawford (2019) abordam a sub-representação desses grupos no campo da IA e suas implicações, reforçando a importância da diversidade para mitigar vieses e discriminação. Ademais, a regulação internacional também tem papel relevante na promoção da igualdade de gênero diante da expansão tecnológica. Organismos como a União Europeia têm liderado iniciativas para a criação de marcos regulatórios que abordam a ética da IA exigindo a eliminação de vieses discriminatórios nos sistemas algorítmicos.

A "Proposta de Regulamento do Parlamento Europeu e do Conselho que estabelece regras harmonizadas em matéria de inteligência artificial" (Comissão Europeia, 2021) é um exemplo de legislação que busca equilibrar a inovação tecnológica com a proteção de direitos fundamentais, incluindo a igualdade de gênero. No entanto, a regulação da inteligência artificial também apresenta desafios significativos. Um deles é equilibrar a necessidade de transparência e explicabilidade dos sistemas algorítmicos com a proteção dos segredos industriais e da propriedade intelectual das empresas. Outro desafio é lidar com a natureza transnacional dos sistemas de IA, que podem ser desenvolvidos em um país, treinados com dados de outro e utilizados globalmente, o que exige um esforço de cooperação internacional para harmonizar as regulações e evitar que as empresas busquem jurisdições menos rigorosas.

Embora a regulação e a governança algorítmica sejam essenciais para mitigar os efeitos da discriminação de gênero, sua implementação prática enfrenta desafios significativos que precisam ser considerados. Em primeiro lugar, destaca-se a complexidade de se criar um marco regulatório eficaz em um contexto global. A natureza transnacional dos sistemas de inteligência artificial, que frequentemente são desenvolvidos em um país, treinados com dados de outro e aplicados mundialmente, dificulta a harmonização das legislações. Países possuem diferentes padrões de proteção de dados e de combate à discriminação, o que pode gerar lacunas jurídicas. Empresas podem, intencionalmente, buscar jurisdições com regulações mais flexíveis, o que compromete a eficácia de medidas protetivas em nível global.

Outro obstáculo relevante é o risco de que a regulação excessiva iniba a inovação tecnológica. A imposição de regras rígidas pode desencorajar empresas a investirem em novas tecnologias, por temerem sanções ou restrições que dificultem a competitividade no mercado global. O dilema entre inovação e proteção de direitos fundamentais representa um dos principais desafios no campo regulatório, exigindo soluções que equilibrem ambos os interesses.

Do ponto de vista técnico, a opacidade de muitos sistemas de IA, especialmente aqueles baseados em redes neurais complexas, dificulta a realização de auditorias eficazes. Os modelos funcionam como verdadeiras "caixas-pretas", dificultando a compreensão até mesmo por parte de seus próprios desenvolvedores. Tal característica técnica compromete não apenas a transparência, mas também a capacidade de identificar e corrigir vieses algorítmicos, tornando desafiador o cumprimento de princípios como a explicabilidade e a responsabilização. Além disso, a exigência de transparência entra em conflito com a proteção de segredos comerciais e direitos de

propriedade intelectual das empresas. A abertura dos códigos e dados para auditorias pode colocar em risco informações sensíveis e estratégicas, criando um impasse entre o interesse público na supervisão e o interesse privado na proteção de inovações.

A eficácia da regulação existente também é limitada por sua capacidade de acompanhar a rápida evolução tecnológica. Leis como a Lei Geral de Proteção de Dados, no Brasil, representam avanços importantes ao estabelecer o direito à revisão de decisões automatizadas e prever o princípio da não discriminação no tratamento de dados pessoais. No entanto, na prática, o exercício desses direitos pode ser dificultado por barreiras técnicas e burocráticas, além da falta de conhecimento da população sobre tais garantias. O processo legislativo, por sua vez, é tipicamente mais lento e reativo, o que cria um descompasso entre a evolução da IA e a capacidade de adaptação das normas regulatórias.

Nesse sentido, a regulação da inteligência artificial demanda abordagens mais dinâmicas e adaptativas. O uso de regulações setoriais pode ser uma solução eficaz, considerando que diferentes áreas, como saúde, segurança pública e mercado financeiro, apresentam riscos e características próprias que exigem respostas regulatórias específicas. Além disso, mecanismos como os *sandboxes* regulatórios permitem a experimentação controlada de novas tecnologias em ambientes supervisionados, possibilitando o teste de soluções inovadoras sem comprometer a segurança e a ética.

A governança algorítmica eficaz não se limita à conformidade legal, mas deve envolver práticas corporativas transparentes, auditorias independentes e a participação ativa da sociedade civil e da academia. A inclusão de grupos diversos no desenvolvimento de tecnologias, incluindo mulheres e minorias de gênero, não é apenas uma questão de justiça social, mas também uma estratégia para enriquecer o processo criativo e reduzir vieses. Assim, a combinação de regulamentação com práticas internas éticas e inclusivas é primordial para enfrentar os desafios apresentados pela IA na promoção da igualdade de gênero.

Katyal (2019) defende uma abordagem holística para enfrentar os desafios da discriminação algorítmica, reforçando a importância dessa interação. Além das medidas regulatórias e de governança, outras ações podem ser adotadas para combater a discriminação algorítmica de gênero. A indústria de tecnologia pode investir em programas de diversidade e inclusão, não apenas para aumentar a representatividade em suas equipes, mas também para incorporar perspectivas diversas no design e desenvolvimento dos sistemas de IA.

A academia pode aprofundar as pesquisas sobre vieses algorítmicos e desenvolver metodologias para auditar e mitigar esses vieses. Já a sociedade civil pode atuar na conscientização sobre o tema, pressionando por maior transparência e responsabilização das empresas de tecnologia. Nesse contexto, o papel do Direito é inegável, servindo como ferramenta de controle e incentivo para a construção de uma sociedade mais igualitária e justa, onde a inovação tecnológica caminhe lado a lado com a promoção dos direitos fundamentais. A regulação e a governança algorítmica, atuando de forma sinérgica, são pilares essenciais para assegurar que a era digital seja marcada pela equidade e pela inclusão.

A título de complementação, de modo a expandir a análise de práticas de mitigação de discriminação algorítmica, apresentando exemplos específicos e a sua eficácia, Schenkel *et al.* (2024) demonstraram problema crítico nos sistemas de aprendizado de máquina, como o *Google Translate*: a reprodução e o reforço de preconceitos sociais, especialmente os de gênero. Como resposta ao desafio, os pesquisadores desenvolveram novo modelo de tradução automática com foco na melhoria da precisão de gênero, que apresentou progresso relevante, elevando a pontuação de acerto de 68,75 para 70,09, o que demonstra maior capacidade de identificar e traduzir corretamente termos com marcação de gênero.

Além disso, observou-se redução de 15,7% na disparidade de desempenho entre traduções associadas a entidades masculinas e femininas. O resultado indica uma representação mais justa e equilibrada entre os gêneros, corrigindo distorções observadas em versões anteriores do sistema. O estudo também foi eficaz na redução de traduções estereotipadas, alcançando uma diminuição de 43% nesse tipo de viés, o que representa um importante passo na direção de uma linguagem mais equitativa, livre de reforços a preconceitos de gênero (Schenkel *et al.* 2024).

Um dos aspectos mais inovadores da pesquisa foi a introdução da possibilidade de tradução em três gêneros, feminino, masculino e neutro, em situações nas quais o idioma de origem (como o inglês) não apresenta marcação de gênero. Do ponto de vista técnico, o estudo utilizou o método *Constrained Beam Search*, que permite preservar a estrutura da frase enquanto assegura a representação correta do gênero. As traduções finais foram refinadas com a ferramenta *SimAlign*, o que contribuiu para a qualidade dos resultados. Como métrica de desempenho, o modelo alcançou uma pontuação BLEU de 48,39, demonstrando precisão e fluidez na produção dos textos traduzidos (Schenkel *et al.* 2024).

Em relação ao Google Tradutor, o modelo proposto demonstrou avanços expressivos: a acurácia de gênero nas traduções foi ampliada de 68,75 para 70,09; houve uma melhora de 15,7% na pontuação que avalia a diferença de precisão entre as unidades masculinas e femininas; além disso, as traduções estereotipadas foram reduzidas em 43%, indicando uma abordagem mais justa e equilibrada na representação de gênero (Schenkel *et al.* 2024).

Dong *et al.* (2024) apresentou nova estrutura metodológica para investigar o preconceito de gênero em modelos de linguagem de larga escala (*LLMs*). A estrutura foi testada em dez *LLMs* distintos, incluindo variações dos modelos LLAMA2, VICUNA, FALCON e OPT. Mesmo em entradas livres de qualquer marcação de gênero ou estereótipos explícitos, os pesquisadores identificaram evidências significativas de preconceito de gênero, tanto em sua forma explícita quanto implícita.

Para mensurar esse viés, o estudo propôs três métricas complementares: o *Gender Attribute Score* (GAS), que avalia o viés explícito, e o *Gender Logits Difference* (GLD) e o *Attribute Distribution Distance* (ADD), voltados à detecção de vieses implícitos. As métricas permitem uma quantificação precisa do preconceito de gênero presente no conteúdo gerado pelos modelos. Também foram investigadas estratégias de mitigação desse viés, com destaque para três abordagens: ajuste de hiperparâmetros, orientação de instruções e ajuste por *Debias Tuning*. Esta última se mostrou a mais eficaz na redução dos preconceitos de gênero, apresentando resultados consistentes em todas as métricas e conjuntos de dados utilizados (Dong *et al.* 2024).

Um achado relevante da pesquisa consistiu na constatação de que modelos maiores ou mais alinhados tendem a apresentar maior grau de viés, indicando que o aumento de capacidade ou sofisticação não garante, por si só, maior neutralidade, ao contrário, pode acentuar padrões discriminatórios. Embora os resultados sejam promissores, os autores reconheceram limitações, especialmente quanto à definição binária de gênero adotada nas métricas. Destacaram, portanto, a necessidade de pesquisas futuras que incorporem uma concepção mais fluida e inclusiva de gênero, a fim de aperfeiçoar a análise e a mitigação desses vieses (Dong *et al.* 2024).

Sob esse enfoque, denota-se que a promoção da igualdade de gênero na era digital requer esforço conjunto e multidisciplinar. A regulação e a governança algorítmica são pilares basilares nesse processo, mas devem ser complementadas por ações da indústria, da academia e da sociedade civil. Pesquisas futuras podem se aprofundar nos mecanismos de mitigação de vieses, nas metodologias de auditoria algorítmica e nos modelos de governança participativa. No campo das políticas públicas, é necessário avançar na

elaboração de marcos regulatórios adaptados aos desafios da IA buscando sempre o equilíbrio entre inovação e proteção dos direitos fundamentais.

Conclusão

A análise do referencial teórico e das discussões apresentadas neste estudo revela que a inteligência artificial pode perpetuar e ampliar desigualdades de gênero. O viés algorítmico decorre de dados desbalanceados, da sub-representação de mulheres nas equipes de desenvolvimento e da ausência de regulação adequada, impactando oportunidades de emprego, acesso a crédito e segurança, dentre outros setores.

Assim, a governança e a regulação algorítmica são elementares para mitigar esses efeitos, mas sua implementação enfrenta adversidades, como a necessidade de equilibrar transparência e inovação. Além disso, a natureza global da IA demanda cooperação internacional para harmonizar normativas e evitar lacunas regulatórias.

Nessa perspectiva, tem-se que a superação da discriminação algorítmica exige esforço pluridisciplinar. A inclusão de mulheres e outras minorias sociais no desenvolvimento de inteligência artificial é essencial para construir tecnologias mais justas e representativas. Dessa maneira, o futuro digital deve ser pautado por escolhas éticas e compromissos sociais, garantindo que a inovação contribua para a equidade de gênero e para a justiça social.

O estudo, no entanto, apresenta limites que merecem ser explicitados. Por adotar um recorte essencialmente teórico e documental, não se explorou de forma empírica como os sistemas algorítmicos operam na realidade brasileira, especialmente nas margens sociais onde a interseccionalidade de opressões é mais intensa. A ausência de entrevistas, estudos de caso localizados ou coleta de dados primários impede uma compreensão mais aprofundada dos impactos concretos vivenciados por mulheres negras, indígenas, trans ou com deficiência.

Diante disso, propõe-se, como agenda de pesquisas futuras, a ampliação da análise empírica sobre os impactos da IA em territórios periféricos e contextos racializados, com destaque para o papel da tecnologia na manutenção de hierarquias sociais. É importante, ainda, ocorrer o aprofundamento da discussão sobre o controle social dos algoritmos, a responsabilização corporativa, a representatividade de gênero e raça nas equipes técnicas, bem como o desenvolvimento de ferramentas jurídicas que articulem regulação algorítmica, justiça de gênero e inclusão digital.

Conclui-se, portanto, que o enfrentamento à discriminação algorítmica não se resume à correção de falhas técnicas, mas demanda projeto coletivo de justiça social. A regulação e a governança algorítmica, se conduzidas com escuta ativa, diversidade epistêmica e responsabilidade pública, podem contribuir para que a inteligência artificial se torne uma aliada na promoção da igualdade de gênero e não um instrumento de reprodução de silenciamentos e desigualdades.

Referências

ABREU, A. J. A.; FURTADO, K. C. S.; SANTOS, R. K. C. Inteligência artificial e preconceito de identidade de gênero: o problema do viés na construção das IA's e a perpetuação das discriminações em sociedades previamente discriminatórias. **COR LGBTQIA+**, v. 1, n. 3, p. 229–247, 2022. Disponível em: https://revistas.ceeinter.com.br/CORLGBTI/article/view/551. Acesso em: 3 fev. 2025.

ADAMS-PRASSL, Jeremias, *et al*. Directly Discriminatory Algorithms. **The Modern Law Review**, vol. 86, no 1, 2023, p. 144–175, Disponível em: doi:10.1111/1468-2230.12759. Acesso em: 20 jun. 2025.

ANDRÉS, Pablo de; GIMENO, Ricardo; CABO, Ruth Mateos de. The gender gap in bank credit access. **Journal of Corporate Finance**, v. 71, p. 101782, 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S0929119920302261. Acesso em: 3 fev. 2025.

ANINZE, A. Artificial Intelligence Life Cycle: The Detection and Mitigation of Bias. **Proceedings of the International Conference on AI Research.**, [*s. l.*], v. 4, n. 1, p. 40–49, 2024. Disponível em: https://papers.academic-conferences.org/index.php/icair/article/download/3131/2881. Acesso em: 21 jun. 2025.

ARANTES, Flávio Antônio Nigro. **Impactos da Lei Geral de Proteção de Dados nas relações de trabalho**: tomada de decisão por inteligência artificial nos processos de recrutamento e seleção. Reflexões sobre o direito do trabalhador de rever o algoritmo de decisão. 2022. Dissertação (Mestrado em Direito) – Universidade Federal de Minas Gerais, Faculdade de Direito, Belo Horizonte. Disponível em: https://repositorio.ufmg.br/handle/1843/47689. Acesso em: 20 jun. 2025.

BEER, David. Power through the algorithm? Participatory web cultures and the technological unconscious. **New Media & Society**, v. 11, n. 6, p. 985–1002, 2009. Disponível em: 10.1177/1461444809336551. Acesso em: 20 jun. 2025.

BORGES, G. S.; FILÓ, M. da C. S. Inteligência artificial, gênero e direitos humanos: o caso Amazon. **Revista Justiça do Direito**, v. 35, n. 3, p. 218–243,

2021. Disponível em: https://doi.org/10.5335/rjd.v35i3.12259. Acesso em: 3 fev. 2025.

BORGESIUS, Frederik J. Zuiderveen. Strengthening legal protection against discrimination by algorithms and artificial intelligence. **The International Journal of Human Rights**, v. 24, n. 10, p. 1572–1593, 2020. Disponível em: https://doi.org/10.1080/13642987.2020.1743976. Acesso em: 9 fev. 2025.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Lei Geral de Proteção de Dados Pessoais – LGPD). Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 3 fev. 2025.

BUOLAMWINI, J.; GEBRU, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: **Conference On Fairness, Accountability and Transparency**, 2018, New York. Anais.... p. 77–91. Disponível em: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. Acesso em: 10 fev. 2025.

CAPITOL TECHNOLOGY UNIVERSITY. **Artificial intelligence and its unique threat to women.** 2024. Disponível em: https://www.captechu.edu/blog/artificial-intelligence-and-its-unique-threat-women. Acesso em: 9 fev. 2025.

COMISSÃO EUROPEIA. **Proposta de regulamento do Parlamento Europeu e do Conselho que estabelece regras harmonizadas em matéria de inteligência artificial.** 2021. Disponível em: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_PT.html. Acesso em: 10 fev. 2025.

COSTA, Diego Carneiro. **O viés do algoritmo e a discriminação por motivos relacionados à sexualidade. 2020.** Dissertação (Mestrado em Direito) – Universidade Federal da Bahia, Salvador, 2020. Disponível em: https://repositorio.ufba.br/handle/ri/34394. Acesso em: 9 fev. 2025.

COUTINHO, Marina de Alencar Araripe. Considerações sobre inteligência artificial e tomada de decisão. In: PEIXOTO, Fabiano Hartmann (org.). **Inteligência artificial:** estudos de inteligência artificial. Curitiba: Alteridade, 2021.

DAVIS, J. L.; WILLIAMS, A.; YANG, M. W. Algorithmic reparation. **Big Data & Society**, v. 2, p. 1–12, 2021. Disponível em: https://doi.org/10.1177/20539517211044808. Acesso em: 3 fev. 2025.

DONG, X.; WANG, Y.; YU, P. S.; CAVERLEE, J. **Disclosure and Mitigation of Gender Bias in LLMs.** [*s. l.*], 2024. Disponível em: https://arxiv.org/pdf/2402.11190. Acesso em: 22 jun. 2025.

FLASINSKI, Mariusz. **Introduction to artificial intelligence**. Cham: Springer, 2016. p. 31.

FONSECA, Bruno. A ilusão da neutralidade algorítmica. **Sociedade & Cultura**, 2024. Disponível em: https://understandingai.iea.usp.br/nota-critica/a-ilusao-da-neutralidade-algoritmica/. Acesso em: 3 fev. 2025.

FOURCADE, Marion; HEALY, Kieran. Classification situations: life-chances in the neoliberal era. **Accounting, Organizations and Society**, v. 38, n. 8, 2013. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0361368213000743. Acesso em: 3 fev. 2025.

FRAZÃO, Ana. Discriminação algorítmica: compreendendo o que são os julgamentos algorítmicos e o seu alcance na atualidade. **JOTA**, 2024. Disponível em: https://www.jota.info/opiniao-e-analise/colunas/constituicao-empresa-e-mercado/discriminacao-algoritmica. Acesso em: 3 fev. 2025.

HAMPTON, L. M. **Black Feminist Musings on Algorithmic Oppression.** [*s. l.*], p. 1, 2021. Disponível em: https://dblp.uni-trier.de/db/journals/corr/corr2101.html#abs-2101-09869. Acesso em: 21 jun. 2025.

JOYCE, Kelly; SMITH-DOERR, Laurel; ALEGRIA, Sharla; BELL, Susan; CRUZ, Taylor; HOFFMAN, Steve G.; NOBLE, Saya; SHESTAKOFSKY, Benjamin. Toward a sociology of artificial intelligence: a call for research on inequalities and structural change. **Socius**, v. 7, 2021. Disponível em: https://doi.org/10.1177/2378023121999581. Acesso em: 3 fev. 2025.

KATYAL, S. K. Private accountability in the age of artificial intelligence. **UCLA Law Review**, v. 66, n. 1, p. 54–141, 2019. Disponível em: https://www.uclalawreview.org/private-accountability-age-algorithm/. Acesso em: 10 fev. 2025.

LEE, Kai-Fu. **Inteligência artificial:** como os robôs estão mudando o mundo. São Paulo: Globo Livros, 2020. p. 27.

LÜTZ, F. Gender equality and artificial intelligence in Europe: addressing direct and indirect impacts of algorithms on gender-based discrimination. **ERA Forum**, v. 23, p. 33–52, 2022. Disponível em: https://doi.org/10.1007/s12027-022-00709-6. Acesso em: 9 fev. 2025.

MCKINSEY & COMPANY. **What is AI (artificial intelligence)?** McKinsey & Company, 3 abr. 2024. Disponível em: https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-ai. Acesso em: 20 jun. 2025.

NADEEM, Ayesha; ABEDIN, Babak; MARJANOVIC, Olivera. Gender bias in AI: a review of contributing factors and mitigating strategies. In: **Australasian Conference On Information Systems – ACIS 2020**, 2020, Wellington. Anais

[...]. Art. 27. Disponível em: https://aisel.aisnet.org/acis2020/27. Acesso em: 09 fev. 2025.

NISHANT, R.; SCHNECKENBERG, D.; RAVISHANKAR, M. The formal rationality of artificial intelligence-based algorithms and the problem of bias. **Journal of Information Technology**, v. 39, n. 1, p. 19–40, 2024. Disponível em: https://doi.org/10.1177/02683962231176842. Acesso em: 9 fev. 2025.

PENCHIKALA, Srini. Analisando e prevenindo o preconceito inconsciente em machine learning. [Tradução de Leonardo Muniz]. **InfoQ** *Brasil*, 22 nov. 2018. Disponível em: https://www.infoq.com/br/articles/machine-learning-unconscious-bias/. Acesso em: 7 fev. 2025.

PLENTZ, Rafael Dobrachinsky. **Redes bayesianas para análise de comportamento aplicadas a telefonia celular.** 2023. 119 f. Dissertação (Mestrado em Mestre em Ciência da Computação) – Universidade Federal de Santa Catarina, Florianópolis, 2023. Disponível em: https://repositorio.ufsc.br/bitstream/handle/123456789/85325/233576.pdf?sequence=1&isAllowed=y. Acesso em: 23 jun. 2025.

PINTO, Henrique Alves. A utilização da inteligência artificial no processo de tomada de decisões: por uma necessária *accountability*. **Revista de Informação Legislativa**, Brasília, v. 57, n. 225, p. 43–60, jan./mar. 2020. Disponível em: https://www12.senado.leg.br/ril/edicoes/57/225/ril_v57_n225_p43. Acesso em: 20 jun. 2025.

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial Intelligence**: A Modern Approach. 3. ed. Upper Saddle River: Prentice Hall, 2009. Disponível em: http://repo.darmajaya.ac.id/5272/1/Artificial%20Intelligence-A%20Modern%20Approach%20(3rd%20Edition)%20(%20PDFDrive%20).pdf. Acesso em: 20 jun. 2025.

SCHENKEL, V.; MELLO, B.; RIGO, S. J.; DE OLIVEIRA RAMOS, G. UnbIAs – Google Translate gender bias mitigation and addition of genderless translation. **Revista de Informática Teórica e Aplicada**, Porto Alegre, v. 31, n. 2, p. 74–90, 2024. Disponível em: https://seer.ufrgs.br/index.php/rita/article/view/139902/92887. Acesso em: 22 jun. 2025.

SCHERTEL MENDES, L.; MATTIUZZO, M. Discriminação algorítmica: conceito, fundamento legal e tipologia. **Direito Público**, v. 16, n. 90, 2019. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/3766. Acesso em: 7 fev. 2025.

SIEGWART, Roland; NOURBAKHSH, Ilha R. **Introduction to autonomous mobile robots**. Cambridge: MIT Press, 2004. p. 7–12.

UNESCO; IRCAI. **Challenging systematic prejudices:** an investigation into gender bias in large language models. 2024. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000388971. Acesso em: 9 fev. 2025.

ULNICANE, I. Intersectionality in Artificial Intelligence: Framing Concerns and Recommendations for Action. **Social Inclusion**, [*s. l.*], 2024. Disponível em: https://www.cogitatiopress.com/socialinclusion/article/view/7543. Acesso em: 21 jun. 2025.

WEST, S. M.; WHITTAKER, M.; CRAWFORD, K. **Discriminating systems:** gender, race and power in AI. New York: AI Now Institute, 2019. Disponível em: https://ainowinstitute.org/discriminatingsystems.html. Acesso em: 9 fev. 2025.

WILLIAMS, T. R. The ethical implications of using generative chatbots in higher education. **Frontiers in Education**, v. 8, 2024. Disponível em: https://www.frontiersin.org/articles/10.3389/feduc.2023.1331607/full. Acesso em: 3 fev. 2025.

WORLD ECONOMIC FORUM. **IA para impacto:** o papel da inteligência artificial na inovação social. Genebra: WEF, 2024. Disponível em: https://reports.weforum.org/docs/WEF_AI_for_Impact_ptbr_2024.pdf. Acesso em: 10 fev. 2025.

Recebido em abril de 2025.

Aprovado em junho de 2025.