Viés e Discriminação no Uso de Aprendizado de Máquina em Governo Digital: Uma Revisão da Literatura

Gabriela B. Coutinho[1], **Igor G. B. Sampaio**[1],
Raissa Barcellos[2], **Flavia Bernardini**[1]

[1] Instituto de Computação – Universidade Federal Fluminense (UFF)

[2]Instituto de Matemática e Estatística – Universidade do Estado do Rio de Janeiro (UERJ)

{cbarrosgabriela, igorgbs}@gmail.com, fcbernardini@ic.uff.br

Abstract. *Artificial Intelligence (AI) has been increasingly adopted in the public sector to enhance the efficiency and accuracy of policies and public services. However, these technologies may reproduce historical inequalities embedded in data, raising concerns about bias, fairness, and discrimination in critical domains such as criminal justice, healthcare, social benefits, and public procurement. Although there is extensive literature on AI in the public sector and numerous studies proposing general bias mitigation strategies, there is a lack of systematic investigations focusing specifically on how bias is addressed in the context of digital government. This gap limits the understanding of how governments, oversight bodies, and society can ensure fair, transparent, and accountable AI-driven public services. This paper presents a Systematic Literature Review (SLR) on how academic research addresses bias and discrimination in AI systems applied to the public sector. The review follows the Kitchenham protocol and the PRISMA model, covering studies retrieved from Scopus and IEEE Xplore, resulting in 372 initial publications and a final sample of 31 articles. The analysis identifies mitigation strategies such as data balancing, synthetic data generation, and interpretable models. However, most approaches remain experimental, with limited replicability and insufficient consideration of sensitive attributes and real-world deployment challenges. The findings highlight a gap between technical solutions and their practical application in government contexts. The study emphasizes the need for auditable, reproducible, and socially grounded approaches, positioning algorithmic fairness as a fundamental requirement for trustworthy digital government and citizen-centered public service delivery.*

Resumo. *A adoção crescente de Inteligência Artificial (IA) no setor público tem ampliado a eficiência e a precisão de políticas e serviços públicos. Entretanto, essas tecnologias podem reproduzir desigualdades históricas presentes nos dados, levantando preocupações relacionadas a viés, justiça e discriminação em áreas críticas como justiça criminal, saúde, benefícios sociais e compras públicas. Apesar da ampla literatura sobre o uso de IA no setor público e das diversas propostas de mitigação de vieses, ainda há escassez de estudos sistemáticos que analisem especificamente como esses vieses são tratados no contexto do*

[0]Corresponding author: raissa.barcellos@ime.uerj.br

governo digital. Essa lacuna limita a compreensão de como governos, órgãos de controle e a sociedade podem garantir sistemas mais justos, transparentes e alinhados à prestação de serviços públicos. Este trabalho apresenta uma Revisão Sistemática da Literatura (RSL) sobre como a pesquisa acadêmica aborda o viés e a discriminação em sistemas de IA aplicados ao setor público. A revisão segue o protocolo de Kitchenham e o modelo PRISMA, considerando estudos das bases Scopus e IEEE Xplore, com 372 publicações iniciais e 31 trabalhos na análise final. Os resultados identificam estratégias como balanceamento de dados, uso de dados sintéticos e modelos interpretáveis. Contudo, a maioria das abordagens permanece em nível experimental, com limitações de replicabilidade e pouca consideração de variáveis sensíveis e de contextos reais de aplicação. Os achados evidenciam uma lacuna entre soluções técnicas e sua aplicação prática no contexto governamental. O estudo destaca a necessidade de abordagens auditáveis, reprodutíveis e socialmente orientadas, posicionando a justiça algorítmica como um requisito fundamental para um governo digital confiável e centrado no cidadão.

1. Introdução

A adoção crescente de algoritmos de aprendizado de máquina no setor público tem transformado processos decisórios, promovendo maior eficiência e precisão nas políticas públicas [Levy et al. 2021]. Inserida na agenda de governo digital, essa transformação busca otimizar a alocação de recursos e aprimorar a prestação de serviços ao cidadão. No entanto, o uso dessas tecnologias levanta preocupações éticas e sociais, especialmente pela possibilidade de reproduzir e amplificar desigualdades históricas presentes nos dados [Zajko 2022]. Estudos demonstram que sistemas algorítmicos podem incorporar vieses dos dados de treinamento, afetando de forma desproporcional grupos historicamente marginalizados [Turner Lee 2018, Mergen et al. 2025]. A busca por equidade e transparência tem impulsionado o desenvolvimento de estratégias de mitigação, como anonimização de dados, balanceamento de amostras e uso de modelos interpretáveis [Aldoseri et al. 2023]. Paralelamente, diretrizes internacionais, como as da União Europeia, reforçam princípios como não discriminação, transparência e *accountability* no uso da IA no setor público [Grupo de Peritos de Alto Nível sobre Inteligência Artificial 2019]. Apesar desses avanços, evidências mostram que vieses algorítmicos continuam impactando decisões em áreas críticas, como justiça criminal, benefícios sociais e contratações públicas, com efeitos diretos sobre cidadãos e sobre a confiança nas instituições.

No contexto do governo digital, tais desafios não podem ser tratados apenas como falhas técnicas, pois envolvem processos decisórios, atores institucionais e implicações sociais amplas. Embora o tema tenha sido explorado sob diferentes perspectivas, ainda há escassez de estudos sistemáticos que investiguem como mitigar vieses em IA considerando as exigências de transparência, *accountability* e prestação de serviços públicos ao longo do ciclo de vida desses sistemas. Assim, esta revisão se posiciona na interseção entre ciência de dados e administração pública, buscando compreender como a literatura articula estratégias de mitigação de vieses com as demandas do governo digital. Diferentemente de revisões amplas sobre IA e fairness, esta RSL concentra-se especificamente no contexto do governo digital. O estudo analisa como a literatura aborda vieses, discriminação e variáveis sensíveis em sistemas de IA utilizados em decisões públicas. Além de

mapear estratégias técnicas de mitigação, busca identificar lacunas relacionadas à governança algorítmica, transparência e aplicabilidade prática dessas soluções no setor público.

Diante desse cenário, este trabalho apresenta uma Revisão Sistemática da Literatura com o objetivo de analisar como a pesquisa acadêmica aborda o viés, a discriminação e o uso de variáveis sensíveis em sistemas de IA no setor público, bem como mapear as estratégias propostas para sua mitigação. A investigação segue o protocolo de RSL [Moher et al. 2009], permitindo identificar padrões, limitações e oportunidades para o desenvolvimento de soluções mais equitativas, transparentes e responsáveis. O restante do trabalho está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve a metodologia; a Seção 4 apresenta os resultados; a Seção 5 discute os achados; e a Seção 6 traz as conclusões e direções futuras.

2. Trabalhos Relacionados

Diversos estudos têm analisado criticamente o uso de IA e aprendizado de máquina no setor público, especialmente no que se refere a questões de viés, justiça e discriminação. Revisões como a de [Balayn et al. 2021] identificam métricas de *fairness*, estratégias de mitigação e lacunas importantes, como a ausência de padrões para auditoria de viés. Casos emblemáticos, como o sistema COMPAS, evidenciam os riscos de discriminação algorítmica. Na área da saúde, [Perez-Downes et al. 2024] apontam que dados desbalanceados comprometem a equidade, enquanto, no contexto das políticas públicas, [Valle-Cruz et al. 2019] destacam o uso ainda incipiente da IA, frequentemente limitado a protótipos e aplicações futuras.

Outros estudos abordam aspectos estruturais e organizacionais do uso da IA no setor público. [Hoekstra et al. 2021] propõem uma tipologia para classificar aplicações de IA, enquanto [Johnson and Reyes 2021] discutem implicações éticas e sociais, incluindo privacidade, transparência e governança. De modo geral, a literatura aponta desafios como exclusão digital, impacto no emprego e reprodução de vieses históricos. Apesar dessas contribuições, os estudos concentram-se, em sua maioria, em análises conceituais amplas ou classificações funcionais das aplicações de IA. Poucos abordam de forma sistemática a interseção entre viés algorítmico, variáveis sensíveis e contextos decisórios governamentais. Nesse sentido, este trabalho foca especificamente na mitigação de vieses em sistemas de IA no setor público, oferecendo um mapeamento estruturado das estratégias propostas na literatura e contribuindo para a promoção de equidade, transparência e responsabilidade no governo digital.

3. Metodologia

Para alcançar os objetivos propostos, realizou-se uma Revisão Sistemática da Literatura (RSL) adaptada, baseada nos procedimentos de [Kitchenham 2004] e no diagrama PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*), conforme [Moher et al. 2009]. O protocolo incluiu a definição da questão de pesquisa, da estratégia de busca e dos critérios de inclusão e exclusão. As buscas foram conduzidas nas bases Scopus e IEEE Xplore utilizando a *string ("machine learning"OR "artificial intelligence") AND ("bias"OR "discrimination") AND "government"*, aplicada aos campos de título, resumo e palavras-chave na Scopus e ao campo de resumo na IEEE. A coleta, realizada em 20 de março de 2025, resultou em 372 publicações, sendo 361 provenientes

da Scopus e 11 da IEEE. Como não foi estabelecido recorte temporal, foram considerados todos os estudos disponíveis até a data da busca. A investigação foi guiada pela seguinte questão de pesquisa: *Como os trabalhos da literatura tratam do viés e da discriminação em tomadas de decisão no contexto governamental?*

Foram considerados elegíveis estudos que abordassem mitigação de vieses, explicabilidade, transparência ou justiça algorítmica em sistemas de IA aplicados ao setor público. Em contrapartida, foram excluídos trabalhos sem relação com decisões governamentais, bem como estudos secundários, utilizados apenas na contextualização teórica e na seção de trabalhos relacionados. O processo de seleção ocorreu em três fases sucessivas. Inicialmente, realizou-se a leitura dos resumos das 372 publicações recuperadas, resultando em 77 estudos potencialmente relevantes. Em seguida, foram analisadas as introduções e conclusões desses trabalhos para verificar aderência aos objetivos da revisão, sendo excluídos 20 estudos indisponíveis em acesso aberto. Ao final, definiu-se um corpus composto por 31 trabalhos submetidos à análise detalhada, com o objetivo de identificar práticas recorrentes, lacunas metodológicas e estratégias de mitigação relacionadas ao viés e à discriminação em sistemas automatizados no setor público.

Os estudos selecionados foram organizados em três categorias analíticas: (i) estudos secundários utilizados para contextualização teórica; (ii) estudos teóricos, voltados a discussões conceituais, normativas e sociais sobre IA no setor público; e (iii) estudos técnico-aplicados, que propõem ou avaliam estratégias computacionais para mitigação de vieses em sistemas automatizados governamentais. Embora estudos secundários não tenham integrado o corpus final da RSL, alguns foram mantidos na Tabela 1 para fins de contextualização do campo. Essa classificação permitiu articular fundamentos teóricos e implicações práticas relacionadas ao uso da IA no contexto governamental. A Figura 1 apresenta o fluxo completo da revisão sistemática segundo o protocolo PRISMA, contemplando as etapas de identificação, triagem, elegibilidade e inclusão dos estudos. O processo resultou na seleção final de 31 trabalhos a partir das 372 publicações inicialmente recuperadas nas bases Scopus e IEEE Xplore.

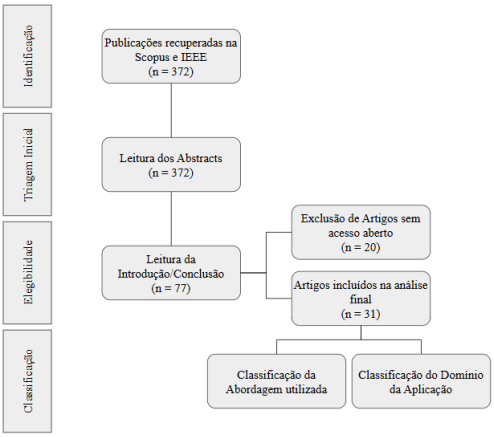

Figura 1. Fluxo de informações através das etapas da revisão usando o diagrama de fluxo PRISMA (Adaptado de [Moher et al. 2009])

4. Resultados

A partir da análise dos estudos selecionados, foi possível identificar padrões recorrentes, abordagens metodológicas distintas e estratégias específicas voltadas à mitigação de vieses em sistemas de IA aplicados ao setor público. Para facilitar a compreensão das contribuições de cada trabalho, elaborou-se a Tabela 1, que sintetiza os principais aspectos de cada estudo, incluindo o domínio de aplicação, a abordagem adotada e a descrição das soluções propostas. Observa-se que a literatura recente sobre mitigação de vieses no setor público revela um crescente esforço interdisciplinar para enfrentar desigualdades estruturais em áreas críticas como saúde, governança, justiça criminal e políticas públicas. Os resultados também evidenciam como desafios relacionados ao viés algorítmico impactam diretamente iniciativas de governo digital, especialmente em processos de tomada de decisão automatizada e prestação de serviços públicos. Para revisar a clareza e a fluidez do presente trabalho, ferramentas de IA Generativa [1] foram empregadas no processo de revisão de texto, auxiliando na otimização e na estruturação dos parágrafos. A seguir, os resultados são apresentados por domínio de aplicação, destacando como cada contexto contribui para a compreensão dos desafios e das estratégias de mitigação de vieses associados ao uso da IA no setor público.

Tabela 1. Visão Geral dos Estudos sobre Viés Algorítmico em IA Governamental

Autores	Tipo de Estudo	Domínio	Contribuição
[Utomo et al. 2022]	Estudo Técnico-Aplicado	Cidades Inteligentes	IA federada com foco em equidade
[Balayn et al. 2021]	Estudo Secundário	Ética, Transparência e Justiça Alg.	Desafios e lacunas na mitigação de vieses
[Bannister and Connolly 2020]	Estudo Teórico	Ética, Transparência e Justiça Alg.	*Framework* de riscos algorítmicos
[Fountain 2022]	Estudo Teórico	Ética, Transparência e Justiça Alg.	Crítica ao viés algorítmico
[Gentzel 2021]	Estudo Teórico	Ética, Transparência e Justiça Alg.	Discriminação em reconhecimento facial
[Johnson and Reyes 2021]	Estudo Secundário	Ética, Transparência e Justiça Alg.	Impactos sociais e éticos da IA
[Landers and Behrend 2023]	Estudo Técnico-Aplicado	Ética, Transparência e Justiça Alg.	Auditoria de IA com foco em *fairness*
[Patrikar et al. 2023]	Estudo Técnico-Aplicado	Ética, Transparência e Justiça Alg.	Mitigação com dados sintéticos (GANs)
[Torres-Berru et al. 2023]	Estudo Técnico-Aplicado	Ética, Transparência e Justiça Alg.	Detecção de viés em licitações
[Valle-Cruz et al. 2019]	Estudo Secundário	Ética, Transparência e Justiça Alg.	Panorama da IA no setor público
[Wei and Zhou 2022]	Estudo Teórico	Ética, Transparência e Justiça Alg.	Problemas éticos em sistemas de IA
[Liu et al. 2019]	Estudo Teórico	Justiça Criminal	Viés em decisões judiciais
[O'Donnell 2019]	Estudo Técnico-Aplicado	Justiça Criminal	Discriminação em policiamento preditivo
[PENG et al. 2024]	Estudo Teórico	Justiça Criminal	Discriminação em governança algorítmica
[Vetrò et al. 2019]	Estudo Técnico-Aplicado	Justiça Criminal	Padrões de viés em dados públicos
[Guevara-Gómez et al. 2021]	Estudo Teórico	Políticas Públicas	Viés de gênero em políticas de IA
[Hoekstra et al. 2021]	Estudo Secundário	Políticas Públicas	Tipologia de aplicações de IA
[Malhotra and Misra 2022]	Estudo Teórico	Políticas Públicas	Regulação e *accountability* em IA
[Mitrou et al. 2021]	Estudo Teórico	Políticas Públicas	Discricionariedade humana em IA

[1] GPT e Gemini

[Nwafor 2024]	Estudo Teórico	Políticas Públicas	Igualdade de gênero em políticas de IA
[Pérez-Suay et al. 2017]	Estudo Técnico-Aplicado	Políticas Públicas	*Fair regression* e redução de viés
[Nascimento et al. 2022]	Estudo Técnico-Aplicado	Redes Sociais	mitigação de vieses em discurso de ódio
[Cheng et al. 2023]	Estudo Técnico-Aplicado	Saúde	Inferência de raça em dados clínicos
[Gao et al. 2021]	Estudo Técnico-Aplicado	Saúde	Modelos preditivos com *Random Forest*
[Lucas et al. 2024]	Estudo Técnico-Aplicado	Saúde	Predição justa em saúde
[McCradden et al. 2020]	Estudo Teórico	Saúde	Viés em dados clínicos
[Newman-Griffis et al. 2022]	Estudo Teórico	Saúde	Viés relacionado à deficiência
[Oniani et al. 2023]	Estudo Teórico	Saúde	Diretrizes éticas para IA generativa
[Perez-Downes et al. 2024]	Estudo Secundário	Saúde	Estratégias de mitigação em *Machine Learning*
[Rodriguez et al. 2024]	Estudo Teórico	Saúde	Riscos de LLMs na saúde
[Zhao 2020]	Estudo Técnico-Aplicado	Seguridade Social	Predição de desemprego com *fairness*

Cidades Inteligentes: Arquiteturas tecnológicas têm sido desenvolvidas com foco em ética e responsabilidade. Uma arquitetura federada confiável de IA (FTAI), aplicada em cidades inteligentes, integrou aprendizado federado com ferramentas como AIF360, AIX360 e ART, priorizando dados balanceados, redução de custos de rede e justiça algorítmica [Utomo et al. 2022].

Ética, Transparência e Justiça Algorítmica: A adoção crescente de IA no setor público levanta preocupações éticas, especialmente quanto à transparência, responsabilização e justiça algorítmica. No Equador, um modelo que combina *machine learning* e processamento de linguagem natural identificou viés de gênero e favoritismo em 32% das licitações públicas [Torres-Berru et al. 2023]. Mesmo sem variáveis sensíveis explícitas, algoritmos podem inferir essas características e reforçar desigualdades [Gentzel 2021]. A análise de 150 incidentes no *AI Incident Database* resultou em uma taxonomia de falhas éticas, categorizando problemas como discriminação, baixo desempenho técnico e riscos à segurança física [Wei and Zhou 2022]. Estratégias de mitigação incluem marcos regulatórios com responsabilização, auditoria pública e regulação independente [Bannister and Connolly 2020], além de soluções técnicas, como uso de GANs para geração de dados sintéticos, que melhoram paridade e igualdade de oportunidade [Patrikar et al. 2023]. Também se destacam a curadoria de dados, auditorias recorrentes e reformas institucionais [Fountain 2022]. Recentemente, propõe-se um *framework* de auditoria que integra psicometria, ciência da computação e ciências sociais, reforçando a importância da colaboração interdisciplinar [Landers and Behrend 2023]. Embora os trabalhos reconheçam a importância da transparência e da justiça algorítmica, há diferenças significativas quanto às estratégias propostas, variando entre abordagens regulatórias, auditorias independentes e soluções estritamente computacionais.

Justiça Criminal: O uso de IA na justiça criminal gera preocupações sobre reprodução de vieses e violação de princípios como transparência, equidade e devido processo legal. Casos como o COMPAS e o *chatbot* Tay mostram como variáveis como localização funcionam como *proxies* discriminatórios [PENG et al. 2024]. Sistemas como o COMPAS operam de forma opaca, perpetuando desigualdades e destacando a neces-

sidade de responsabilização [Liu et al. 2019]. Mesmo sem variáveis sensíveis explícitas, o uso de *proxies* como histórico socioeconômico mantém padrões discriminatórios [O'Donnell 2019]. Estudos empíricos com bases como *Credit Card Default*, COMPAS e *Student Alcohol Consumption* reforçam a necessidade de modelos mais justos e transparentes [Vetrò et al. 2019]. De forma geral, os estudos sobre justiça criminal convergem ao apontar que variáveis *proxy* e opacidade algorítmica representam riscos significativos para a equidade em decisões automatizadas. Entretanto, enquanto alguns trabalhos enfatizam soluções técnicas de mitigação, outros defendem maior supervisão humana e mecanismos institucionais de responsabilização.

Políticas Públicas: A ética e a justiça algorítmica são centrais nas discussões sobre IA em políticas públicas. Modelos que aplicam o *Hilbert-Schmidt Independence Criterion* (HSIC) promovem justiça e mantêm a acurácia [Pérez-Suay et al. 2017]. Críticas ao modelo de IA como agente racional maximizador de utilidade defendem a manutenção da discricionariedade humana para garantir imparcialidade e responsabilidade [Mitrou et al. 2021]. A perspectiva de gênero segue pouco considerada em políticas de IA, como mostram estudos em países como Egito, Ruanda e Maurício, que reforçam a necessidade de dados desagregados e abordagens interseccionais [Nwafor 2024]. Na Índia, a falta de responsabilização ilustra o "problema das muitas mãos", exigindo marcos regulatórios que priorizem transparência e diversidade [Malhotra and Misra 2022]. Documentos da União Europeia e da Espanha também indicam a urgência de regulamentações com perspectivas feministas para garantir equidade de gênero [Guevara-Gómez et al. 2021].

Redes Sociais: Estudos em redes sociais revelam que algoritmos podem amplificar estereótipos sociais, de gênero e raciais, especialmente na detecção de discurso de ódio. No *Twitter*, estratégias como aprendizado por comitê e substituição de termos sensíveis por *tokens* neutros foram eficazes na redução do viés de gênero, tornando os sistemas mais justos e éticos [Nascimento et al. 2022].

Saúde: Na saúde, oito estudos destacam propostas para reduzir disparidades e promover equidade. Um *framework* que combina otimização *in-processing*, *Equalized Odds Disparity* e SHAP, aplicado à previsão de conclusão de tratamentos para transtornos por uso de substâncias (SUD), ajustou um modelo de regressão logística para minimizar disparidades entre grupos sensíveis [Lucas et al. 2024]. O uso de LLMs também tem sido explorado, embora enfrente riscos de reprodução de vieses e exclusão digital. Estratégias como engenharia de *prompts* focada na equidade, fontes culturalmente adaptadas, transparência e monitoramento contínuo foram propostas para mitigar esses riscos [Rodriguez et al. 2024]. Outros trabalhos defendem *framework*s éticos robustos, baseados em não maleficência, relevância, prestação de contas, transparência e justiça, incluindo práticas como coleta de dados demográficos e auditorias locais [McCradden et al. 2020]. A questão da deficiência é abordada com um *framework* que considera escopo, dados, uso e definição operacional, defendendo design participativo liderado por pessoas com deficiência [Newman-Griffis et al. 2022]. Também se destaca o *framework* "*GREAT PLEA*", que adapta princípios éticos de IA militar para IA generativa na saúde, com foco em governabilidade, equidade e empatia [Oniani et al. 2023]. Um estudo com dados da China revelou que o sistema de seguro médico beneficia desproporcionalmente grupos de maior renda, evidenciando como modelos podem reproduzir desigualdades socioeconômicas [Gao et al. 2021]. Os estudos da área da saúde apresentam

maior diversidade de estratégias de mitigação, combinando abordagens técnicas, princípios éticos e mecanismos de governança. Apesar disso, observa-se que grande parte das propostas ainda permanece em nível experimental, com pouca validação em contextos reais de prestação de serviços públicos.

Seguridade Social: Na seguridade social, modelos preditivos podem reproduzir desigualdades estruturais. Em Portugal, uma abordagem com *XGBoost* e SHAP aplicada à previsão de desemprego de longo prazo revelou discriminação contra idosos e imigrantes. O estudo utilizou validação cruzada, balanceamento de classes e auditoria de viés com Aequitas, reforçando a necessidade de modelos transparentes e justos [Zhao 2020].

Em conjunto, os estudos analisados demonstram um crescente interesse na mitigação de vieses em sistemas de IA aplicados ao setor público. Observa-se convergência quanto à importância de transparência, auditabilidade e equidade, embora existam divergências em relação às estratégias adotadas, que variam entre soluções técnicas, mecanismos regulatórios e abordagens centradas em governança. Além disso, a literatura ainda apresenta limitações relacionadas à replicabilidade, validação empírica e aplicação prática das propostas em contextos reais de governo digital.

5. Discussão

Ao analisar os estudos selecionados, observa-se que, embora haja avanços importantes, ainda existem fragilidades significativas na forma como o tema tem sido abordado. Um dos pontos que chama atenção é a baixa preocupação com a replicabilidade dos experimentos. Muitos trabalhos apresentam propostas de modelos, *frameworks* e metodologias, mas não disponibilizam os códigos, os dados ou sequer detalham de forma suficiente os procedimentos adotados [Utomo et al. 2022, Patrikar et al. 2023, Zhao 2020]. Essa limitação reduz a capacidade de auditoria independente e dificulta a consolidação de mecanismos institucionais de controle no setor público. Tal fragilidade torna-se particularmente crítica no contexto do governo digital, em que sistemas baseados em IA impactam diretamente a prestação de serviços públicos e devem atender a requisitos de transparência, auditabilidade e *accountability*. Nesse cenário, torna-se fundamental que órgãos públicos estabeleçam mecanismos formais de supervisão, auditoria e avaliação contínua de sistemas algorítmicos empregados em processos decisórios automatizados.

Além disso, a forma como as variáveis sensíveis são tratadas aparece de maneira superficial na maior parte dos estudos [Torres-Berru et al. 2023, O'Donnell 2019, Gao et al. 2021]. Em muitos casos, a estratégia adotada é simplesmente excluir essas variáveis dos modelos, sem discutir de forma mais aprofundada como *proxies* presentes nos dados continuam carregando esses atributos de maneira implícita. Como consequência, padrões discriminatórios podem ser mantidos mesmo quando, aparentemente, as variáveis sensíveis foram removidas [Liu et al. 2019, Cheng et al. 2023, Malhotra and Misra 2022], comprometendo a equidade nas decisões automatizadas que afetam diretamente cidadãos e seu acesso a serviços públicos. Tais limitações podem afetar diretamente a distribuição de recursos públicos, o acesso a benefícios sociais e a confiança da população nas instituições governamentais.

Verifica-se que a literatura ainda privilegia abordagens predominantemente técnicas, com menor atenção às dimensões institucionais, jurídicas e organizacionais envolvidas na implementação de IA no setor público. Como consequência, muitos estu-

dos propõem mecanismos de mitigação sem considerar desafios relacionados à governança, supervisão humana, *accountability* e capacidade institucional dos órgãos públicos. Adicionalmente, um ponto recorrente é a limitação dos conjuntos de dados empregados nos estudos. É comum que as bases de dados não reflitam adequadamente a diversidade social, apresentando amostras desbalanceadas, recortes geográficos restritos ou sub-representação de grupos minorizados [Nascimento et al. 2022, Lucas et al. 2024, Guevara-Gómez et al. 2021]. Essa fragilidade, que raramente considera uma abordagem interseccional, compromete a capacidade dos modelos de generalizar resultados e de promover equidade na prestação de serviços públicos em diferentes contextos sociais. Embora diversas estratégias técnicas, como balanceamento de classes, uso de dados sintéticos ou ajustes no treinamento, sejam propostas[Pérez-Suay et al. 2017, Rodriguez et al. 2024, Nwafor 2024], a maioria dessas soluções permanece restrita ao campo experimental. Poucos estudos demonstram, na prática, a efetividade dessas abordagens em cenários reais de governo, o que evidencia um distanciamento entre o desenvolvimento técnico e sua aplicação em contextos institucionais.

Esses achados reforçam uma lacuna crítica na literatura: a necessidade de pesquisas que não apenas avancem em termos técnicos, mas que também considerem as dimensões institucionais, sociais e organizacionais que permeiam o uso da IA no setor público. Nesse sentido, a mitigação de vieses deve ser compreendida como uma questão central para a governança no contexto do governo digital, envolvendo não apenas modelos e dados, mas também processos, atores e mecanismos de controle capazes de garantir decisões mais justas, transparentes e alinhadas ao interesse público [Mitrou et al. 2021, Bannister and Connolly 2020]. Assim, no contexto do governo digital, a mitigação de vieses deve ser tratada não apenas como um problema técnico, mas como um requisito essencial para fortalecer a legitimidade, a transparência e a confiança nas decisões automatizadas do setor público.

6. Conclusão e Trabalhos Futuros

Este trabalho, através de uma RSL, apresentou um panorama sobre como a literatura aborda o viés e a discriminação em sistemas de IA aplicados ao setor público. Os resultados indicam que, embora haja avanços em estratégias técnicas de mitigação, como otimização de modelos e uso de dados sintéticos, a maioria das soluções ainda permanece em nível experimental. Evidencia-se uma lacuna entre a mitigação técnica e a implementação de uma governança algorítmica robusta. No contexto do governo digital, isso implica a necessidade de estruturas institucionais capazes de monitorar, auditar e justificar decisões automatizadas que impactam direitos e serviços públicos. A baixa replicabilidade dos estudos e o tratamento superficial das variáveis sensíveis dificultam a aplicação dessas soluções na prática, indicando que a justiça algorítmica deve ser tratada como um requisito organizacional no contexto do governo digital.

Como agenda futura, destaca-se a necessidade de avançar da experimentação para a aplicação de soluções em contextos reais do setor público, com mecanismos auditáveis e alinhados a princípios de transparência e *accountability*. Isso inclui a adoção de práticas de governança algorítmica, como auditorias de viés, avaliações de impacto e integração com marcos regulatórios existentes. Além disso, a explicabilidade deve ser considerada um requisito essencial, garantindo que sistemas de IA sejam compreensíveis não apenas para especialistas, mas também para gestores públicos e cidadãos. Por fim, pesquisas

futuras podem explorar os impactos da IA generativa no setor público e fortalecer práticas de ciência aberta, como a disponibilização de dados e artefatos de pesquisa, contribuindo para o desenvolvimento de sistemas de IA mais transparentes, auditáveis e alinhados aos princípios do governo digital.

Referências

Aldoseri, A., Al-Khalifa, K. N., and Hamouda, A. M. (2023). Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. *Applied Sciences*, 13(12):7082.

Balayn, A., Lofi, C., and Houben, G.-J. (2021). Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, 30(5):739–768.

Bannister, F. and Connolly, R. (2020). Administration by algorithm: A risk management framework. *Information Polity*, 25(4):471–490.

Cheng, L., Gallegos, I. O., Ouyang, D., Goldin, J., and Ho, D. (2023). How redundant are redundant encodings? blindness in the wild and racial disparity when race is unobserved. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 667–686.

Fountain, J. E. (2022). The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. *Government Information Quarterly*, 39(2):101645.

Gao, J., Chu, D., and Ye, T. (2021). Empirical analysis of beneficial equality of the basic medical insurance for migrants in china. *Discrete Dynamics in Nature and Society*, 2021(1):7708605.

Gentzel, M. (2021). Biased face recognition technology used by government: A problem for liberal democracy. *Philosophy & Technology*, 34(4):1639–1663.

Grupo de Peritos de Alto Nível sobre Inteligência Artificial (2019). Orientações Éticas para uma ia de confiança. Relatório técnico, Comissão Europeia. Disponível em: `https://op.europa.eu/pt/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1`.

Guevara-Gómez, A., de Zárate-Alcarazo, L. O., and Criado, J. I. (2021). Feminist perspectives to artificial intelligence: Comparing the policy frames of the european union and spain. *Information Polity*, 26(2):173–192.

Hoekstra, M., Van Veenstra, A. F., and Chideock, C. (2021). A typology for applications of public sector ai. In *EGOV-CeDEM-ePart-**, pages 121–128.

Johnson, K. N. and Reyes, C. L. (2021). Exploring the implications of artificial intelligence. *J. Int'l & Comp. L.*, 8:315.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Landers, R. N. and Behrend, T. S. (2023). Auditing the ai auditors: A framework for evaluating fairness and bias in high stakes ai predictive models. *American Psychologist*, 78(1):36.

Levy, K., Chasalow, K. E., and Riley, S. (2021). Algorithms and decision-making in the public sector. *Annual Review of Law and Social Science*, 17(1):309–334.

Liu, H.-W., Lin, C.-F., and Chen, Y.-J. (2019). Beyond state v loomis: artificial intelligence, government algorithmization and accountability. *International journal of law and information technology*, 27(2):122–141.

Lucas, M. M., Wang, X., Chang, C.-H., Yang, C. C., Braughton, J. E., and Ngo, Q. M. (2024). An explainablefair framework for prediction of substance use disorder treatment completion. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 157–166. IEEE.

Malhotra, P. and Misra, A. (2022). Accountability and responsibility of artificial intelligence decision-making models in indian policy landscape. In *AISafety@ IJCAI*.

McCradden, M. D., Joshi, S., Anderson, J. A., Mazwi, M., Goldenberg, A., and Zlotnik Shaul, R. (2020). Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *Journal of the American Medical Informatics Association*, 27(12):2024–2027.

Mergen, A., Çetin-Kılıç, N., and Özbilgin, M. F. (2025). Artificial intelligence and bias towards marginalised groups: Theoretical roots and challenges. In *AI and Diversity in a Datafied World of Work: Will the Future of Work be Inclusive?*, volume 12, pages 17–38. Emerald Publishing Limited.

Mitrou, L., Janssen, M., and Loukis, E. (2021). Human control and discretion in ai-driven decision-making in government. In *Proceedings of the 14th International Conference on Theory and Practice of Electronic Governance*, pages 10–16.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group*, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269.

Nascimento, F. R., Cavalcanti, G. D., and Da Costa-Abreu, M. (2022). Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications*, 201:117032.

Newman-Griffis, D., Rauchberg, J. S., Alharbi, R., Hickman, L., and Hochheiser, H. (2022). Definition drives +design: Disability models and mechanisms of bias in ai technologies. *arXiv preprint arXiv:2206.08287*.

Nwafor, I. E. (2024). Gender mainstreaming into african artificial intelligence policies: Egypt, rwanda and mauritius as case studies. *Law, Technology and Humans*, 6(2):53–68.

O'Donnell, R. M. (2019). Challenging racist predictive policing algorithms under the equal protection clause. *NYUL Rev.*, 94:544.

Oniani, D., Hilsman, J., Peng, Y., Poropatich, R. K., Pamplin, J. C., Legault, G. L., and Wang, Y. (2023). Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *NPJ Digital Medicine*, 6(1):225.

Patrikar, A. M., Mahenthiran, A., and Said, A. (2023). Leveraging synthetic data for ai bias mitigation. In *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications*, volume 12529, pages 185–190. SPIE.

PENG, L., ZHANG, Q., and LI, T. (2024). Risk of ai algorithmic discrimination embedded in government data governance and its prevention and control. *Journal of library and information science in agriculture*, 36(5):23–31.

Perez-Downes, J. C., Tseng, A. S., McConn, K. A., Elattar, S. M., Sokumbi, O., Sebro, R. A., Allyse, M. A., Dangott, B. J., Carter, R. E., and Adedinsewo, D. (2024). Mitigating bias in clinical machine learning models. *Current Treatment Options in Cardiovascular Medicine*, 26(3):29–45.

Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. (2017). Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 339–355. Springer.

Rodriguez, J. A., Alsentzer, E., and Bates, D. W. (2024). Leveraging large language models to foster equity in healthcare. *Journal of the American Medical Informatics Association*, 31(9):2147–2150.

Torres-Berru, Y., Lopez-Batista, V. F., and Zhingre, L. C. (2023). A data mining approach to detecting bias and favoritism in public procurement. *Intelligent Automation & Soft Computing*, 36(3).

Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 16(3):252–260.

Utomo, S., John, A., Rouniyar, A., Hsu, H.-C., and Hsiung, P.-A. (2022). Federated trustworthy ai architecture for smart cities. In *2022 IEEE International Smart Cities Conference (ISC2)*, pages 1–7. IEEE.

Valle-Cruz, D., Alejandro Ruvalcaba-Gomez, E., Sandoval-Almazan, R., and Ignacio Criado, J. (2019). A review of artificial intelligence in government and its potential from a public policy perspective. In *Proceedings of the 20th annual international conference on digital government research*, pages 91–99.

Vetrò, A., Santangelo, A., Beretta, E., and De Martin, J. C. (2019). Ai: from rational agents to socially responsible agents. *Digital policy, regulation and governance*, 21(3):291–304.

Wei, M. and Zhou, Z. (2022). Ai ethics issues in real world: Evidence from ai incident database. *arXiv preprint arXiv:2206.07635*.

Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, 16(3):e12962.

Zhao, L. F. (2020). Data-driven approach for predicting and explaining the risk of long-term unemployment. In *E3S Web of Conferences*, volume 214, page 01023. EDP Sciences.