REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

ÉTICA, VIESES COGNITIVOS E ALGORÍTMICOS EM SISTEMAS DE INTELIGÊNCIA ARTIFICIAL APLICADOS À SEGURANÇA PÚBLICA

ETHICS, COGNITIVE AND ALGORITHMIC BIASES IN ARTIFICIAL INTELLIGENCE SYSTEMS APPLIED TO PUBLIC SECURITY

ÉTICA, SESGOS COGNITIVOS Y ALGORÍTMICOS EN SISTEMAS DE INTELIGENCIA ARTIFICIAL APLICADOS A LA SEGURIDAD PÚBLICA

Frank Eduardo da Mata Cascaes
Mestre em Segurança Pública, Universidade Estadual do Amazonas, Brasil
E-mail: frankcascaes@gmail.com

Resumo

A inteligência artificial (IA) tem sido incorporada à segurança pública, à persecução penal e à justiça criminal, com aplicações em policiamento preditivo, reconhecimento facial e avaliação de risco de reincidência. Este artigo é uma revisão narrativa crítica atualizada, que toma como ponto de partida o mapeamento sistemático de Nobre e Leite (2025) e o complementa com literatura acadêmica selecionada, documentação institucional e fontes normativas. O objetivo é analisar como vieses de diferentes níveis — cognitivos humanos, de dados e modelagem, institucionais e sociotécnicos, e de interação humano–IA — se articulam ao longo do ciclo de vida desses sistemas e quais desafios éticos e jurídicos produzem. A análise qualitativa organiza a literatura em eixos temáticos e distingue evidência empírica, conceitual, institucional e normativa. Os resultados indicam que o risco ético não decorre apenas de "preconceitos programados", mas de uma cadeia sociotécnica na qual registros históricos, escolhas de modelagem, protocolos institucionais, usuários humanos e mecanismos de retroalimentação podem produzir ou ampliar desigualdades. As evidências sobre policiamento preditivo são heterogêneas e não autorizam conclusões universais. Conclui-se que a mitigação deve começar antes do algoritmo, pela verificação de necessidade, legalidade e proporcionalidade, e prosseguir com validação, auditoria e mecanismos de contestação, em um cenário brasileiro no qual o tratamento de dados para fins de segurança pública ainda carece de legislação específica.

Palavras-chave: Ética; Vieses Cognitivos; Vieses Algorítmicos; Inteligência Artificial; Segurança Pública.

Abstract

Artificial intelligence (AI) has been incorporated into public security, criminal prosecution and criminal justice, with applications in predictive policing, facial recognition and recidivism risk assessment. This article is an updated critical narrative review that takes the systematic mapping by Nobre and Leite (2025) as its starting point and supplements it with selected academic literature, institutional documents and normative sources. It aims to analyze how biases at different levels — human cognitive biases, data and modeling biases, institutional and sociotechnical biases, and human–AI interaction biases — interact throughout the life cycle of these systems, and which ethical and legal challenges they raise. The qualitative analysis organizes the literature into thematic axes and distinguishes empirical, conceptual, institutional and normative evidence. The findings indicate that ethical risk does not stem only from "programmed prejudice", but from a sociotechnical chain in which historical records, modeling choices, institutional protocols, human users and feedback mechanisms can produce or amplify inequalities. Evidence on predictive policing is heterogeneous and does not support universal conclusions. The article concludes that mitigation must begin before the algorithm, by assessing necessity, legality and proportionality, and continue with validation, auditing and contestation

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

mechanisms, in a Brazilian context where data processing for public security purposes still lacks specific legislation.

Keywords: Ethics; Cognitive Biases; Algorithmic Biases; Artificial Intelligence; Public Security.

Resumen

La inteligencia artificial (IA) se ha incorporado a la seguridad pública, la persecución penal y la justicia penal, con aplicaciones en vigilancia predictiva, reconocimiento facial y evaluación del riesgo de reincidencia. Este artículo es una revisión narrativa crítica actualizada que toma como punto de partida el mapeo sistemático de Nobre y Leite (2025) y lo complementa con literatura académica seleccionada, documentación institucional y fuentes normativas. El objetivo es analizar cómo sesgos de distintos niveles — cognitivos humanos, de datos y modelado, institucionales y sociotécnicos, y de interacción humano–IA — se articulan a lo largo del ciclo de vida de estos sistemas y qué desafíos éticos y jurídicos generan. El análisis cualitativo organiza la literatura en ejes temáticos y distingue evidencia empírica, conceptual, institucional y normativa. Los resultados indican que el riesgo ético no proviene solo de "prejuicios programados", sino de una cadena sociotécnica en la que registros históricos, decisiones de modelado, protocolos institucionales, usuarios humanos y mecanismos de retroalimentación pueden producir o ampliar desigualdades. La evidencia sobre la vigilancia predictiva es heterogénea y no autoriza conclusiones universales. Se concluye que la mitigación debe comenzar antes del algoritmo, verificando necesidad, legalidad y proporcionalidad, y continuar con validación, auditoría y mecanismos de impugnación, en un contexto brasileño en el que el tratamiento de datos con fines de seguridad pública aún carece de legislación específica.

Palabras clave: Ética; Sesgos Cognitivos; Sesgos Algorítmicos; Inteligencia Artificial; Seguridad Pública.

1. Introdução

1.1 Contextualização Histórica e Social

A inteligência artificial (IA) tornou-se instrumento relevante na segurança pública, influenciando desde a previsão de locais de ocorrência criminal até a identificação biométrica em espaços públicos. Instrumentos atuariais de avaliação de risco antecedem a IA contemporânea: o COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), desenvolvido nos Estados Unidos no final da década de 1990, é utilizado para estimar o risco de reincidência em decisões da justiça criminal (Angwin et al., 2016). No contexto brasileiro, o uso de reconhecimento facial por secretarias estaduais de segurança ganhou escala na segunda metade da década de 2010; em 2019, o monitoramento da Rede de Observatórios da Segurança registrou, ao longo de cinco meses, 151 prisões associadas à tecnologia em quatro estados, das quais 90,5% de pessoas negras (Nunes, 2019). Socialmente, essa integração reflete uma transição para sociedades

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

de controle, como descrito por Deleuze (1992), nas quais dados pessoais são processados para antecipar comportamentos, com riscos de discriminação.

A expansão desses sistemas envolve investimentos públicos e parcerias com fornecedores privados e, como documentado nos Estados Unidos, pode ocorrer com baixa transparência e regulação insuficiente, o que dificulta o controle público (Garvie, Bedoya e Frankle, 2016). Juridicamente, o marco brasileiro apresenta uma lacuna específica: a Lei Geral de Proteção de Dados Pessoais (LGPD, Lei nº 13.709/2018) não se aplica ao tratamento realizado para fins exclusivos de segurança pública e de investigação e repressão de infrações penais (art. 4º, III), remetendo a matéria a legislação específica (art. 4º, § 1º) ainda não editada. Na União Europeia, o Regulamento Geral sobre a Proteção de Dados (RGPD/GDPR) é complementado, para as autoridades de aplicação da lei, pela Diretiva (UE) 2016/680, e, mais recentemente, pelo Regulamento (UE) 2024/1689 sobre inteligência artificial (União Europeia, 2016a; 2016b; 2024).

Os vieses cognitivos — padrões sistemáticos de desvio no julgamento humano, descritos por Tversky e Kahneman (1974) — constituem apenas uma das camadas do problema. Vieses algorítmicos podem surgir da seleção de dados, da escolha da variável-alvo, da mensuração, do desenho do modelo e de sua implantação, sem que seja possível identificar um viés cognitivo específico de quem desenvolveu o sistema (Barocas e Selbst, 2016; Suresh e Guttag, 2021). No reconhecimento facial, a avaliação do *National Institute of Standards and Technology* (NIST) encontrou diferenciais demográficos na maioria dos algoritmos testados, com grande variação entre sistemas (Grother, Ngan e Hanaoka, 2019); em tarefa distinta — a classificação de gênero a partir de imagens faciais —, Buolamwini e Gebru (2018) identificaram erros de até 34,7% para mulheres de pele mais escura, contra até 0,8% para homens de pele mais clara.

No plano internacional, o policiamento preditivo tornou-se objeto de controvérsia: estudos de simulação indicam risco de concentração racialmente desigual do patrulhamento (Lum e Isaac, 2016), enquanto ensaios de campo relatam resultados distintos quanto à eficácia e às disparidades nas prisões (Mohler et al., 2015; Brantingham, Valasik e Mohler, 2018; Ferguson, 2017). Essa contextualização

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

revela a necessidade de examinar os vieses em IA não como falha isolada de um programa, mas como resultado de um ciclo sociotécnico que conecta instituições, dados, modelos e decisões humanas, com implicações para a justiça social e os direitos fundamentais.

1.2 Delimitação do Escopo

Este artigo adota uma delimitação ampla, mas explícita, que abrange três campos relacionados e juridicamente distintos: (i) a **segurança pública** em sentido estrito, que compreende policiamento ostensivo e preventivo, vigilância e identificação biométrica em espaços públicos; (ii) a **persecução penal**, que abrange investigação criminal e identificação de suspeitos; e (iii) a **justiça criminal**, que inclui instrumentos de avaliação de risco empregados em decisões de fiança, sentença e execução penal. O COMPAS, por exemplo, situa-se predominantemente neste último campo, e não no patrulhamento policial. Casos de administração pública automatizada fora desses campos, como sistemas de detecção de fraudes em benefícios sociais, são mencionados apenas como analogias, sinalizadas como tal no texto.

2. Problema de Pesquisa

O problema central desta pesquisa reside no modo como vieses de diferentes naturezas — cognitivos, algorítmicos, institucionais e de interação humano–IA — se combinam em sistemas de IA aplicados à segurança pública, à persecução penal e à justiça criminal, podendo perpetuar desigualdades sociais e violar princípios como igualdade, devido processo e presunção de inocência. A questão de pesquisa guia-se pela seguinte formulação: como esses níveis de viés se articulam ao longo do ciclo sociotécnico dos sistemas de IA nesses campos, quais desafios éticos e jurídicos produzem e que salvaguardas podem preveni-los ou mitigá-los?

Essa questão emerge de evidências de naturezas distintas: estudos que documentam a contaminação de bases policiais por práticas ilegais ou discriminatórias (Richardson, Schultz e Crawford, 2019), avaliações técnicas de desempenho desigual de algoritmos de reconhecimento facial (Grother, Ngan e Hanaoka, 2019) e registros brasileiros sobre o perfil racial das pessoas presas com uso dessa tecnologia (Nunes, 2019). No Brasil, o problema é agravado pela ausência

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

da legislação específica prevista no art. 4º, § 1º, da LGPD: o Anteprojeto de Lei de Proteção de Dados para Segurança Pública e Investigação Criminal, entregue à Câmara dos Deputados em novembro de 2020 por comissão de juristas por ela instituída, não foi convertido em lei (Brasil, 2020b).

3. Objetivos

3.1 Objetivo Geral

Analisar, por meio de revisão narrativa crítica atualizada, os desafios éticos associados a vieses cognitivos e algorítmicos em sistemas de IA aplicados à segurança pública, à persecução penal e à justiça criminal.

3.2 Objetivos Específicos

- Diferenciar conceitualmente vieses cognitivos humanos, vieses de dados e modelagem, vieses institucionais e sociotécnicos e vieses de interação humano–IA.
- Descrever como esses níveis se articulam e se reforçam ao longo do ciclo sociotécnico, com ênfase em policiamento preditivo, avaliação de risco criminal e reconhecimento facial.
- Avaliar impactos éticos, sociais e jurídicos, considerando os marcos normativos brasileiro, europeu e internacional.
- Propor uma sequência de salvaguardas que inclua necessidade, legalidade e proporcionalidade antes das estratégias técnicas de mitigação.
- Identificar lacunas empíricas e normativas no contexto brasileiro.

4. Justificativa

A relevância acadêmica e social desta pesquisa justifica-se pela expansão da IA na segurança pública em um contexto de salvaguardas normativas incompletas. Decisões automatizadas ou apoiadas por algoritmos nesses campos podem afetar a igualdade (Constituição Federal, art. 5º, *caput*), o devido processo legal, o contraditório e a ampla defesa (art. 5º, LIV e LV), a presunção de inocência (art. 5º, LVII) e a proteção de dados pessoais, reconhecida como direito fundamental autônomo pelo Supremo Tribunal Federal no referendo da medida cautelar na ADI

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

6.387 (Brasil, 2020a) e incorporada ao texto constitucional pela Emenda Constitucional nº 115/2022 (art. 5º, LXXIX). No plano internacional, o Alto Comissariado das Nações Unidas para os Direitos Humanos recomendou moratória no uso de reconhecimento biométrico remoto em espaços públicos até que se demonstre o cumprimento de padrões de privacidade, proteção de dados e não discriminação (OHCHR, 2021).

Academicamente, o estudo contribui para o campo do direito digital e da ética da IA, dialogando com trabalhos como o de Requião (2024), sobre a relação entre IA e vieses cognitivos em decisões judiciais. Sua contribuição específica consiste em articular, em um mesmo quadro analítico, os diferentes níveis de viés e a cadeia sociotécnica pela qual se reforçam. Na prática, pretende subsidiar gestores, operadores do direito e legisladores no debate sobre a legislação específica de proteção de dados em segurança pública e sobre o marco regulatório da IA em tramitação no Congresso Nacional.

5. Metodologia Resumida

Trata-se de revisão narrativa crítica atualizada, de abordagem qualitativa, que toma como ponto de partida o mapeamento sistemático de Nobre e Leite (2025) — Trabalho de Conclusão de Curso de Bacharelado em Engenharia de Software da Universidade de Brasília — e o complementa com literatura acadêmica selecionada, documentação institucional e fontes normativas. O artigo não constitui revisão sistemática original e não reproduz os resultados quantitativos daquele mapeamento como resultados próprios. O percurso metodológico, os critérios de seleção e a classificação das fontes segundo sua natureza estão detalhados na seção 8.

6. Fundamentação Teórica

6.1 Fundações Éticas na IA e Segurança Pública

A ética em IA na segurança pública é ancorada em correntes filosóficas clássicas, adaptadas a contextos tecnológicos. O utilitarismo, proposto por Bentham (1789) e Mill (1863), avalia ações pela maximização do bem-estar coletivo, o que suscita a questão de saber se sistemas de IA que prometem reduzir crimes justificam

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

efeitos discriminatórios (Bostrom e Yudkowsky, 2011). Na segurança pública, isso se aplica a algoritmos que otimizam a alocação de recursos policiais, mas podem concentrar ônus sobre determinados grupos e territórios.

A deontologia, inspirada em Kant (1785), enfatiza deveres como transparência e não discriminação, independentemente de resultados (Floridi e Sanders, 2004). No Brasil, essa perspectiva encontra eco na LGPD, cujo art. 20 assegura a revisão de decisões tomadas unicamente com base em tratamento automatizado nas relações por ela reguladas; como se verá, esse dispositivo não alcança diretamente o tratamento realizado para fins exclusivos de segurança pública, excluído pelo art. 4º, III. A ética das virtudes, de Aristóteles (século IV a.C.), foca no desenvolvimento de caráter virtuoso em projetistas e usuários de IA, promovendo a justiça como virtude central (Vallor, 2016).

Essas correntes convergem em propostas contemporâneas de governança. Floridi et al. (2018), no projeto AI4People, sintetizam os princípios de beneficência, não maleficência, autonomia e justiça, oriundos da bioética, e acrescentam a explicabilidade como princípio próprio da IA. Floridi e Cowls (2019) consolidam esses cinco princípios em um quadro unificado, a partir da comparação de diversos documentos internacionais. A UNESCO (2021), em recomendação adotada por seus Estados-membros, afirma a proporcionalidade e a supervisão humana como exigências centrais e estabelece que sistemas de IA não devem ser usados para pontuação social ou vigilância em massa. Coeckelbergh (2020) observa que tais princípios só adquirem sentido quando articulados a questões concretas de responsabilidade, poder e justiça, o que é especialmente sensível na segurança pública.

6.2 Correntes Teóricas Éticas: Utilitarismo, Deontologia e Virtudes

O utilitarismo prioriza resultados, mas críticos como Rawls (1971) argumentam que ele pode sacrificar os interesses de minorias, como em aplicações de IA que concentram a vigilância em áreas pobres. A deontologia exige regras firmes, mas enfrenta desafios em cenários ambíguos, como decisões em tempo real na atividade policial (Jobin et al., 2019). A ética das virtudes integra ambas, propondo formação ética para desenvolvedores, conforme as diretrizes do IEEE (2019).

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

Binns (2018) mostra que as discussões técnicas sobre justiça algorítmica reproduzem, muitas vezes sem saber, debates antigos da filosofia política, como a tensão entre igualdade de oportunidades e igualdade de resultados, o que reforça a necessidade de explicitar qual concepção de justiça orienta cada sistema. Russell (2019) desloca a questão para o problema do alinhamento de valores, defendendo máquinas que permaneçam incertas quanto às preferências humanas e, por isso, abertas à correção. Danaher (2016), por sua vez, descreve a ameaça da algocracia: quando decisões públicas passam a ser organizadas por algoritmos opacos, a participação e a compreensão dos cidadãos ficam limitadas, comprometendo a legitimidade dos procedimentos decisórios do Estado.

6.3 Debates Acadêmicos Atuais sobre Ética em IA

Debates atuais incluem a tensão entre inovação e regulação. Crawford (2021) critica o "mito da neutralidade" da IA, enquanto Zuboff (2019) discute o capitalismo de vigilância. No âmbito da segurança pública, o Alto Comissariado das Nações Unidas para os Direitos Humanos recomendou que os Estados proíbam aplicações de IA incompatíveis com o direito internacional dos direitos humanos e adotem moratória para o reconhecimento biométrico remoto em espaços públicos até que salvaguardas adequadas estejam em vigor (OHCHR, 2021).

A proliferação de diretrizes éticas não resolveu esses dilemas. Mittelstadt et al. (2016) mapearam as preocupações éticas associadas aos algoritmos, distinguindo problemas epistêmicos, como evidências inconclusivas, inescrutáveis ou equivocadas, de problemas normativos, como resultados injustos e efeitos transformadores sobre a autonomia. Hagendorff (2020), ao avaliar dezenas de diretrizes, concluiu que elas raramente dispõem de mecanismos de cumprimento e exercem pouca influência sobre a prática de desenvolvimento. Mittelstadt (2019) acrescenta que princípios, isoladamente, não garantem uma IA ética, pois o campo não possui, como a medicina, objetivos comuns, deveres fiduciários consolidados e mecanismos efetivos de responsabilização profissional.

6.4 Conceito de Inteligência Artificial

A IA é definida como o campo dedicado a sistemas capazes de realizar tarefas que requerem inteligência humana, como aprendizado e raciocínio (Russell e Norvig,

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

2020). Na segurança pública, as aplicações existentes correspondem à chamada IA fraca, voltada a tarefas específicas como o reconhecimento facial ou a previsão de locais de ocorrência; a IA forte, de capacidade generalizada, permanece hipotética.

6.5 Evolução Histórica da IA na Segurança Pública

A história da IA remonta a Turing (1950), mas, na segurança pública, ganhou tração a partir da década de 1980, com sistemas especialistas para análise criminal. Nos anos 2000 e 2010, o aprendizado de máquina passou a predominar, com ferramentas como o PredPol (2011), nos Estados Unidos, que utiliza dados históricos de ocorrências para prever locais de maior probabilidade criminal. No Brasil, a criação do Sistema Nacional de Informações de Segurança Pública (Sinesp), pela Lei nº 12.681/2012, ampliou a integração de bases de dados de segurança pública (Brasil, 2012); como essas bases derivam em grande parte de registros policiais, carregam as seletividades das práticas que as produziram, o que é relevante para qualquer uso analítico ou preditivo posterior.

Os estudos empíricos sobre essa geração de sistemas trazem resultados divergentes, examinados no Eixo II da seção 9. Joh (2014) aponta que essas ferramentas tensionam categorias jurídicas tradicionais, como a suspeita fundada que justifica a abordagem policial.

6.6 Modelos de IA Aplicados: Aprendizado de Máquina e Redes Neurais

O aprendizado de máquina envolve algoritmos que "aprendem" padrões a partir de dados e que, por isso, são suscetíveis a reproduzir distorções presentes nos conjuntos de treinamento (Mehrabi et al., 2021). Redes neurais profundas, base dos sistemas contemporâneos de reconhecimento facial, podem apresentar desempenho desigual entre grupos demográficos quando os dados de treinamento e de avaliação são desbalanceados; a magnitude dessas diferenças, contudo, varia amplamente entre algoritmos e deve ser verificada sistema a sistema (Grother, Ngan e Hanaoka, 2019).

Esses modelos são frequentemente opacos. Burrell (2016) distingue três formas de opacidade: o segredo corporativo deliberado, o analfabetismo técnico do público e a opacidade inerente à escala e à complexidade do aprendizado de máquina. Pasquale (2015) descreve uma sociedade da caixa-preta, na qual algoritmos que

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

afetam direitos permanecem protegidos do escrutínio público. O risco de variáveis substitutas enviesadas é ilustrado, por analogia, por Obermeyer et al. (2019): um algoritmo de saúde amplamente utilizado nos Estados Unidos, ao adotar o custo do tratamento como indicador de necessidade, subestimou a gravidade clínica de pacientes negros. Na segurança pública, o uso de registros de prisão como indicador de criminalidade pode produzir distorção análoga, pois esses registros medem a atividade policial tanto quanto a ocorrência de crimes.

6.7 Níveis Analíticos do Viés: Distinções Conceituais

A literatura frequentemente reúne sob o termo "viés" fenômenos que pertencem a níveis analíticos distintos. Viés de confirmação, ancoragem, efeito de enquadramento e viés de automação referem-se a padrões de julgamento humano; viés histórico, de representação, de mensuração, de agregação, de avaliação e de implantação referem-se a propriedades de dados, modelos e contextos de uso. Esses fenômenos se relacionam, mas não são sinônimos: um conjunto de dados pode produzir discriminação algorítmica sem que se identifique um viés cognitivo específico de qualquer desenvolvedor, e um modelo tecnicamente bem calibrado pode ser empregado por um agente humano de maneira enviesada. Para evitar essa confusão, este artigo distingue quatro níveis, sintetizados no Quadro 1.

Quadro 1 – Níveis analíticos do viés em sistemas de IA aplicados à segurança pública

Nível	Definição	Exemplos na segurança pública	Referências
Vieses cognitivos humanos	Padrões sistemáticos de desvio no julgamento de pessoas (desenvolvedores, gestores, policiais, juízes).	Confirmação, ancoragem, enquadramento de uma pontuação como "alto risco".	Tversky e Kahneman (1974; 1981); Kahneman (2011)
Vieses de dados e modelagem	Distorções originadas na amostragem, na escolha da variável-alvo, na mensuração, na agregação, no treinamento e na avaliação do modelo.	Registros de prisão usados como indicador de criminalidade; base de imagens desbalanceada.	Suresh e Guttag (2021); Barocas e Selbst (2016); Mehrabi et al. (2021)
Vieses institucionais e sociotécnicos	Práticas, prioridades e relações de poder que definem o que é registrado, onde se policia e para que o sistema é adotado.	"Dados sujos" de práticas ilegais; seletividade territorial do policiamento; protocolos de uso.	Friedman e Nissenbaum (1996); Richardson, Schultz e Crawford (2019); Almeida (2019)
Vieses de interação humano–IA	Distorções que surgem quando agentes interpretam, aceitam ou contestam recomendações algorítmicas, e retroalimentação daí resultante.	Viés de automação diante de alerta facial; desvios desiguais da recomendação; laços de retroalimentação.	Skitka, Mosier e Burdick (1999); Green e Chen (2019); Ensign et al. (2018)

Fonte: elaboração própria, com base nas obras indicadas.

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

O primeiro nível, dos **vieses cognitivos humanos**, diz respeito ao julgamento de pessoas — desenvolvedores, gestores, policiais, promotores, juízes. O segundo, dos **vieses de dados e modelagem**, localiza-se nas etapas técnicas do ciclo de vida do aprendizado de máquina e pode existir independentemente de qualquer intenção ou erro de julgamento individual. O terceiro, dos **vieses institucionais e sociotécnicos**, refere-se às práticas, prioridades e distribuições de poder que determinam quais eventos são registrados, onde se policia e para que fim o sistema é adquirido. O quarto, dos **vieses de interação humano–IA**, surge no momento em que uma recomendação algorítmica é interpretada, aceita, ignorada ou contestada por um agente humano.

Somente após essa distinção é possível analisar como os níveis se reforçam mutuamente. Registros policiais produzidos sob práticas seletivas (nível institucional) tornam-se dados de treinamento não representativos (nível de dados); o modelo aprende essa distribuição (nível de modelagem); o agente que recebe a previsão tende a aceitá-la e a interpretá-la segundo expectativas prévias (níveis de interação e cognitivo); e as ações resultantes geram novos registros que retroalimentam o ciclo. A seção 7 desenvolve essa cadeia.

6.8 Vieses Cognitivos Humanos e Interação Humano–IA

Vieses cognitivos são desvios sistemáticos no processamento humano de informações (Kahneman, 2011). Tversky e Kahneman (1974) descreveram as heurísticas de representatividade, disponibilidade e ancoragem, esta última caracterizada pela dependência excessiva de um valor inicial de referência. O viés de confirmação — a tendência a buscar e privilegiar informações que confirmam crenças prévias — provém de tradição de pesquisa distinta, mas é igualmente relevante na atividade policial e investigativa. Na segurança pública, esses vieses atuam sobre quem define o problema a ser resolvido pelo sistema, quem seleciona e rotula dados e quem utiliza as previsões.

Três fenômenos merecem destaque no nível da interação humano–IA. O primeiro é o viés de automação, descrito por Skitka, Mosier e Burdick (1999), que consiste na tendência de aceitar recomendações automatizadas mesmo diante de informações contrárias, gerando erros de omissão e de comissão. O segundo é o

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

efeito de enquadramento, pelo qual Tversky e Kahneman (1981) demonstraram que a forma de apresentação de uma mesma informação altera a decisão; uma pontuação de risco apresentada como "alto risco" tende a orientar o agente de modo diverso de uma probabilidade numérica equivalente. O terceiro é o ruído, definido por Kahneman, Sibony e Sunstein (2021) como a variabilidade indesejada entre julgamentos que deveriam ser idênticos, frequentemente invocado para justificar a automação, mas que não se confunde com o viés e não é por ela automaticamente resolvido. Green e Chen (2019) mostraram, em experimento, que pessoas assistidas por uma avaliação de risco dela se desviaram de maneira racialmente desigual, o que evidencia que a supervisão humana, por si só, não neutraliza vieses e pode introduzir outros.

6.9 Classificações Doutrinárias dos Vieses Algorítmicos

Friedman e Nissenbaum (1996) já distinguiam, em sistemas computacionais, o viés preexistente, originado em instituições e práticas sociais, o viés técnico, decorrente de limitações do projeto, e o viés emergente, que surge do uso do sistema em contexto diverso do original. Suresh e Guttag (2021) propõem um modelo das fontes de dano ao longo do ciclo de vida do aprendizado de máquina, identificando vieses histórico, de representação, de mensuração, de agregação, de aprendizado, de avaliação e de implantação, o que permite localizar com maior precisão onde a intervenção corretiva é necessária.

Nobre e Leite (2025), em mapeamento sistemático sobre desafios éticos associados a vieses em algoritmos de IA, organizaram sua busca em torno de três categorias: viés de agregação (combinação inadequada de dados heterogêneos), viés de interação com o usuário (ciclos em que o comportamento dos usuários e as respostas do sistema se retroalimentam) e viés comportamental (reprodução de padrões de comportamento e julgamento humanos). Essas categorias atravessam os níveis do Quadro 1: a agregação pertence ao nível de dados e modelagem; a interação combina os níveis de interação humano–IA e de retroalimentação institucional; e o viés comportamental aproxima-se, mas não se identifica inteiramente, do nível cognitivo.

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

6.10 Da Transferência de Vieses ao Ciclo Sociotécnico

A reprodução de desigualdades por sistemas algorítmicos ocorre, em grande medida, por meio de dados históricos que refletem práticas sociais e institucionais (Barocas e Selbst, 2016). Caliskan, Bryson e Narayanan (2017) demonstraram que modelos de linguagem treinados em grandes corporações absorvem associações estereotipadas semelhantes às medidas em testes de associação implícita com seres humanos, evidenciando que esse processo ocorre mesmo sem intenção dos desenvolvedores. No campo policial, Richardson, Schultz e Crawford (2019) mostram que jurisdições norte-americanas investigadas por práticas ilegais ou discriminatórias produziram "dados sujos", isto é, registros contaminados por essas práticas, que depois alimentaram sistemas preditivos. Harcourt (2007) já advertia que métodos atuariais aplicados ao policiamento podem gerar um efeito catraca, concentrando a atenção sobre grupos já supervigiados e ampliando sua representação na população carcerária, sem ganhos correspondentes na redução do crime.

Por isso, a expressão "transferência de vieses cognitivos para a IA" é insuficiente para descrever o fenômeno. Mais do que um viés psicológico transposto para o código, trata-se de um ciclo sociotécnico no qual instituições, dados, modelos e usuários se influenciam reciprocamente.

7. O Ciclo Sociotécnico do Viés na Segurança Pública

7.1 Da Instituição à Retroalimentação: o Caso do Policiamento Preditivo

A cadeia pela qual o viés se forma e se reproduz pode ser descrita em seis etapas: instituição → geração dos dados → modelagem → implantação → interpretação humana → retroalimentação institucional. No policiamento preditivo, essa cadeia pode assumir a seguinte forma:

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

Registros policiais historicamente desiguais
↓
Base de treinamento não representativa da criminalidade real
↓
Previsão de maior risco no território já policiado
↓
Maior patrulhamento nesse território
↓
Mais registros no mesmo território
↓
Realimentação do modelo

Fonte: elaboração própria, com base em Lum e Isaac (2016), Ensign et al. (2018) e Richardson, Schultz e Crawford (2019).

Lum e Isaac (2016) simularam a aplicação de um algoritmo de policiamento preditivo a dados de crimes de drogas em Oakland e constataram que o sistema direcionaria o patrulhamento desproporcionalmente para bairros de maioria negra, embora estimativas de saúde pública indicassem consumo distribuído de forma mais equilibrada pela cidade. Ensign et al. (2018) formalizaram esse fenômeno como laço de retroalimentação descontrolado: como o sistema só observa crimes nos locais para onde envia policiais, suas previsões tendem a se confirmar e a se intensificar, independentemente da distribuição real da criminalidade. Nesse mecanismo, o problema não é apenas um viés cognitivo transferido ao algoritmo, mas a interação entre práticas institucionais, dados produzidos por essas práticas e decisões operacionais baseadas nas previsões.

7.2 Tipologias Jurídicas da Discriminação Algorítmica

No direito brasileiro, Mendes e Mattiuzzo (2019) propõem uma tipologia da discriminação algorítmica que distingue a discriminação por erro estatístico, por generalização, pelo uso de informações sensíveis e a discriminação limitadora do exercício de direitos, categorias úteis para enquadrar juridicamente os níveis de viés descritos no Quadro 1. Em relatório ao Conselho da Europa, Zuiderveen Borgesius (2018) observa que a legislação antidiscriminatória e a de proteção de dados oferecem proteção relevante, mas insuficiente, sobretudo quando a discriminação ocorre por meio de variáveis aparentemente neutras que funcionam como substitutas de características protegidas.

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

7.3 Do Diferencial de Desempenho ao Dano: o Caso do Reconhecimento Facial

No reconhecimento facial, diferenças de desempenho do algoritmo entre grupos demográficos não significam, automaticamente, prisões injustas. Entre o diferencial técnico e o dano concreto interpõem-se fatores como a qualidade da imagem capturada, o limiar de similaridade adotado, o tamanho e a composição da base de referência, a existência e a qualidade da revisão humana, a realização de investigação complementar e o protocolo institucional que define o que um alerta autoriza fazer. Cada um desses elementos pode atenuar ou agravar o risco: um limiar permissivo aplicado a uma base extensa, sem revisão qualificada e com um protocolo que trate o alerta como fundamento suficiente para abordagem ou prisão, converte uma diferença estatística em dano individual.

Essa precisão também se aplica à interpretação de casos jurídicos. No caso SyRI, o Tribunal Distrital de Haia declarou, em 2020, a incompatibilidade de um sistema governamental de detecção de fraudes em benefícios sociais com o direito à vida privada previsto no art. 8º da Convenção Europeia de Direitos Humanos, em razão de sua falta de transparência e de salvaguardas (Rechtbank Den Haag, 2020). Embora situado no campo da administração pública, e não da segurança pública, o caso ilustra os riscos de sistemas opacos de perfilamento estatal.

8. Metodologia

8.1 Tipo de Pesquisa

Pesquisa bibliográfica e documental, exploratória e descritiva, configurada como revisão narrativa crítica atualizada. Optou-se por essa modalidade, e não pela revisão sistemática, por duas razões. Primeiro, o objetivo é interpretativo e conceitual — articular níveis de viés e salvaguardas —, o que exige integrar literatura empírica, conceitual, institucional e normativa, heterogênea demais para síntese sistemática. Segundo, o estudo não pretende estimar a frequência de temas ou de tipos de viés na literatura, finalidade para a qual a revisão sistemática seria indispensável. Assim, o artigo não segue o protocolo PRISMA e não apresenta síntese quantitativa própria.

O ponto de partida é o mapeamento sistemático de Nobre e Leite (2025), Trabalho de Conclusão de Curso de Bacharelado em Engenharia de Software

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

apresentado à Universidade de Brasília em julho de 2025. Segundo as autoras, a busca foi realizada na base Scopus no início de 2025 e resultou em 814 registros, dos quais 457 estudos foram selecionados para análise. Esse mapeamento não tem como objeto específico a segurança pública; por isso, é utilizado como fonte secundária para enquadramento conceitual e identificação de literatura, e seus resultados são sempre atribuídos às autoras, não sendo apresentados como resultados desta pesquisa.

8.2 Abordagem

Qualitativa, mediante análise temática interpretativa das fontes selecionadas. Não se adota abordagem mista, uma vez que o estudo não produz contagens, distribuições ou estatísticas próprias; os dados numéricos mencionados no texto provêm das fontes citadas e são a elas atribuídos.

8.3 Instrumentos e Procedimentos

8.3.1 Fontes e Estratégia de Identificação

A identificação das fontes combinou três procedimentos: (a) consulta ao mapeamento de Nobre e Leite (2025), cuja busca foi realizada na base Scopus, e às suas referências; (b) busca retrospectiva (*backward snowballing*) nas referências de obras seminais sobre viés algorítmico, policiamento preditivo, reconhecimento facial e avaliação de risco, como Barocas e Selbst (2016), Richardson, Schultz e Crawford (2019), Suresh e Guttag (2021) e Mehrabi et al. (2021); e (c) consulta direta a documentos oficiais nos portais das respectivas instituições (NIST, ACNUDH, UNESCO, EUR-Lex, Planalto, Senado Federal, Câmara dos Deputados, STF e ANPD). Não foi realizada busca sistemática independente em bases de dados; por essa razão, não se apresentam strings booleanas, datas de busca, contagens de registros por base, etapas de deduplicação e triagem ou número de revisores, elementos próprios da revisão sistemática e não aplicáveis a este desenho. A única base bibliográfica subjacente ao ponto de partida é a Scopus, conforme informado por Nobre e Leite (2025).

8.3.2 Critérios de Seleção

Foram selecionadas fontes que atendessem a três critérios: (i) pertinência temática, isto é, tratar de IA ou decisão automatizada e de viés, discriminação ou

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

direitos fundamentais em segurança pública, persecução penal ou justiça criminal, ou oferecer fundamento conceitual diretamente aplicável a esses campos; (ii) contribuição reconhecida, empírica ou conceitual, para o debate; e (iii) rastreabilidade, entendida como a possibilidade de o leitor identificar e localizar exatamente o documento citado. Obras seminais foram incluídas independentemente da data; para a literatura empírica e normativa, priorizou-se o período de 2016 a 2026. Estudos não empíricos não foram excluídos, pois a revisão narrativa incorpora literatura conceitual e normativa; essas fontes, porém, são distinguidas das evidências empíricas, conforme o Quadro 2. Estatísticas institucionais cuja fonte exata não pôde ser verificada foram excluídas.

8.3.3 Classificação das Fontes segundo sua Natureza

Para evitar atribuir peso empírico equivalente a materiais de naturezas distintas, as fontes foram classificadas em quatro categorias, apresentadas no Quadro 2. O mapeamento de Nobre e Leite (2025) é tratado à parte, como fonte secundária de enquadramento, e não como evidência primária.

Quadro 2 – Classificação das fontes segundo sua natureza e função no argumento

Categoria	Função no argumento	Exemplos de fontes utilizadas
Evidência empírica	Sustenta afirmações sobre ocorrência, magnitude e mecanismos de viés; sintetizada no Quadro 3.	Angwin et al. (2016); Larson et al. (2016); Lum e Isaac (2016); Mohler et al. (2015); Brantingham, Valasik e Mohler (2018); Buolamwini e Gebru (2018); Grother, Ngan e Hanaoka (2019); Green e Chen (2019); Brayne (2021); Nunes (2019)
Literatura conceitual	Fornece categorias analíticas, fundamentos éticos e interpretações teóricas; não é tratada como prova empírica.	Kant (1785); Rawls (1971); Friedman e Nissenbaum (1996); Barocas e Selbst (2016); Suresh e Guttag (2021); Floridi et al. (2018); Almeida (2019); Benjamin (2019)
Documentação institucional	Registra posições, diagnósticos e recomendações de organismos públicos e da sociedade civil.	OHCHR (2021); UNESCO (2021); ANPD (2024); Garvie, Bedoya e Frankle (2016); Zuiderveen Borgesius (2018); Dieterich, Mendoza e Brennan (2016)
Fontes normativas	Definem o regime jurídico aplicável e seus limites; interpretadas segundo seu texto e estágio de vigência.	Constituição Federal; LGPD; EC nº 115/2022; ADI 6.387; PL nº 2.338/2023; RGPD; Diretiva (UE) 2016/680; Regulamento (UE) 2024/1689; caso SyRI

Fonte: elaboração própria.

8.3.4 Procedimento de Análise

As fontes foram analisadas por meio de codificação temática interpretativa, organizada em sete eixos: (I) fontes e mecanismos de viés; (II) policiamento preditivo e sistemas de avaliação de risco; (III) reconhecimento facial e biometria; (IV) interação humano–algoritmo e viés de automação; (V) impactos jurídicos e sociais; (VI)

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

estratégias de mitigação e governança; e (VII) contexto brasileiro e lacunas empíricas. Para os estudos empíricos, extraíram-se autor e ano, país, aplicação de IA, contexto, método, principal resultado, nível de viés envolvido e limitação, informações reunidas na matriz do Quadro 3.

8.4 Limitações do Estudo

Por se tratar de revisão narrativa, a seleção das fontes envolve julgamento dos autores e não é reprodutível no mesmo grau de uma revisão sistemática, estando sujeita a viés de seleção. Somam-se o viés de publicação, a restrição a textos em português, inglês e espanhol e a predominância de evidências empíricas produzidas nos Estados Unidos, cuja transposição ao contexto brasileiro exige cautela. O mapeamento de Nobre e Leite (2025), usado como ponto de partida, não é específico da segurança pública e utilizou uma única base. Por fim, o estudo não produz dados primários e, no Brasil, há escassez de dados públicos auditados sobre o desempenho dos sistemas em uso.

9. Resultados e Discussão

Os resultados são apresentados em sete eixos temáticos. O Quadro 3 reúne a matriz das principais evidências empíricas discutidas, que servem de base para os eixos seguintes.

Quadro 3 – Matriz das principais evidências empíricas discutidas

Estudo	País / campo	Aplicação de IA	Método	Principal resultado	Nível de viés	Limitação
Angwin et al. (2016); Larson et al. (2016)	EUA (Broward, FL) / justiça criminal	COMPAS: risco de reincidência	Análise estatística de registros de réus	Falsos positivos de 44,9% (negros) vs. 23,5% (brancos) entre não reincidentes	Dados e modelagem; institucional	Critério de equidade contestado pelo fabricante; um único condado
Dressel e Farid (2018)	EUA / justiça criminal	COMPAS vs. julgamento leigo	Experimento on-line e modelos simples	Leigos e modelo de duas variáveis com acurácia semelhante ao COMPAS	Modelagem	Participantes sem contexto decisório real
Kleinberg et al. (2018)	EUA (Nova York) / justiça criminal	Decisões de fiança	Aprendizado de máquina com dados administrativos	Potencial de reduzir prisões provisórias sem aumento de crimes e com menor disparidade racial	Modelagem; interação	Estimativas simuladas; dependem de desenho e auditoria

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

Estudo	País / campo	Aplicação de IA	Método	Principal resultado	Nível de viés	Limitação
Green e Chen (2019)	EUA / justiça criminal	Avaliação de risco com humano no circuito	Experimento com participantes on-line	Desvios da recomendação racialmente desiguais	Interação humano–IA	Participantes leigos
Mohler et al. (2015)	EUA (Los Angeles) e Reino Unido (Kent) / segurança pública	Previsão de pontos quentes	Ensaios de campo controlados	Previsões superiores às de analistas; associação com redução de crimes	Modelagem; implantação	Autores vinculados ao desenvolvimento da ferramenta
Brantingham, Valasik e Mohler (2018)	EUA (Los Angeles) / segurança pública	Policiamento preditivo	Ensaio randomizado	Sem diferença significativa nas taxas de prisão por grupo racial-étnico	Institucional; implantação	Mede prisões, não outras formas de contato; autores vinculados
Lum e Isaac (2016)	EUA (Oakland) / segurança pública	Policiamento preditivo (drogas)	Simulação com dados policiais e de saúde pública	Patrulhamento concentrado em bairros de maioria negra	Dados; retroalimentação	Simulação, não implantação real
Ensign et al. (2018)	Modelo teórico / segurança pública	Policiamento preditivo	Modelo formal	Laços de retroalimentação descontrolados	Retroalimen-taçã o	Resultado teórico
Richardson, Schultz e Crawford (2019)	EUA (13 jurisdições) / segurança pública	Sistemas preditivos	Estudos de caso jurídico-documentais	"Dados sujos" de práticas ilegais alimentaram sistemas	Institucional; dados	Não mede efeitos quantitativos
Brayne (2021)	EUA (LAPD) / segurança pública	Policiamento orientado por dados	Etnografia	Ampliação da vigilância e deslocamento da discricionariedade	Institucional; interação	Um único departamento
Buolamwini e Gebru (2018)	Sistemas comerciais / não policial	Classificação de gênero por imagem facial	Avaliação com base de referência balanceada	Erro de até 34,7% (mulheres de pele escura) vs. 0,8% (homens de pele clara)	Dados e modelagem	Não avalia identificação facial nem uso policial
Grother, Ngan e Hanaoka (2019)	EUA (NIST) / avaliação técnica	Reconhecimento facial 1:1 e 1:N (189 algoritmos)	Avaliação padronizada	Diferenciais demográficos na maioria; em 1:1, falsos positivos 10 a 100 vezes maiores para rostos asiáticos e afro-americanos em diversos algoritmos	Dados e modelagem	Condições controladas; grande variação entre algoritmos
Nunes (2019)	Brasil (BA, RJ, SC, PB) / segurança pública	Reconhecimento facial	Monitoramento de casos divulgados	151 prisões; 90,5% de pessoas negras	Institucional; implantação	Não mede taxa de erro; depende de divulgação
Obermeyer et al. (2019)	EUA / saúde (analogia)	Gestão de saúde populacional	Análise de dados	Variável substituta (custo) subestimou necessidade de pacientes negros	Modelagem (variável-alvo)	Fora da segurança pública

Fonte: elaboração própria, com base nos estudos indicados.

9.1 Eixo I – Fontes e Mecanismos de Viés

A literatura analisada converge em que os vieses em sistemas de IA aplicados à segurança pública não têm origem única. Sistemas de apoio à decisão costumam

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

ser adotados com a expectativa de maior consistência ou apoio analítico; contudo, nenhuma dessas tecnologias é neutra, pois todo sistema incorpora escolhas sobre o objetivo a otimizar, a definição da variável-alvo, os dados utilizados, o limiar decisório e a ponderação dos custos de cada tipo de erro. Quando algoritmos são treinados com registros que refletem práticas institucionais anteriores, marcadas por desigualdades raciais, sociais ou territoriais, podem reproduzir esses padrões sob aparência de objetividade (Barocas e Selbst, 2016; Richardson, Schultz e Crawford, 2019). A escolha de variáveis substitutas, como registros de prisão em lugar de criminalidade efetiva, é um dos mecanismos mais relevantes, pois mede a atividade policial tanto quanto o fenômeno que se pretende prever (Obermeyer et al., 2019, por analogia).

9.2 Eixo II – Policiamento Preditivo e Sistemas de Avaliação de Risco

As evidências sobre policiamento preditivo são heterogêneas. Mohler et al. (2015) relataram, em experimentos de campo controlados em Los Angeles e em Kent, que previsões algorítmicas de pontos quentes superaram as de analistas criminais e se associaram a reduções de crimes. Brantingham, Valasik e Mohler (2018), em ensaio randomizado, não encontraram diferença significativa nas taxas de prisão por grupo racial e étnico entre as condições de policiamento preditivo e de controle. Lum e Isaac (2016), por outro lado, demonstraram por simulação que dados históricos de crimes de drogas poderiam gerar concentração racialmente desigual do policiamento, e Ensign et al. (2018) formalizaram o mecanismo de retroalimentação subjacente. A etnografia de Brayne (2021) no Departamento de Polícia de Los Angeles mostra que o policiamento orientado por dados amplia a vigilância sobre pessoas sem envolvimento criminal e transfere, em vez de eliminar, a discricionariedade policial.

Esses resultados não são necessariamente contraditórios, pois analisam desenhos, dados e desfechos diferentes: ensaios que medem taxas de prisão em crimes patrimoniais não respondem à mesma pergunta que simulações sobre crimes de drogas, cujos registros dependem fortemente da iniciativa policial. A literatura não autoriza, portanto, concluir universalmente nem que o policiamento preditivo gera discriminação, nem que é isento de viés; o risco depende do tipo de crime previsto, da fonte dos dados, do desfecho medido e do contexto institucional de implantação.

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

No campo da justiça criminal, a controvérsia sobre o COMPAS é paradigmática. A análise da ProPublica sobre réus do condado de Broward, na Flórida, constatou que, entre os réus que não reincidiram em dois anos, 44,9% dos negros haviam sido classificados como de maior risco, contra 23,5% dos brancos — ou seja, uma taxa de falsos positivos quase duas vezes maior (Angwin et al., 2016; Larson et al., 2016). Separadamente, em análise ajustada por antecedentes, idade e gênero, a ProPublica relatou que réus negros eram 45% mais propensos a receber pontuação de risco mais elevada. A empresa responsável pelo COMPAS contestou a interpretação, sustentando que o instrumento apresentava acurácia e valor preditivo equivalentes entre os grupos (Dieterich, Mendoza e Brennan, 2016).

A controvérsia revelou limites técnicos mais profundos. Dressel e Farid (2018) mostraram que o sistema não era mais preciso que previsões feitas por pessoas leigas recrutadas pela internet, e que um modelo simples com apenas duas variáveis alcançava desempenho equivalente. Chouldechova (2017) e Kleinberg, Mullainathan e Raghavan (2017) demonstraram matematicamente que, quando as taxas-base de reincidência diferem entre grupos, é impossível satisfazer simultaneamente a calibração e a igualdade das taxas de erro — o que explica por que ProPublica e fabricante, adotando critérios distintos, chegaram a conclusões opostas. Corbett-Davies et al. (2017) evidenciaram que impor restrições de equidade pode ter custos em termos de segurança pública, e Berk et al. (2021) concluem que tais escolhas envolvem compromissos inevitáveis entre critérios de justiça e de precisão. Em sentido favorável à automação, Kleinberg et al. (2018) estimaram que algoritmos aplicados a decisões de fiança poderiam reduzir a população presa provisoriamente sem aumento da criminalidade, inclusive com redução de disparidades raciais, desde que cuidadosamente desenhados e auditados. Esses achados indicam que a definição do critério de equidade é uma decisão normativa e política, e não apenas técnica.

9.3 Eixo III – Reconhecimento Facial e Biometria

No reconhecimento facial, é necessário distinguir tarefas e contextos. O estudo *Gender Shades* avaliou três sistemas comerciais de classificação de gênero a partir de imagens faciais e encontrou erro de até 34,7% para mulheres de pele mais escura,

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

contra até 0,8% para homens de pele mais clara (Buolamwini e Gebru, 2018). Trata-se de evidência relevante sobre desempenho desigual por raça e gênero, mas não de avaliação de identificação facial policial ou de reconhecimento um-para-muitos. Para o reconhecimento facial propriamente dito, a referência mais apropriada é a avaliação do NIST, que analisou 189 algoritmos e encontrou diferenças demográficas na maioria deles, ressalvando grande variação entre sistemas: em verificações um-para-um, diversos algoritmos apresentaram taxas de falsos positivos para rostos asiáticos e afro-americanos de 10 a 100 vezes superiores às observadas para rostos caucasianos, embora os algoritmos de melhor desempenho apresentassem diferenciais muito menores (Grother, Ngan e Hanaoka, 2019). Garvie, Bedoya e Frankle (2016) estimaram que metade dos adultos norte-americanos já integrava redes de reconhecimento facial policial, em grande medida sem regulação.

No Brasil, o monitoramento da Rede de Observatórios da Segurança registrou, em 2019, 151 prisões associadas ao reconhecimento facial na Bahia, no Rio de Janeiro, em Santa Catarina e na Paraíba, das quais 90,5% de pessoas negras (Nunes, 2019). Esse dado não mede, por si só, taxas de erro dos algoritmos, pois reflete também onde as câmeras são instaladas, quais bases de procurados são utilizadas e como os alertas são tratados; justamente por isso, ilustra a cadeia descrita na seção 7.3, na qual o desempenho técnico se combina com decisões institucionais. A Autoridade Nacional de Proteção de Dados, em seu Radar Tecnológico sobre biometria e reconhecimento facial, apontou o risco de discriminação decorrente de vieses nas bases de treinamento e os impactos da vigilância em espaços públicos (ANPD, 2024). Silva (2022) situa esse uso no quadro mais amplo do racismo algorítmico, e Buolamwini (2023) descreve o padrão como "olhar codificado", no qual prioridades e omissões de quem desenvolve os sistemas se inscrevem na tecnologia. Permanece ausente, contudo, uma avaliação pública e independente da acurácia dos sistemas efetivamente utilizados pelas polícias brasileiras.

9.4 Eixo IV – Interação Humano–Algoritmo e Viés de Automação

Em segurança pública, não basta perguntar se o algoritmo possui viés: é necessário investigar como agentes humanos interpretam, aceitam ou contestam suas recomendações. O viés de automação (Skitka, Mosier e Burdick, 1999) pode levar um

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

policial a tratar um alerta de reconhecimento facial como identificação confirmada, ou um juiz a atribuir à pontuação de risco peso superior ao das demais provas. O efeito de enquadramento (Tversky e Kahneman, 1981) agrava esse risco quando o sistema apresenta categorias como "alto risco" em lugar de probabilidades com margem de incerteza. Inversamente, Green e Chen (2019) mostraram que a supervisão humana pode introduzir disparidades próprias, pois participantes se desviaram das recomendações de forma racialmente desigual. Brayne (2021) acrescenta que a discricionariedade policial não desaparece com os algoritmos, mas se desloca para as escolhas sobre quando consultá-los e como agir a partir deles.

A consequência é que a "revisão humana" só funciona como salvaguarda quando é qualificada, dispõe de tempo e informação suficientes, tem autoridade efetiva para divergir do sistema e é documentada. Uma revisão meramente formal tende a legitimar, e não a corrigir, os erros do sistema.

9.5 Eixo V – Impactos Jurídicos e Sociais

Decisões automatizadas ou apoiadas por algoritmos podem afetar princípios constitucionais como a igualdade, o devido processo legal e a presunção de inocência. A ausência de explicabilidade e transparência dificulta o exercício do contraditório e da ampla defesa, especialmente quando algoritmos são utilizados em decisões de risco criminal ou de monitoramento policial. A literatura jurídica fornece bases para enfrentar esse problema. Citron (2008) propõe o devido processo tecnológico, diante da constatação de que sistemas automatizados na administração pública podem suprimir as garantias de notificação e de audiência. Citron e Pasquale (2014) estendem essa proposta às pontuações preditivas, defendendo transparência, auditoria e possibilidade de contestação. Crawford e Schultz (2014) sugerem um devido processo de dados para os danos decorrentes de previsões, e Zarsky (2016) sistematiza as objeções de eficiência e de justiça às decisões automatizadas opacas. Završnik (2020) sintetiza os riscos ao julgamento justo, à paridade de armas e à presunção de inocência na justiça criminal automatizada.

No plano social, a análise dialoga com a noção de racismo estrutural de Almeida (2019), segundo a qual o racismo integra a organização econômica, política e jurídica da sociedade e não se reduz a atos individuais. Alexander (2010) descreve

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

o encarceramento em massa norte-americano como um sistema de controle racial, e Benjamin (2019) mostra que tecnologias apresentadas como neutras podem reproduzir essa lógica sob a forma de um novo código de discriminação. Eubanks (2018) documenta como ferramentas automatizadas de gestão pública perfilam, vigiam e punem desproporcionalmente as populações pobres, que se tornam o primeiro campo de testes dessas tecnologias, e Noble (2018) demonstra que mesmo sistemas cotidianos, como mecanismos de busca, reforçam estereótipos raciais e de gênero. Couldry e Mejias (2019) enquadram esse processo como colonialismo de dados, no qual populações vulneráveis se tornam objeto de extração e experimentação tecnológica sem garantias adequadas. É plausível, ainda, que a percepção de tratamento injusto reduza a confiança nas instituições de segurança e dificulte a cooperação entre população e Estado, hipótese que demanda investigação empírica específica no Brasil.

Há, por fim, uma dimensão operacional. Sistemas enviesados podem direcionar recursos policiais de forma ineficiente, concentrando o patrulhamento em determinadas regiões e negligenciando outras com necessidade real de intervenção. Combater vieses é, portanto, não apenas questão ética, mas também estratégica para a gestão pública.

9.6 Eixo VI – Estratégias de Mitigação e Governança

A mitigação não deve ser tratada apenas como problema técnico. Em segurança pública, algumas questões são anteriores ao algoritmo: determinado sistema deveria ser utilizado? A coleta dos dados é juridicamente legítima? O benefício esperado justifica a intrusão nos direitos afetados? Existe meio menos restritivo para alcançar o mesmo fim? Propõe-se, assim, uma sequência de salvaguardas em oito etapas, na qual as medidas técnicas ocupam posição intermediária:

necessidade → legalidade → proporcionalidade → desenho → validação → implantação → auditoria → contestação

Fonte: elaboração própria.

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

As três primeiras etapas são jurídico-políticas e podem conduzir à decisão de não adotar o sistema. As etapas de desenho e validação envolvem escolhas técnicas documentadas. Instrumentos como os cartões de modelo de Mitchell et al. (2019) e as fichas de dados de Gebru et al. (2021) permitem registrar a finalidade, os limites e o desempenho por subgrupo de cada sistema. Critérios formais de equidade, como a equidade individual de Dwork et al. (2012), segundo a qual indivíduos semelhantes devem receber tratamento semelhante, e a igualdade de oportunidades de Hardt, Price e Srebro (2016), oferecem parâmetros mensuráveis, discutidos de forma sistemática por Barocas, Hardt e Narayanan (2023). Selbst et al. (2019) advertem, contudo, para as armadilhas da abstração: soluções técnicas desconectadas do contexto social em que o sistema opera podem falhar mesmo quando satisfazem as métricas formais. Na implantação, são relevantes protocolos que definam o valor de um alerta, formação crítica dos agentes e revisão humana qualificada. A auditoria algorítmica independente e periódica (Raji et al., 2020), a revisão das bases de dados e comitês multidisciplinares com participação da sociedade civil completam o ciclo. A contestação, por fim, exige que a pessoa afetada saiba que um sistema foi utilizado e disponha de meios para questionar seu resultado.

Nesse ponto, o debate sobre transparência exige precisão. Rudin (2019) sustenta que, em decisões de alto impacto, devem ser adotados modelos intrinsecamente interpretáveis, em vez de modelos opacos acompanhados de explicações posteriores, que podem ser imprecisas, e Doshi-Velez e Kim (2017) propõem critérios rigorosos para avaliar a interpretabilidade. Quanto ao RGPD, convém distinguir três coisas: os deveres de informação e transparência do responsável pelo tratamento, inclusive sobre a existência de decisões automatizadas e a "lógica subjacente" (arts. 13 a 15); os direitos relativos a decisões exclusivamente automatizadas, como o de obter intervenção humana e contestar a decisão (art. 22); e uma explicação técnica integral do modelo, que o regulamento não assegura. Wachter, Mittelstadt e Floridi (2017) argumentam que o RGPD não institui propriamente um direito geral à explicação, mas um direito mais limitado à informação, e Wachter, Mittelstadt e Russell (2018) sugerem explicações contrafactuais, que indicam o que precisaria mudar para alterar o resultado, como forma viável de

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

contestação. Para as autoridades de aplicação da lei, aplica-se a Diretiva (UE) 2016/680, cujo art. 11 proíbe, em regra, decisões baseadas exclusivamente em tratamento automatizado que produzam efeitos adversos, salvo autorização legal com garantias adequadas, incluindo ao menos o direito de obter intervenção humana (União Europeia, 2016b).

O Regulamento (UE) 2024/1689 adota abordagem baseada em risco e deve ser descrito dentro dos limites exatos de suas proibições. Ele não proíbe genericamente toda avaliação de risco criminal por IA: proíbe sistemas destinados a avaliar ou prever o risco de uma pessoa cometer uma infração penal com base exclusivamente em perfilamento ou na avaliação de traços e características de personalidade (art. 5º, n.º 1, alínea *d*), ressalvados os sistemas que apoiam a avaliação humana baseada em fatos objetivos e verificáveis diretamente ligados a uma atividade criminosa. Outros sistemas empregados na aplicação da lei são classificados como de alto risco (Anexo III, ponto 6) e permanecem admissíveis sob requisitos específicos. A identificação biométrica à distância em tempo real em espaços acessíveis ao público para fins policiais tampouco é simplesmente proibida: sujeita-se a um regime de proibição com exceções estreitamente definidas — como a busca de vítimas de determinados crimes, a prevenção de ameaça iminente e a localização de suspeitos de infrações graves —, condicionadas à necessidade, à proporcionalidade e à autorização prévia de autoridade judiciária ou administrativa independente (art. 5º, n.º 1, alínea *h*, e n.ºs 2 a 7) (União Europeia, 2024). A Recomendação da UNESCO (2021), por sua vez, estabelece abordagem centrada em direitos humanos, proporcionalidade, supervisão humana e rejeição do uso da IA para pontuação social e vigilância em massa.

9.7 Eixo VII – Contexto Brasileiro e Lacunas Empíricas

No Brasil, o enquadramento jurídico exige precisão. A LGPD não se aplica ao tratamento de dados pessoais realizado para fins exclusivos de segurança pública, defesa nacional, segurança do Estado ou atividades de investigação e repressão de infrações penais (art. 4º, III); o § 1º do mesmo artigo determina que esse tratamento seja regido por legislação específica, que deverá prever medidas proporcionais e estritamente necessárias ao interesse público, observados o devido processo legal,

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

os princípios gerais de proteção e os direitos do titular previstos na LGPD. Por isso, não é correto afirmar genericamente que a LGPD exige explicabilidade algorítmica em segurança pública, nem que o direito de revisão de decisões automatizadas do art. 20 se aplica diretamente às decisões policiais automatizadas; o que se pode sustentar é que os princípios e direitos da LGPD devem orientar a legislação específica e a interpretação conforme a Constituição, sobretudo após o reconhecimento da proteção de dados como direito fundamental (Brasil, 2020a; 2022). O anteprojeto de lei elaborado em 2020 por comissão de juristas da Câmara dos Deputados para disciplinar esse tratamento ainda não foi convertido em lei (Brasil, 2020b).

No plano do marco regulatório da IA, o Projeto de Lei nº 2.338/2023 foi aprovado pelo Senado Federal, na forma de substitutivo, em 10 de dezembro de 2024, e encontra-se, na data desta revisão (setembro de 2026), em tramitação na Câmara dos Deputados, em Comissão Especial; não se trata, portanto, de legislação vigente (Brasil, 2023; Senado Federal, 2024). O texto aprovado pelo Senado classifica como de alto risco determinados sistemas de identificação biométrica e restringe o uso de identificação biométrica à distância em tempo real em espaços públicos a hipóteses excepcionais. O Projeto de Lei nº 21/2020, aprovado pela Câmara em 2021, representou etapa relevante do debate regulatório, mas foi declarado prejudicado com a aprovação do PL nº 2.338/2023 no Senado, devendo ser compreendido como antecedente histórico, e não como a principal proposição em andamento (Brasil, 2020c; Senado Federal, 2024). O Marco Civil da Internet (Lei nº 12.965/2014), por sua vez, não disciplina especificamente sistemas de IA.

Quanto às lacunas empíricas, o próprio mapeamento de Nobre e Leite (2025) aponta a sub-representação da produção científica de países africanos e de parte da América do Sul no debate sobre vieses algorítmicos, com predominância de autores dos Estados Unidos e da Europa. No Brasil, persistem a escassez de dados públicos e auditados sobre o desempenho dos sistemas de reconhecimento facial e de análise preditiva em uso, a ausência de relatórios de impacto obrigatórios e a carência de estudos sobre o comportamento de agentes diante de recomendações algorítmicas. Fatores como desigualdade regional, racismo estrutural e fragilidade institucional

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

tornam indispensável a análise crítica e contextualizada da aplicação de IA na segurança pública (Almeida, 2019; Silva, 2022; Machado e Silva, 2024).

Em síntese, os eixos analisados atendem aos objetivos da pesquisa ao demonstrar que o risco ético não decorre apenas de "preconceitos programados", mas de uma cadeia sociotécnica na qual dados históricos, decisões de modelagem, instituições, usuários humanos e mecanismos de retroalimentação podem produzir ou ampliar desigualdades.

10. Conclusão

A presente revisão narrativa crítica evidenciou que os riscos éticos associados à IA na segurança pública, na persecução penal e na justiça criminal não se explicam por um único tipo de viés. Vieses cognitivos humanos, vieses de dados e modelagem, vieses institucionais e sociotécnicos e vieses de interação humano–IA operam em níveis distintos, mas se reforçam ao longo de uma cadeia que vai da produção institucional dos registros à retroalimentação dos modelos. A distinção entre esses níveis é a principal contribuição conceitual do estudo, pois permite localizar onde intervir e evita tanto a atribuição do problema a "preconceitos programados" quanto a crença em uma tecnologia neutra.

As evidências empíricas examinadas são heterogêneas: indicam desempenho desigual de algoritmos de reconhecimento facial, com grande variação entre sistemas, e controvérsias metodológicas legítimas sobre avaliação de risco e policiamento preditivo, que não autorizam conclusões universais. No plano normativo, o estudo mostrou que a mitigação deve começar antes do algoritmo — pela verificação de necessidade, legalidade e proporcionalidade — e prosseguir com desenho documentado, validação, implantação sob protocolos claros, auditoria independente e mecanismos efetivos de contestação. No Brasil, a ausência da legislação específica prevista no art. 4º, § 1º, da LGPD e a tramitação ainda pendente do PL nº 2.338/2023 tornam essas salvaguardas especialmente urgentes.

Reconhece-se a limitação do trabalho por basear-se em revisão narrativa de fontes secundárias, indicando a necessidade de pesquisas empíricas que avaliem o desempenho e os efeitos dos sistemas efetivamente utilizados pelas forças de

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

segurança brasileiras, inclusive em contextos regionais específicos, como a realidade amazônica, e em estudos comparativos com outras regiões do mundo. Conclui-se que o desenvolvimento de uma inteligência artificial ética na segurança pública exige abordagem multidisciplinar, integrando tecnologia, direito e políticas públicas, de modo a assegurar inovação responsável e respeito aos direitos fundamentais.

Referências

Alexander, Michelle. ***The New Jim Crow: Mass Incarceration in the Age of Colorblindness***. The New Press, 2010.

Almeida, Silvio. ***Racismo estrutural***. Pólen, 2019.

Angwin, Julia, et al. **"Machine Bias."** *ProPublica*, 23 maio 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em 20 fev. 2026.

Autoridade Nacional de Proteção de Dados (ANPD). ***Radar Tecnológico, n. 2: Biometria e reconhecimento facial***. ANPD, jun. 2024.

Barocas, Solon, e Andrew D. Selbst. **"Big Data's Disparate Impact."** *California Law Review*, vol. 104, 2016, pp. 671-732.

Barocas, Solon, Moritz Hardt, e Arvind Narayanan. ***Fairness and Machine Learning: Limitations and Opportunities***. MIT Press, 2023.

Benjamin, Ruha. ***Race After Technology: Abolitionist Tools for the New Jim Code***. Polity, 2019.

Bentham, Jeremy. ***An Introduction to the Principles of Morals and Legislation***. T. Payne, 1789.

Berk, Richard, et al. **"Fairness in Criminal Justice Risk Assessments: The State of the Art."** *Sociological Methods & Research*, vol. 50, no. 1, 2021, pp. 3-44.

Binns, Reuben. **"Fairness in Machine Learning: Lessons from Political Philosophy."** *Proceedings of Machine Learning Research*, vol. 81, 2018, pp. 149-159.

Bostrom, Nick, e Eliezer Yudkowsky. **"The Ethics of Artificial Intelligence."** *The Cambridge Handbook of Artificial Intelligence*, organizado por Keith Frankish e William Ramsey, Cambridge UP, 2011.

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

Brantingham, P. Jeffrey, Matthew Valasik, e George O. Mohler. **"Does Predictive Policing Lead to Biased Arrests? Results from a Randomized Controlled Trial."** *Statistics and Public Policy*, vol. 5, no. 1, 2018, pp. 1-6.

Brasil. ***Constituição da República Federativa do Brasil de 1988***. Diário Oficial da União, 5 out. 1988.

Brasil. ***Lei nº 12.681, de 4 de julho de 2012: institui o Sistema Nacional de Informações de Segurança Pública, Prisionais e sobre Drogas (Sinesp)***. Diário Oficial da União, 2012.

Brasil. ***Lei nº 12.965, de 23 de abril de 2014: estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil (Marco Civil da Internet)***. Diário Oficial da União, 2014.

Brasil. ***Lei nº 13.709, de 14 de agosto de 2018: Lei Geral de Proteção de Dados Pessoais (LGPD)***. Diário Oficial da União, 2018.

Brasil. Supremo Tribunal Federal (Plenário). ***Referendo na Medida Cautelar na Ação Direta de Inconstitucionalidade 6.387/DF***. Relatora: Min. Rosa Weber, 2020a. Julgamento em 7 maio 2020.

Brasil. Câmara dos Deputados. Comissão de Juristas. ***Anteprojeto de Lei de Proteção de Dados para Segurança Pública e Investigação Criminal***. Câmara dos Deputados, nov. 2020b.

Brasil. Câmara dos Deputados. ***Projeto de Lei nº 21, de 2020: estabelece fundamentos, princípios e diretrizes para o desenvolvimento e a aplicação da inteligência artificial no Brasil***. Câmara dos Deputados, 2020c.

Brasil. ***Emenda Constitucional nº 115, de 10 de fevereiro de 2022: inclui a proteção de dados pessoais entre os direitos e garantias fundamentais***. Diário Oficial da União, 2022.

Brasil. Senado Federal. ***Projeto de Lei nº 2.338, de 2023: dispõe sobre o uso da inteligência artificial***. Senado Federal, 2023.

Brayne, Sarah. ***Predict and Surveil: Data, Discretion, and the Future of Policing***. Oxford UP, 2021.

Buolamwini, Joy. ***Unmasking AI: My Mission to Protect What Is Human in a World of Machines***. Random House, 2023.

Buolamwini, Joy, e Timnit Gebru. **"Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification."** *Proceedings of Machine Learning Research*, vol. 81, 2018, pp. 77-91.

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

Burrell, Jenna. **"How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms."** *Big Data & Society*, vol. 3, no. 1, 2016, pp. 1-12.

Caliskan, Aylin, Joanna J. Bryson, e Arvind Narayanan. **"Semantics Derived Automatically from Language Corpora Contain Human-like Biases."** *Science*, vol. 356, no. 6334, 2017, pp. 183-186.

Chouldechova, Alexandra. **"Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments."** *Big Data*, vol. 5, no. 2, 2017, pp. 153-163.

Citron, Danielle Keats. **"Technological Due Process."** *Washington University Law Review*, vol. 85, no. 6, 2008, pp. 1249-1313.

Citron, Danielle Keats, e Frank Pasquale. **"The Scored Society: Due Process for Automated Predictions."** *Washington Law Review*, vol. 89, no. 1, 2014, pp. 1-33.

Coeckelbergh, Mark. ***AI Ethics***. MIT Press, 2020.

Corbett-Davies, Sam, et al. **"Algorithmic Decision Making and the Cost of Fairness."** *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2017, pp. 797-806.

Couldry, Nick, e Ulises A. Mejias. ***The Costs of Connection: How Data Is Colonizing Human Life***. Stanford UP, 2019.

Crawford, Kate. ***Atlas of AI***. Yale UP, 2021.

Crawford, Kate, e Jason Schultz. **"Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms."** *Boston College Law Review*, vol. 55, no. 1, 2014, pp. 93-128.

Danaher, John. **"The Threat of Algocracy: Reality, Resistance and Accommodation."** *Philosophy & Technology*, vol. 29, no. 3, 2016, pp. 245-268.

Deleuze, Gilles. **"Postscript on the Societies of Control."** *October*, vol. 59, 1992, pp. 3-7.

Dieterich, William, Christina Mendoza, e Tim Brennan. ***COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity***. Northpointe Inc., 8 jul. 2016.

Doshi-Velez, Finale, e Been Kim. **"Towards a Rigorous Science of Interpretable Machine Learning."** *arXiv*, 2017, arXiv:1702.08608.

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

Dressel, Julia, e Hany Farid. **"The Accuracy, Fairness, and Limits of Predicting Recidivism."** *Science Advances*, vol. 4, no. 1, 2018, eaao5580.

Dwork, Cynthia, et al. **"Fairness through Awareness."** *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ACM, 2012, pp. 214-226.

Ensign, Danielle, et al. **"Runaway Feedback Loops in Predictive Policing."** *Proceedings of Machine Learning Research*, vol. 81, 2018, pp. 160-171.

Eubanks, Virginia. ***Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor***. St. Martin's Press, 2018.

Ferguson, Andrew Guthrie. ***The Rise of Big Data Policing***. NYU Press, 2017.

Floridi, Luciano, e Jeff W. Sanders. **"On the Morality of Artificial Agents."** *Minds and Machines*, vol. 14, 2004, pp. 349-379.

Floridi, Luciano, e Josh Cowls. **"A Unified Framework of Five Principles for AI in Society."** *Harvard Data Science Review*, vol. 1, no. 1, 2019.

Floridi, Luciano, et al. **"AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations."** *Minds and Machines*, vol. 28, no. 4, 2018, pp. 689-707.

Friedman, Batya, e Helen Nissenbaum. **"Bias in Computer Systems."** *ACM Transactions on Information Systems*, vol. 14, no. 3, 1996, pp. 330-347.

Garvie, Clare, Alvaro Bedoya, e Jonathan Frankle. ***The Perpetual Line-Up: Unregulated Police Face Recognition in America***. Georgetown Law Center on Privacy & Technology, 2016.

Gebru, Timnit, et al. **"Datasheets for Datasets."** *Communications of the ACM*, vol. 64, no. 12, 2021, pp. 86-92.

Green, Ben, e Yiling Chen. **"Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments."** *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2019, pp. 90-99.

Grother, Patrick, Mei Ngan, e Kayee Hanaoka. ***Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects***. NISTIR 8280, National Institute of Standards and Technology, dez. 2019, doi:10.6028/NIST.IR.8280.

Hagendorff, Thilo. **"The Ethics of AI Ethics: An Evaluation of Guidelines."** *Minds and Machines*, vol. 30, no. 1, 2020, pp. 99-120.

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

Harcourt, Bernard E. ***Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age***. U of Chicago P, 2007.

Hardt, Moritz, Eric Price, e Nathan Srebro. **"Equality of Opportunity in Supervised Learning."** *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 3315-3323.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. ***Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems***. 1. ed., IEEE, 2019.

Jobin, Anna, et al. **"The Global Landscape of AI Ethics Guidelines."** *Nature Machine Intelligence*, vol. 1, 2019, pp. 389-399.

Joh, Elizabeth E. **"Policing by Numbers: Big Data and the Fourth Amendment."** *Washington Law Review*, vol. 89, no. 1, 2014, pp. 35-68.

Kahneman, Daniel. ***Thinking, Fast and Slow***. Farrar, Straus and Giroux, 2011.

Kahneman, Daniel, Olivier Sibony, e Cass R. Sunstein. ***Noise: A Flaw in Human Judgment***. Little, Brown Spark, 2021.

Kant, Immanuel. ***Groundwork of the Metaphysics of Morals***. 1785. Cambridge UP, 1998.

Kleinberg, Jon, et al. **"Human Decisions and Machine Predictions."** *The Quarterly Journal of Economics*, vol. 133, no. 1, 2018, pp. 237-293.

Kleinberg, Jon, Sendhil Mullainathan, e Manish Raghavan. **"Inherent Trade-Offs in the Fair Determination of Risk Scores."** *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017, pp. 43:1-43:23.

Larson, Jeff, et al. **"How We Analyzed the COMPAS Recidivism Algorithm."** *ProPublica*, 23 maio 2016.

Lum, Kristian, e William Isaac. **"To Predict and Serve?"** *Significance*, vol. 13, no. 5, 2016, pp. 14-19.

Machado, Helena, e Susana Silva. ***Desafios sociais e éticos da inteligência artificial no século XXI***. UMinho Editora, 2024, doi:10.21814/uminho.ed.130.

Mehrabi, Ninareh, et al. **"A Survey on Bias and Fairness in Machine Learning."** *ACM Computing Surveys*, vol. 54, no. 6, 2021, pp. 1-35.

Mendes, Laura Schertel, e Marcela Mattiuzzo. **"Discriminação algorítmica: conceito, fundamento legal e tipologia."** *Revista Direito Público*, vol. 16, no. 90, 2019, pp. 39-64.

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

Mill, John Stuart. ***Utilitarianism***. Parker, Son and Bourn, 1863.

Mitchell, Margaret, et al. **"Model Cards for Model Reporting."** *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2019, pp. 220-229.

Mittelstadt, Brent. **"Principles Alone Cannot Guarantee Ethical AI."** *Nature Machine Intelligence*, vol. 1, no. 11, 2019, pp. 501-507.

Mittelstadt, Brent Daniel, et al. **"The Ethics of Algorithms: Mapping the Debate."** *Big Data & Society*, vol. 3, no. 2, 2016, pp. 1-21.

Mohler, George O., et al. **"Randomized Controlled Field Trials of Predictive Policing."** *Journal of the American Statistical Association*, vol. 110, no. 512, 2015, pp. 1399-1411.

Noble, Safiya Umoja. ***Algorithms of Oppression: How Search Engines Reinforce Racism***. NYU Press, 2018.

Nobre, Amanda Jeniffer Pereira, e Ana Carolina Rodrigues Leite. ***Desafios éticos associados aos vieses em algoritmos de inteligência artificial: um mapeamento sistemático da literatura***. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Software), Faculdade de Ciências e Tecnologias em Engenharia, Universidade de Brasília, 2025, bdm.unb.br/handle/10483/42354. Acesso em 27 set. 2026.

Nunes, Pablo. **"Novas ferramentas, velhas práticas: reconhecimento facial e policiamento no Brasil."** *Retratos da violência: cinco meses de monitoramento, análises e descobertas*, Rede de Observatórios da Segurança / CESeC, nov. 2019.

Obermeyer, Ziad, et al. **"Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations."** *Science*, vol. 366, no. 6464, 2019, pp. 447-453.

Office of the United Nations High Commissioner for Human Rights (OHCHR). ***The Right to Privacy in the Digital Age: Report of the United Nations High Commissioner for Human Rights***. A/HRC/48/31, Nações Unidas, 13 set. 2021.

Pasquale, Frank. ***The Black Box Society: The Secret Algorithms That Control Money and Information***. Harvard UP, 2015.

Raji, Inioluwa Deborah, et al. **"Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing."** *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2020, pp. 33-44.

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

Rawls, John. ***A Theory of Justice***. Harvard UP, 1971.

Rechtbank Den Haag (Tribunal Distrital de Haia). ***NJCM c.s. v. De Staat der Nederlanden (SyRI)***. ECLI:NL:RBDHA:2020:865, 5 fev. 2020.

Requião, Maurício. **"Inteligência artificial, vieses cognitivos e decisões judiciais."** *Revista do Programa de Pós-Graduação em Direito da UFBA*, vol. 34, 2024, doi:10.9771/rppgd.v34i0.63797.

Richardson, Rashida, Jason M. Schultz, e Kate Crawford. **"Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice."** *New York University Law Review Online*, vol. 94, 2019, pp. 15-55.

Rudin, Cynthia. **"Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead."** *Nature Machine Intelligence*, vol. 1, no. 5, 2019, pp. 206-215.

Russell, Stuart. ***Human Compatible: Artificial Intelligence and the Problem of Control***. Viking, 2019.

Russell, Stuart, e Peter Norvig. ***Artificial Intelligence: A Modern Approach***. 4. ed., Pearson, 2020.

Selbst, Andrew D., et al. **"Fairness and Abstraction in Sociotechnical Systems."** *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2019, pp. 59-68.

Senado Federal. **"Senado aprova regulamentação da inteligência artificial; texto vai à Câmara."** *Agência Senado*, 10 dez. 2024, www12.senado.leg.br/noticias/materias/2024/12/10/senado-aprova-regulamentacao-da-inteligencia-artificial-texto-vai-a-camara. Acesso em 27 set. 2026.

Silva, Tarcízio. ***Racismo algorítmico: inteligência artificial e discriminação nas redes digitais***. Edições Sesc São Paulo, 2022.

Skitka, Linda J., Kathleen L. Mosier, e Mark Burdick. **"Does Automation Bias Decision-Making?"** *International Journal of Human-Computer Studies*, vol. 51, no. 5, 1999, pp. 991-1006.

Suresh, Harini, e John Guttag. **"A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle."** *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*, ACM, 2021.

Turing, Alan M. **"Computing Machinery and Intelligence."** *Mind*, vol. 59, 1950, pp. 433-460.

CC BY

REMUNOM
ISSN 2178-6925

Received: 13/08/2026 - **Accepted:** 24/09/2026
Vol: 13.20
DOI: 10.66104/aczpy604
Pages: 1-32

Tversky, Amos, e Daniel Kahneman. **"Judgment under Uncertainty: Heuristics and Biases."** *Science*, vol. 185, no. 4157, 1974, pp. 1124-1131.

Tversky, Amos, e Daniel Kahneman. **"The Framing of Decisions and the Psychology of Choice."** *Science*, vol. 211, no. 4481, 1981, pp. 453-458.

UNESCO. ***Recommendation on the Ethics of Artificial Intelligence***. UNESCO, 23 nov. 2021.

União Europeia. ***Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016 (Regulamento Geral sobre a Proteção de Dados)***. Jornal Oficial da União Europeia, L 119, 2016a.

União Europeia. ***Diretiva (UE) 2016/680 do Parlamento Europeu e do Conselho, de 27 de abril de 2016, relativa à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais pelas autoridades competentes para efeitos de prevenção, investigação, deteção ou repressão de infrações penais***. Jornal Oficial da União Europeia, L 119, 2016b.

União Europeia. ***Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024 (Regulamento da Inteligência Artificial)***. Jornal Oficial da União Europeia, L, 12 jul. 2024.

Vallor, Shannon. ***Technology and the Virtues***. Oxford UP, 2016.

Wachter, Sandra, Brent Mittelstadt, e Chris Russell. **"Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR."** *Harvard Journal of Law & Technology*, vol. 31, no. 2, 2018, pp. 841-887.

Wachter, Sandra, Brent Mittelstadt, e Luciano Floridi. **"Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation."** *International Data Privacy Law*, vol. 7, no. 2, 2017, pp. 76-99.

Zarsky, Tal. **"The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making."** *Science, Technology, & Human Values*, vol. 41, no. 1, 2016, pp. 118-132.

Završnik, Aleš. **"Criminal Justice, Artificial Intelligence Systems, and Human Rights."** *ERA Forum*, vol. 20, no. 4, 2020, pp. 567-583.

Zuboff, Shoshana. ***The Age of Surveillance Capitalism***. PublicAffairs, 2019.

Zuiderveen Borgesius, Frederik. ***Discrimination, Artificial Intelligence, and Algorithmic Decision-Making***. Council of Europe, 2018.

CC BY