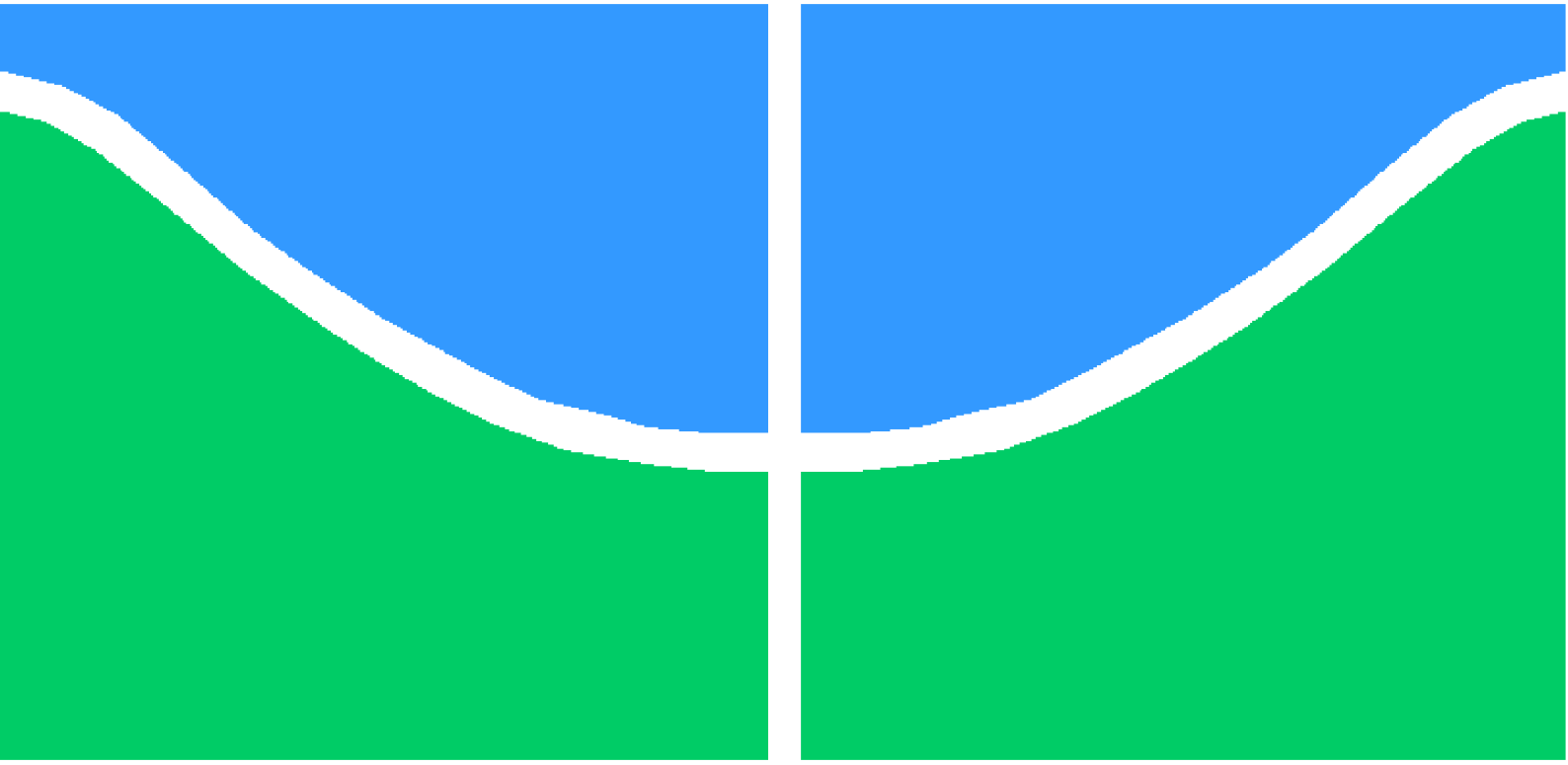


Universidade de Brasília - UnB

Faculdade de Ciências e Tecnologias em Engenharia - FCTE

Engenharia de Software

Desafios Éticos Associados aos Vieses em Algoritmos de Inteligência Artificial: Um Mapeamento Sistemático da Literatura

Autora: Amanda Jeniffer Pereira Nobre
e Ana Carolina Rodrigues Leite

Orientador: Dr. Andre Luiz Peron Martins Lanna

Brasília, DF

2025

Amanda Jeniffer Pereira Nobre
e Ana Carolina Rodrigues Leite

Desafios Éticos Associados aos Vieses em Algoritmos de Inteligência Artificial: Um Mapeamento Sistemático da Literatura

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Universidade de Brasília - UnB

Faculdade de Ciências e Tecnologias em Engenharia - FCTE

Orientador: Dr. Andre Luiz Peron Martins Lanna

Brasília, DF
2025

Amanda Jeniffer Pereira Nobre
e Ana Carolina Rodrigues Leite
Desafios Éticos Associados aos Vieses em Algoritmos de Inteligência Artificial: Um Mapeamento Sistemático da Literatura/ Amanda Jeniffer Pereira Nobre
e Ana Carolina Rodrigues Leite. – Brasília, DF, 2025
67 p. : il. (algumas color.) ; 30 cm.

Orientador: Dr. Andre Luiz Peron Martins Lanna

Trabalho de Conclusão de Curso – Universidade de Brasília - UnB
Faculdade de Ciências e Tecnologias em Engenharia - FCTE , 2025.

1. Ética 2. Inteligência Artificial. 3. Educação Superior. III. Universidade de Brasília. IV. Faculdade de Ciências e Tecnologias em Engenharia - FCTE. V. Desafios Éticos Associados aos Vieses em Algoritmos de Inteligência Artificial: Um Mapeamento Sistemático da Literatura

CDU -

Amanda Jeniffer Pereira Nobre
e Ana Carolina Rodrigues Leite

Desafios Éticos Associados aos Vieses em Algoritmos de Inteligência Artificial: Um Mapeamento Sistemático da Literatura

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Trabalho aprovado. Brasília, DF, 21 de julho de 2025: Brasília, DF, 21 de julho de 2025:

Dr. Andre Luiz Peron Martins Lanna
Orientador

Dr. Daniel Sundfeld Lima
Convidado 1

Dr. Zanei Ramos Barcellos
Convidado 2

Brasília, DF
2025

Ana Carolina Rodrigues Leite

Dedico este trabalho à minha família, que sempre me incentivou a nunca desistir dos meus sonhos. Seu apoio constante e encorajamento foram fundamentais para que eu chegasse até aqui. E a Deus, por sempre permitir os meus sonhos e por me guiar com sua graça e sabedoria ao longo desta jornada.

Amanda Jeniffer Pereira Nobre

Dedico este trabalho aos meus pais, que sob muito sol, me permitiram chegar até aqui, na sombra.

Agradecimentos

Ana Carolina Rodrigues Leite

A Deus, por ter me dado força, sabedoria e coragem ao longo de toda essa jornada acadêmica. Sua presença constante me guiou e sustentou nos momentos de dificuldade e incerteza, me proporcionando fé e inspiração para seguir em frente.

Aos meus pais, JeneJames e Marilene, pelo amor incondicional, apoio inabalável e pelos inúmeros sacrifícios feitos ao longo dos anos. Sem essa base sólida que vocês me proporcionaram, esse sonho não seria possível. Agradeço por acreditarem em mim e serem meus maiores incentivadores.

Ao meu namorado, Augusto, que sempre me deu o suporte e apoio recorrente em momentos que tive dificuldade em pedir ajuda e ele estava sempre firme, me incentivando a não desistir.

À Universidade de Brasília (UnB), por ser o palco do meu desenvolvimento acadêmico e pessoal. Agradeço aos professores, colegas e funcionários que contribuíram para minha formação e crescimento. Esta instituição não foi apenas um local de aprendizado, mas também um espaço de grandes experiências e descobertas.

Agradeço ao orientador por guiar nosso caminho.

Amanda Jeniffer Pereira Nobre

Aos meus pais, Jorge e Aleteia, cuja força e suporte foram pilares nos momentos mais desafiadores. Minha mãe, verdadeira heroína, sempre presente com seu apoio incondicional e incentivo, especialmente nas horas de desânimo e cansaço. Meu pai, que apesar de todas as minhas dificuldades, sempre me fortaleceu e foi uma figura de imenso valor em minha jornada.

Aos meus irmãos, por seu constante incentivo e compreensão durante minha ausência enquanto me dedicava a este trabalho.

Aos meus queridos amigos, pela amizade inabalável e pelo apoio contínuo que me deram ao longo de todo o período de dedicação a este projeto. Sua presença constante foi um alicerce fundamental.

Ao professor André Lanna, pela orientação e dedicação. Sua paciência e sabedoria foram essenciais para a conclusão deste trabalho.

"A oportunidade dança com aqueles que já estão no salão." (H. Jackson Brown Jr.)

Resumo

O avanço da inteligência artificial (IA) tem proporcionado transformações significativas em diversas áreas, como: saúde, segurança e mercado de trabalho. No entanto, a presença de vieses algorítmicos nesses sistemas tem levantado preocupações éticas, uma vez que podem amplificar desigualdades sociais e reforçar discriminações históricas. Este trabalho tem como objetivo realizar um mapeamento sistemático da literatura para identificar os desafios éticos associados aos vieses em algoritmos de IA. A pesquisa adota uma abordagem quantitativa, analisando a produção científica sobre o tema, mapeando indicadores de produção, como a evolução temporal, a distribuição geográfica e os tipos de publicação, bem como a estrutura intelectual do campo, por meio de seus autores mais atuantes, focos temáticos e trabalhos de maior impacto. Os resultados do mapeamento sistemático fornecem *insights* valiosos que respondem diretamente aos objetivos propostos, detalhando a estrutura intelectual, a evolução e os focos temáticos da pesquisa sobre os desafios éticos em IA.

Palavras-chave: Ética. Inteligência Artificial. Viés. Mapeamento Sistemático.

Abstract

The advancement of artificial intelligence (AI) has brought about significant transformations in various fields, such as healthcare, security, and the job market. However, the presence of algorithmic biases in these systems has raised ethical concerns, as they can amplify social inequalities and reinforce historical discrimination. This study aims to conduct a systematic mapping of the literature to identify the ethical challenges associated with biases in AI algorithms. The research adopts a quantitative approach, analyzing the scientific production on the topic, mapping production indicators such as temporal evolution, geographical distribution, and publication types, as well as the intellectual structure of the field through its most active authors, thematic focuses, and most impactful works. The results of the systematic mapping provide valuable insights that directly address the proposed objectives by detailing the intellectual structure, evolution, and thematic focuses of the research on ethical challenges in AI.

Key-words: Ethics. Artificial Intelligence. Bias. Systematic Mapping.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

IA	Inteligência Artificial
MSL	Mapeamento Sistemático da Literatura
MIT	Instituto de Tecnologia de Massachusetts
XAI	Inteligência Artificial Explicável
TCC	Trabalho de Conclusão de Curso
UNB	Universidade de Brasília

Sumário

1 Introdução

1.1 Contextualização

A inteligência artificial (IA) oferece diversas oportunidades para melhorar a vida cotidiana, com avanços que vão desde diagnósticos médicos e previsões climáticas mais precisas até a automação de veículos (RUSSELL; NORVIG, 2020). Iniciativas como o *AI for Humanitarian Action da Microsoft* e o *AI for Good*, em parceria com o Google, exemplificam o potencial da IA para enfrentar desafios globais em áreas como direitos humanos, desastres naturais e preservação ambiental (COLUMBIA, 2024; AWARDS, 2024).

No entanto, essas inovações também trazem desafios significativos, segundo Russell e Norvig (2020) o impacto econômico da automação pode agravar as desigualdades sociais, com a concentração de riqueza nas mãos de poucos, e ameaçar o desenvolvimento de países emergentes. Esses desafios evidenciam a necessidade de um debate ético em torno do uso da IA, especialmente no que diz respeito à transparência e confiança, visto que a opacidade dos algoritmos pode dificultar a compreensão de suas decisões (BURRELL, 2016). As ferramentas de IA podem, inclusive, perpetuar ou amplificar preconceitos existentes caso aconteça de estarem, de alguma forma, enviesadas. (NOBLE, 2018).

Os vieses em inteligência artificial se manifestam em várias dimensões, com impactos que se estendem desde os dados de origem até as interações humano-máquina e os resultados gerados pelos sistemas. Destacam-se o viés de interação do usuário, que reflete ciclos de retroalimentação entre o design do sistema e o comportamento do usuário; o viés comportamental, no qual padrões cognitivos humanos, como preconceitos inconscientes ou aversões, são modelados e amplificados pelos sistemas; e o viés de agregação, que emerge quando dados heterogêneos são agrupados, resultando na perda de granularidade e na incapacidade de capturar variações individuais importantes. A falta de tratamento adequado a esses vieses resulta na perpetuação de preconceitos e no reforço de padrões discriminatórios. (MEHRABI et al., 2021).

1.2 Justificativa

A relevância do tema dos vieses em inteligência artificial está diretamente ligada às questões éticas, sociais e tecnológicas que impactam a sociedade. Com o uso crescente de IA em setores essenciais como saúde, segurança pública, finanças e recursos humanos, torna-se importante compreender e mitigar os vieses que emergem desses algoritmos já que, conforme Obermeyer et al. (2019) e Mehrabi et al. (2021), esses sistemas podem

reproduzir e amplificar preconceitos históricos e desigualdades sociais, colocando em risco os valores de equidade e justiça.

Uma lacuna importante no campo é a falta de metodologias para identificar e corrigir vieses, especialmente em contextos sensíveis e complexos. Estudos como os de Selbst et al. (2019) abordam a opacidade algorítmica e a necessidade de avaliações mais abstratas de justiça, enquanto Barocas, Hardt e Narayanan (2019) ressaltam a importância de perspectivas interdisciplinares para superar esses desafios. Além disso, Buolamwini e Gebru (2018) destacam o impacto desproporcional de vieses sobre grupos sub-representados, como mulheres e minorias étnicas, reforçando a urgência de práticas mais inclusivas no desenvolvimento de IA.

Este estudo tem como objetivo mapear e sintetizar informações sobre as pesquisas existentes relacionadas aos desafios éticos dos vieses algorítmicos na IA. Binns (2018) e Raji et al. (2020) se destacam a relevância de práticas como auditorias algorítmicas e maior transparência nos processos decisórios, abordagens que servem de inspiração para esta pesquisa. Além disso, o estudo busca organizar dados e tendências de publicações, contribuindo para debates mais informados e para o desenvolvimento de tecnologias mais inclusivas e equitativas.

A pesquisa também se propõe a destacar a relevância do tema para países em desenvolvimento, que frequentemente enfrentam desafios únicos e são negligenciados em estudos globais. Trabalhos como os de Suresh e Guttag (2021) mostram que sistemas de IA podem interagir com desigualdades sociais preexistentes, exigindo abordagens adaptadas a esses contextos. Assim, este estudo busca ampliar o alcance dessas discussões e propor soluções que considerem a diversidade das realidades sociais.

Por fim, a relevância desse tema transcende o campo acadêmico, impactando diretamente o desenvolvimento social e as interações humanas. Investigar vieses em IA é essencial para garantir que essas tecnologias sejam ferramentas de progresso e equidade, em vez de perpetuadoras de desigualdades. Trabalhos como os de Cath et al. (2018) reforçam a importância de incorporar valores éticos ao desenvolvimento tecnológico, papel em que a pesquisa acadêmica se torna fundamental.

1.3 Questão de Pesquisa

Para guiar este trabalho, foi definida a seguinte questão de pesquisa:

Quais os principais desafios éticos associados aos vieses de agregação, de interação com o usuário e comportamental em algoritmos de IA atualmente?

Essa questão busca identificar e organizar as discussões existentes sobre os desafios éticos relacionados aos vieses em sistemas de IA, mantendo o foco do estudo na realização

de uma análise quantitativa das publicações, que permita mapear padrões e tendências na produção científica sobre o tema.

Ao mapear a literatura de forma estruturada, o trabalho contribuiu para uma melhor compreensão do panorama atual, oferecendo *insights* sobre como a comunidade científica tem abordado os desafios éticos dos vieses em IA. Dessa forma, o estudo serve como base para orientar pesquisas futuras e para a formulação de diretrizes que promovam o desenvolvimento de sistemas de IA mais justos, transparentes e inclusivos.

1.4 Objetivos

1.4.1 Objetivo geral

Mapear os desafios éticos dos vieses em algoritmos de IA, identificando os desafios éticos discutidos e caracterizando o cenário científico por meio de suas principais publicações, autores, evolução temporal, distribuição geográfica, focos temáticos, tipos de publicação e trabalhos de maior impacto.

1.4.2 Objetivos específicos

1. Identificar os autores e grupos de pesquisa mais produtivos que atuam na área de vieses algorítmicos e ética em IA;

2. Analisar a evolução temporal da produção científica para identificar os períodos de maior volume de publicações e o crescimento do interesse acadêmico sobre vieses em IA;

3. Mapear os principais países e regiões responsáveis pelas publicações sobre os desafios éticos dos vieses em IA, mapeando a distribuição geográfica da produção científica;

4. Identificar os focos temáticos e as frentes de pesquisa por meio da análise de frequência e coocorrência de palavras-chave, investigando sua evolução para refletir mudanças de enfoque no debate sobre vieses;

5. Classificar os tipos de estudos predominantes (artigos de conferência, periódicos, dissertações, etc.) para determinar os principais canais de disseminação do conhecimento sobre vieses em IA;

6. Levantar os trabalhos fundamentais e de maior impacto por meio da análise de citações, a fim de compreender as contribuições teóricas e conceituais que moldaram a pesquisa sobre vieses em IA.

1.5 Organização do trabalho

Este trabalho de conclusão de curso está organizado nos seguintes capítulos:

- **Capítulo 1 - Introdução**: este capítulo oferece uma contextualização sobre o tema, incluindo a justificativa da pesquisa e os objetivos tanto principal quanto os específicos para esse trabalho.

- **Capítulo 2 - Referencial Teórico**: apresenta conceitos e estudos que apoiaram no desenvolvimento desta pesquisa.

- **Capítulo 3 - Mapeamento Sistemático da Literatura**: apresenta o delineamento do mapeamento sistemático da literatura realizado, detalhando a metodologia adotada na pesquisa, incluindo a construção do protocolo e sua execução.

- **Capítulo 4 - Resultados**: fornece uma apresentação detalhada dos resultados obtidos através do mapeamento sistemático da literatura, permitindo a análise e interpretação dos dados coletados.

- **Capítulo 5 - Considerações finais**: expõe as conclusões derivadas desta pesquisa, discute suas limitações e sugere possíveis direções para trabalhos futuros.

2 Referencial Teórico

O presente capítulo estabelece o referencial teórico que serve como alicerce para este trabalho, tendo como objetivo definir e contextualizar os conceitos que formam a base para a análise dos vieses algorítmicos em inteligência artificial.

2.1 Fundações Éticas para a Análise Tecnológica

A ética, como campo de estudo filosófico, busca compreender o que constitui o comportamento moralmente correto e avaliar as justificativas para essas normas (JOHNSON, 2008). No contexto do uso de tecnologias, especialmente da IA e outras inovações avançadas, torna-se essencial explorar como diferentes teorias éticas podem orientar decisões e práticas. Entre as principais teorias éticas estão o utilitarismo, a deontologia e a ética das virtudes, cada uma oferecendo perspectivas distintas para avaliar e guiar a ação moral.

O utilitarismo, inicialmente proposto por Jeremy Bentham e posteriormente desenvolvido por John Stuart Mill, fundamenta-se no princípio da maximização do bem-estar (BOSTROM; YUDKOWSKY, 2011). De acordo com essa teoria, uma ação é moralmente correta se produzir o maior benefício possível para o maior número de pessoas. No contexto tecnológico, o utilitarismo pode ser aplicado na avaliação das consequências do uso de IA. Por exemplo, ao implementar sistemas de IA na educação, como assistentes virtuais para personalização da aprendizagem, a avaliação utilitarista consideraria se essas tecnologias ampliam significativamente o acesso à educação, aumentam a eficiência do ensino ou promovem a igualdade de oportunidades educacionais.

Em contraste, a deontologia, associada principalmente ao filósofo Immanuel Kant, enfatiza a moralidade das ações em si mesmas, independentemente de suas consequências (KANT; QUINTELA, 1997). Para a deontologia, certas ações são intrinsecamente corretas ou erradas e devem ser seguidas, mesmo que sua violação possa trazer benefícios a curto prazo. No uso de tecnologias, isso implica que princípios éticos como o respeito à privacidade, a justiça na distribuição de recursos tecnológicos e a transparência nos processos decisórios são imperativos morais que não devem ser transgredidos (FLORIDI; SANDERS, 2004).

Já a ética das virtudes, abordagem desenvolvida por filósofos como Aristóteles, concentra-se no desenvolvimento de características de caráter moralmente excelentes (FLORES, 2021). Essa teoria sugere que o foco principal deve estar na formação de virtudes como honestidade, coragem, compaixão e sabedoria, tanto nos indivíduos quanto

nas organizações. No contexto tecnológico, a ética das virtudes pode ser aplicada ao desenvolvimento e uso de IA que promovam tais valores. Por exemplo, sistemas de IA que incentivam a colaboração, a integridade acadêmica e a equidade no acesso à informação podem ser considerados moralmente desejáveis sob essa perspectiva.

A aplicação desses ideais éticos ao uso de tecnologias, incluindo a IA, não apenas orienta a tomada de decisões éticas, mas também possibilita a criação de políticas e diretrizes que promovam um uso responsável e benéfico dessas ferramentas (BOSTROM; YUDKOWSKY, 2011). Por exemplo, no desenvolvimento de algoritmos de IA voltados para aplicações educacionais, a consideração das consequências utilitárias pode ajudar a prever e mitigar impactos negativos, como a amplificação de desigualdades educacionais. Simultaneamente, princípios deontológicos assegurariam o respeito à privacidade dos alunos e garantiriam que os processos automatizados sejam transparentes e responsáveis. Além disso, a promoção de virtudes morais por meio da tecnologia pode fortalecer a integridade do sistema educacional, cultivando um ambiente de aprendizagem ético e inclusivo (FLORIDI; SANDERS, 2004).

Para implementar efetivamente essas teorias éticas, é importante considerar os contextos culturais, sociais e legais específicos, adaptando princípios éticos universais as realidades práticas do uso de tecnologias em diferentes setores (JOHNSON, 2008). Estudos como *"The Ethics of Artificial Intelligence"* (BOSTROM; YUDKOWSKY, 2011) e *"On the Morality of Artificial Agents"* (FLORIDI; SANDERS, 2004) fornecem *insights* valiosos sobre como as teorias éticas tradicionais podem ser aplicadas e adaptadas ao contexto da IA e de outras tecnologias emergentes.

2.2 Inteligência Artificial: Conceitos e Contextos

A IA é um campo multidisciplinar que busca desenvolver sistemas capazes de executar tarefas que tradicionalmente requereriam inteligência humana. De acordo com Russell e Norvig (2020), a IA pode ser abordada sob diferentes perspectivas, como a reprodução do comportamento humano ou a busca pela racionalidade, ou seja, a capacidade de tomar decisões boas com base em critérios lógicos e matemáticos.

Um marco nesse campo foi o teste de Turing, proposto por Alan Turing em 1950, que mede a inteligência de uma máquina pela sua capacidade de se passar por humano em uma conversa escrita. Para superar esse teste, seria necessário dominar habilidades como processamento de linguagem natural, raciocínio automatizado e aprendizado de máquina (RUSSELL; NORVIG, 2020).

Contudo, muitos pesquisadores concentram-se em princípios fundamentais da inteligência em vez de replicar o comportamento humano. "A busca pelo "voo artificial" teve sucesso quando engenheiros e inventores pararam de imitar pássaros e começaram a

usar túneis de vento e aprender sobre aerodinâmica" (RUSSELL; NORVIG, 2020, p. 20). Essa evolução demonstra como o avanço da engenharia aeronáutica ocorreu ao estudar aerodinâmica em vez de tentar copiar exatamente o voo das aves. A integração de áreas como psicologia, matemática, estatística e economia tornou a IA um dos campos mais dinâmicos da ciência contemporânea.

Uma dessas áreas é a filosofia, que historicamente abordou questões fundamentais sobre a mente, o conhecimento e a racionalidade. Aristóteles desenvolveu um sistema lógico que permitia gerar conclusões a partir de premissas, influenciando pensadores como Ramon Llull e Leibniz, que imaginaram máquinas baseadas em raciocínio mecânico (RUSSELL; NORVIG, 2020). Há séculos, filósofos investigam a mente humana e a possibilidade de mentes não humanas, questões centrais no estudo da inteligência. Alguns acreditam que máquinas podem realizar todas as tarefas humanas, enquanto outros argumentam que certos comportamentos, como criatividade e escolhas morais, estão além das capacidades das máquinas (NEGNEVITSKY, 2005).

Voltando para a atualidade, o campo da IA vive uma fase de avanços acelerados, impulsionada principalmente pela ascensão dos modelos de linguagem de grande escala (LLMs) e de IA generativa. Tecnologias que antes pareciam distantes, como a geração de texto coerente e imagens, composições musicais e códigos de programação funcionais, tornaram-se amplamente acessíveis através de ferramentas como o ChatGPT da OpenAI, Gemini do Google, entre outros. Essa nova onda de IA demonstrou uma capacidade surpreendente de aprendizado e adaptação, impactando diretamente desde a automação de tarefas cotidianas até a forma como empresas desenvolvem produtos e tomam decisões.

Tais progressos têm catalisado uma integração cada vez mais profunda da IA em diversos setores. Na ciência, por exemplo, modelos como o AlphaFold 3, desenvolvido pelo Google DeepMind, revolucionaram a biologia ao prever com alta precisão a estrutura tridimensional de proteínas e outras moléculas biológicas, acelerando a pesquisa de novos medicamentos e o entendimento de doenças (ABRAMSON et al., 2024). De acordo com o relatório *AI Index* de 2025 da Universidade de Stanford, o investimento privado em IA continua a crescer massivamente e a adoção da tecnologia por organizações aumentou significativamente, com muitas relatando melhorias mensuráveis na produtividade (MASLEJ et al., 2024). Esse cenário indica que a IA deixou de ser um campo puramente experimental para se tornar uma força transformadora e presente na sociedade contemporânea.

2.3 Os Desafios Éticos da Inteligência Artificial

A ascensão da IA trouxe consigo avanços tecnológicos significativos em diversas áreas, como: saúde, educação e transporte. No entanto, o uso extensivo dessa tecnologia tem levantado preocupações éticas, especialmente em relação aos vieses algorítmicos e

as implicações sociais de sua implementação (JOBIN; IENCA; VAYENA, 2019). Esses desafios decorrem, em grande parte, da forma como os sistemas de IA são desenvolvidos, treinados e aplicados em contextos que podem perpetuar ou amplificar desigualdades existentes.

Entretanto, os desafios éticos associados à IA vão além do viés algorítmico, englobando também questões de transparência, responsabilidade e privacidade. A falta de explicabilidade dos sistemas – conhecida como "caixa-preta" – dificulta a compreensão dos critérios adotados nas decisões, o que gera desconfiança entre os usuários e complica a responsabilização dos desenvolvedores (FLORIDI et al., 2018). Ademais, o uso intensivo de dados pessoais aumenta o risco de violações de privacidade, especialmente em contextos com regulamentação insuficiente (ZUBOFF, 2019).

Outro desafio importante é a representatividade dos dados utilizados para treinar os algoritmos. Pesquisas indicam que modelos de IA apresentam desempenho inferior ao processar informações de populações sub-representadas, como pessoas negras e mulheres, por conta da predominância de dados oriundos de grupos majoritários (BUOLAMWINI; GEBRU, 2018). Essa limitação contribui para a ocorrência de discriminações em áreas sensíveis, como o reconhecimento facial, onde erros na identificação de indivíduos negros foram constatados (RAJI; BUOLAMWINI, 2019).

A responsabilidade na criação e implementação de sistemas de IA também merece atenção. A ausência de um *framework* ético durante o desenvolvimento pode permitir que os sistemas gerem impactos sociais negativos e decisões automatizadas sem mecanismos adequados de contestação (CRAWFORD et al., 2019). Essa situação reforça a necessidade de incorporar critérios éticos de modo criterioso, envolvendo tanto os desenvolvedores quanto as organizações responsáveis.

A opacidade dos modelos, sobretudo aqueles baseados em *deep learning*, limita a capacidade de auditar e corrigir decisões inadequadas. O desenvolvimento de métodos de explicabilidade oferece uma alternativa para revelar os processos internos desses modelos, aumentando a transparência e facilitando a responsabilização (DOSHI-VELEZ; KIM, 2017). No entanto, a aplicação prática dessas técnicas ainda apresenta desafios técnicos que precisam ser superados.

Por fim, a ausência de regulamentações globais padronizadas agrava os desafios éticos em IA. Iniciativas, como a Estratégia de IA da União Europeia (COMMISSION, 2020), apontam para a necessidade de diretrizes claras, mas há uma lacuna entre os princípios propostos e sua implementação prática. Essa realidade ressalta a urgência de normativas que garantam a equidade, a segurança e a proteção dos direitos humanos no desenvolvimento e uso dos sistemas de IA.

2.4 Vieses Algorítmicos: Origens, Tipos e Impactos

Os vieses em IA podem surgir de diversas fontes, incluindo os dados utilizados para treinar os modelos e as decisões tomadas durante o *design* do sistema (O'NEIL, 2016). Dados históricos frequentemente refletem desigualdades sociais, econômicas e culturais, o que pode levar os algoritmos a reproduzir essas disparidades. Um exemplo notável é o uso de algoritmos em processos seletivos de emprego que, baseados em dados históricos enviesados, desfavorecem mulheres e minorias (BINNS, 2018).

Além disso, os vieses podem ser introduzidos de forma inadvertida por desenvolvedores que não consideram variáveis contextuais ou culturais ao criar esses sistemas (MEHRABI et al., 2021). Esse fenômeno reforça a necessidade de diversificar as equipes de desenvolvimento, promovendo uma perspectiva mais ampla sobre os impactos sociais e éticos da tecnologia.

Os **vieses de agregação** ocorrem quando dados heterogêneos são combinados de forma inadequada, resultando em modelos que não refletem a diversidade da população. Por exemplo, em sistemas de saúde, a agregação de dados de diferentes grupos étnicos sem considerar variações biológicas pode levar a tratamentos menos eficazes para minorias. Um estudo de Obermeyer et al. (2019) mostrou que algoritmos de previsão de necessidades médicas priorizavam pacientes brancos em detrimento de pacientes negros, mesmo quando estes últimos tinham condições de saúde mais graves. Esse tipo de viés também é observado em sistemas financeiros, onde a agregação de dados de diferentes regiões pode penalizar comunidades de baixa renda ou minorias étnicas, persistindo desigualdades econômicas (BAROCAS; SELBST, 2016). Para mitigar esses vieses, é essencial segmentar os dados em subgrupos representativos e realizar avaliações de impacto específicas para cada população (MEHRABI et al., 2021).

Já os **vieses de interação com o usuário** surgem quando sistemas de IA aprendem com o comportamento dos usuários, perpetuando estereótipos e preconceitos. Esse tipo de viés é comum em plataformas de mídia social, onde algoritmos de recomendação priorizam conteúdo engajador, muitas vezes amplificando desinformação ou discursos de ódio. Pariser (2011) descreveu esse fenômeno como "filtro bolha" ou "câmaras de eco" onde os usuários são expostos apenas a informações que reforçam suas visões pré-existentes, levando à polarização. Em sistemas de recomendação, como os utilizados pela Netflix ou Spotify, os algoritmos podem reforçar estereótipos de gênero ou raça ao sugerir conteúdo com base em padrões históricos enviesados (SURESH; GUTTAG, 2021). Para combater esses vieses, é importante diversificar as recomendações e aumentar a transparência algorítmica, permitindo que os usuários entendam e controlem como as recomendações são geradas (RAJI et al., 2020).

Por fim, os **vieses comportamentais** referem-se a tendências inconscientes ou

padrões de comportamento humano que são capturados e replicados por sistemas de IA. Esses vieses podem ser introduzidos durante a coleta de dados ou através de interações com os usuários, resultando em modelos que refletem e amplificam preconceitos sociais. Um exemplo relevante é o viés racial em sistemas de reconhecimento facial, em que modelos treinados com dados predominantemente de pessoas brancas apresentam taxas de erro significativamente maiores para pessoas negras (BUOLAMWINI; GEBRU, 2018). Outro exemplo é o uso de algoritmos em processos seletivos de emprego, que podem perpetuar discriminação de gênero ou raça ao aprender com decisões históricas enviesadas (BINNS, 2018). Para mitigar esses vieses, é fundamental realizar auditorias regulares para identificar e corrigir problemas, além de garantir que os conjuntos de dados utilizados para treinamento sejam representativos de todas as populações relevantes (MEHRABI et al., 2021).

Esses três tipos de vieses ilustram a complexidade dos desafios éticos na IA e a necessidade de abordagens sólidas para garantir que os sistemas sejam justos, transparentes e responsáveis. A combinação de práticas consolidadas de governança de dados, diversidade nas equipes de desenvolvimento e regulamentações claras é essencial para mitigar os impactos negativos desses vieses e promover uma IA ética e inclusiva.

2.5 Estratégias para Mitigação dos Vieses

A mitigação dos vieses na IA requer uma abordagem multifacetada. É fundamental implementar práticas consolidadas de governança de dados, desenvolver *frameworks* éticos e promover auditorias regulares nos sistemas de IA para identificar e corrigir possíveis problemas (RAJI et al., 2020). Além disso, é essencial educar os desenvolvedores e demais *stakeholders* sobre ética na tecnologia, incentivando um *design* que considere os impactos sociais e garanta maior responsabilidade.

Para tanto, políticas claras devem ser estabelecidas para a coleta, armazenamento e utilização dos dados. Estudos apontam para a importância de adotar padrões globais na governança de dados, assegurando que os conjuntos utilizados para treinar modelos de IA sejam representativos e minimizem a introdução de vieses (JOBIN; IENCA; VAYENA, 2019). A implementação de controles internos e auditorias periódicas, conforme sugerido por Raji et al. (2020), é importante para monitorar e corrigir possíveis desvios nos processos.

A capacitação dos profissionais envolvidos no desenvolvimento e implementação de sistemas de IA é outro aspecto fundamental. Programas de treinamento que integrem aspectos técnicos e as implicações sociais das tecnologias emergentes podem ampliar a consciência sobre os riscos associados a decisões automatizadas (CRAWFORD et al., 2019). Essa formação contribui para a criação de um ambiente de inovação que priorize

práticas responsáveis e a transparência nos processos.

A transparência dos algoritmos também desempenha um papel importante na mitigação dos desafios éticos. Técnicas de explicabilidade, conforme discutido por Doshi-Velez e Kim (2017), permitem uma compreensão mais aprofundada dos mecanismos internos dos modelos de IA. Essa clareza não só facilita a identificação de vieses e erros, como também fortalece a confiança dos usuários e a eficácia das auditorias, proporcionando uma base sólida para ajustes e melhorias contínuas.

Por fim, a integração de políticas públicas e regulamentações é indispensável para alinhar o desenvolvimento da IA com princípios éticos e de justiça social. Iniciativas como a Estratégia de IA da União Europeia demonstram a necessidade de normativas que garantam a equidade e a proteção dos direitos individuais (COMMISSION, 2020). A cooperação entre governos, setor privado e academias científicas pode estabelecer um ambiente regulatório que sustente a inovação responsável e promova a inclusão no uso das tecnologias emergentes.

3 Mapeamento Sistemático da Literatura

Com o objetivo de compreender os desafios éticos associados aos vieses em algoritmos de IA, este estudo é a realização de um mapeamento sistemático da literatura (MSL). Para isso, serão examinadas publicações científicas e acadêmicas relevantes buscando mapear aspectos como os principais autores, países, instituições, palavras-chave, tipos de estudos e trabalhos mais influentes, além de analisar a evolução temporal das publicações.

3.1 Classificação Metodológica

O estudo dos desafios éticos associados aos vieses em algoritmos de IA exige uma abordagem metodológica estruturada para garantir uma análise ampla. A classificação metodológica deste estudo é definida conforme sua natureza, abordagem, objetivos e procedimentos técnicos.

Esta pesquisa é aplicada, pois busca investigar os desafios éticos relacionados aos vieses em algoritmos de IA e analisar como a literatura tem tratado essas questões. O foco está na identificação e análise dos dilemas éticos, bem como na sistematização dos resultados obtidos, fornecendo uma base sólida para futuras pesquisas (JOBIN; IENCA; VAYENA, 2019; CRAWFORD, 2021).

A abordagem metodológica adotada é quantitativa, justificando-se pela complexidade do tema e pela necessidade de compreender as percepções e implicações dos vieses algorítmicos em diferentes contextos. A análise quantitativa envolve a coleta e categorização de dados a partir dos estudos selecionados, registrando informações na ferramenta *Parsifal* [1], utilizada para realizar revisões e mapeamentos sistemáticos. Trata-se de uma ferramenta que auxilia na organização e classificação dos estudos, contribuindo para a consistência metodológica do trabalho (KITCHENHAM; CHARTERS, 2007).

Após a seleção dos estudos, um formulário é preenchido e estruturado, contendo categorias como: tipos de estudo, ano de publicação, etc. Esse processo assegura uma coleta de dados padronizada, permitindo comparações e análises sobre as publicações a respeito dos vieses na IA (BUOLAMWINI; GEBRU, 2018).

Os objetivos desta pesquisa são exploratórios e descritivos. Primeiramente, o estudo explora a literatura existente para mapear os principais autores, anos da publicação, países, tipos de publicação, palavras-chave e as métricas de citação nos estudos associados. Em seguida, descreve essas informações e analisa como diferentes estudos e tendências

[1] <https://parsif.al/>

podem contribuir para o avanço e melhorias para a mitigação dos impactos dos vieses algorítmicos (FLORIDI et al., 2018).

O mapeamento sistemático da literatura é conduzido por meio de um conjunto de procedimentos bem definidos, que incluem: planejamento, execução da busca em bases utilizando uma *string* de busca refinada, aplicação dos critérios de inclusão e exclusão, obtenção e análise dos dados selecionados e síntese dos resultados obtidos. Esse método possibilita a identificação das principais tendências, lacunas e contribuições acadêmicas, garantindo uma visão fundamentada da produção científica existente. O mapeamento segue critérios de inclusão e exclusão, assegurando a seleção dos estudos analisados. Esse processo serve como base para a compreensão dos desafios e para a proposição de direções futuras de pesquisa e práticas voltadas à mitigação de vieses em IA (RAJI et al., 2020).

Esta pesquisa aplicada, de abordagem quantitativa e com objetivos exploratórios e descritivos, utiliza o MSL para investigar os desafios éticos relacionados aos vieses em algoritmos de IA. A estrutura metodológica adotada garante uma análise do tema, contribuindo para a compreensão e servindo como base para o debate de busca por soluções que tornem a IA mais transparente, equitativa e responsável.

3.2 Plano Metodológico Adotado

O MSL é uma metodologia que busca organizar, classificar e sintetizar a produção científica em um determinado campo de estudo. Diferente da revisão sistemática, que se aprofunda na análise de evidências para responder a questões específicas, essa abordagem tem como objetivo fornecer uma visão geral sobre o panorama atual, identificando tendências, lacunas e padrões na literatura (KITCHENHAM, 2004). Amplamente utilizado em áreas complexas e multidisciplinares, como a IA, sua aplicação permite investigar de forma estruturada os desafios éticos e os vieses algorítmicos.

Segundo Kitchenham e Charters (2007), o mapeamento sistemático segue um processo estruturado dividido em três fases principais: planejamento, condução e síntese dos resultados. Na fase de planejamento, são definidos os objetivos da pesquisa, as questões a serem investigadas e o protocolo metodológico, incluindo os critérios de inclusão e exclusão dos estudos. Nessa etapa, também são selecionadas as bases de dados a serem consultadas e as estratégias de busca utilizadas para garantir a abrangência e a reprodutibilidade da pesquisa (PETERSEN et al., 2008).

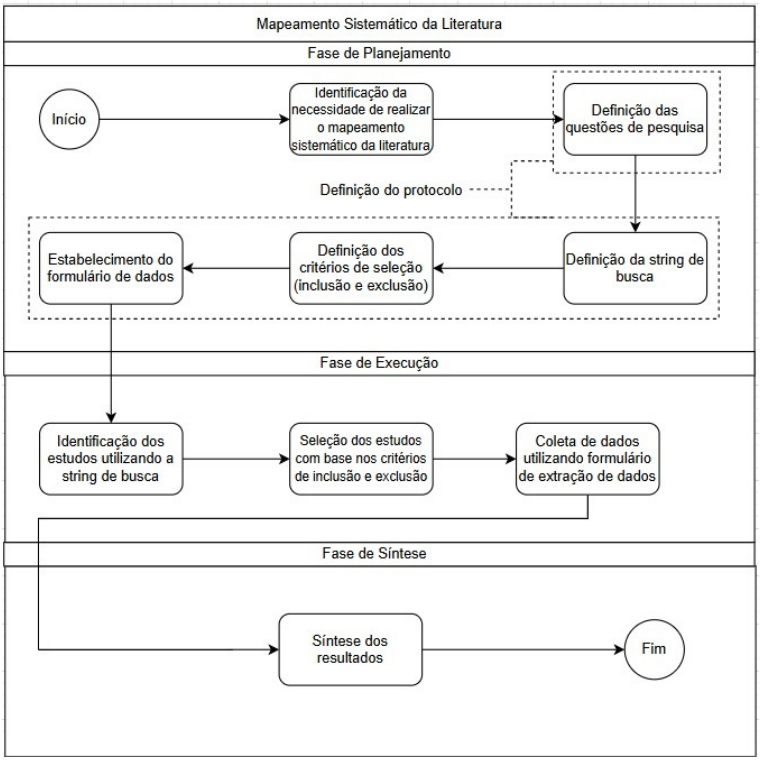
A fase de condução envolve a execução da busca nas bases selecionadas, a aplicação dos critérios de inclusão e exclusão, a categorização dos estudos e a extração dos dados relevantes. Para garantir a qualidade e a confiabilidade da análise, essa etapa pode incluir a revisão por pares e a organização dos resultados em tabelas ou gráficos, facilitando a identificação de padrões e lacunas na literatura (BRERETON et al., 2007). Além disso,

a categorização dos estudos permite compreender quais metodologias têm sido aplicadas e quais desafios éticos são mais frequentemente abordados no contexto dos vieses em IA.

Por fim, a fase de síntese dos resultados consiste na estruturação e apresentação dos achados do mapeamento de forma estruturada, destacando as tendências identificadas, as lacunas na pesquisa e as oportunidades para estudos futuros. A transparência no relato é essencial para que outros pesquisadores possam reproduzir ou expandir o estudo, contribuindo para a consolidação do conhecimento na área (OKOLI, 2015). Além disso, os resultados podem fornecer subsídios para a formulação de diretrizes éticas e estratégias de mitigação de vieses em algoritmos de IA (JOBIN; IENCA; VAYENA, 2019).

Dessa forma, o mapeamento sistemático se apresenta como a metodologia escolhida para compreender a evolução das discussões sobre ética e viés em IA, fornecendo uma base estruturada para futuras investigações e no desenvolvimento de práticas mais responsáveis no uso dessas tecnologias.

Figura 1 – Fases da pesquisa



Fonte: Autor

3.3 Fase de Planejamento

3.3.1 Identificação da necessidade de realizar o mapeamento sistemático

Conforme ilustrado na Figura 1, a fase de planejamento inicia-se com a identificação da necessidade de realizar uma revisão ou mapeamento da literatura. Esse processo é

fundamental para consolidar e sintetizar o conhecimento existente sobre um determinado tema. Segundo Felizardo (2017), essa etapa é essencial para garantir uma visão abrangente do estado da pesquisa e identificar lacunas que justifiquem novos estudos.

No contexto deste trabalho, a etapa de planejamento envolveu a definição da área de pesquisa e a delimitação do tema a ser analisado e apresentado. Com o objetivo de delinear o escopo da pesquisa, foi iniciada esta jornada com a leitura do capítulo 27.3 – *The Ethics of AI* – do livro *Artificial Intelligence: A Modern Approach* (RUSSELL; NORVIG, 2020). Essa leitura preliminar proporcionou uma visão ampla sobre os múltiplos subtemas relacionados à ética em IA. Durante discussões subsequentes, identificou-se como área de interesse as questões de justiça e transparência na IA, aspectos que despertaram maior curiosidade e relevância para o estudo.

Prosseguiu-se com a leitura do artigo *A Survey on Bias and Fairness in Machine Learning* (MEHRABI et al., 2021), que aprofundou a compreensão sobre os diferentes tipos de vieses presentes na IA. Este artigo destacou três categorias principais de vieses: *Data to Algorithm*, *Algorithm to User* e *User to Data*. Dentro dessas categorias, foram selecionados subtipos específicos para direcionar a pesquisa, incluindo *Aggregation Bias, Algorithmic Bias, Behavioral Bias, User Interaction Bias, Emergent Bias, Historical Bias e Population Bias*.

Dentre estes, foi priorizado o estudo dos tipos de vieses: *Aggregation Bias, User Interaction Bias e Behavioral Bias*. Essa priorização foi definida empiricamente após a análise inicial, que revelou estes vieses como particularmente relevantes devido a complexidade de seu conteúdo conceitual, a magnitude de seu impacto em aplicações práticas e a diversidade de suas formas de manifestação. A partir dessa definição, foi conduzida uma busca exploratória por artigos focados nesses subtipos, selecionando de dois a três estudos para cada viés. A análise desses artigos permitiu categorizar a manifestação dos vieses e identificar padrões e desafios éticos associados.

Com base nessas descobertas preliminares, foi realizada uma busca inicial utilizando uma *string* simples, que contemplava os principais conceitos identificados. Essa busca permitiu avaliar a abrangência das publicações disponíveis e definir os próximos passos da investigação, assegurando um mapeamento consistente e direcionado.

A metodologia adotada seguiu uma abordagem em camadas para a seleção dos estudos. Inicialmente, as palavras-chave foram analisadas para uma triagem preliminar. Em seguida, os resumos dos estudos foram avaliados para determinar sua relevância. Aqueles que continuaram pertinentes passaram por uma leitura mais detalhada da introdução e da conclusão. Somente os estudos que demonstraram real alinhamento com o tema ao longo dessas etapas foram lidos integralmente. Esse processo de seleção e análise confirmou a necessidade de realizar um MSL.

3.3.2 Definição do problema e questões de pesquisa

A IA tem se mostrado uma ferramenta poderosa para transformar setores como saúde, educação, finanças e justiça. No entanto, sua aplicação tem revelado desafios éticos significativos, especialmente relacionados aos vieses algorítmicos, que podem perpetuar ou amplificar desigualdades sociais existentes. Esses vieses surgem de diversas fontes, como dados desbalanceados, falhas no *design* dos algoritmos e a falta de diversidade nas equipes de desenvolvimento, resultando em decisões injustas ou discriminatórias (OBERMEYER et al., 2019; SELBST et al., 2019).

Diante desse cenário, o ponto central deste trabalho reside no mapeamento dos desafios éticos no contexto de vieses em algoritmos de IA, com foco na identificação de tendências de publicações, autores relevantes e na compreensão dos impactos dos vieses nas diferentes áreas de aplicação.

A questão de pesquisa que guia este estudo é: "Quais os principais desafios éticos associados aos vieses de agregação, de interação com o usuário e comportamental em algoritmos de IA atualmente?". Essa questão busca identificar e organizar, por meio de um mapeamento sistemático da literatura, as principais informações sobre desafios éticos associados aos vieses em algoritmos de IA atualmente, com foco nos seguintes pontos principais:

- Identificar os autores e grupos de pesquisa mais produtivos na área, a fim de mapear os principais polos de produção de conhecimento sobre vieses algorítmicos e ética em IA.

- Analisar a evolução temporal das publicações, identificando os períodos de maior produção científica e diagnosticando o crescimento do interesse acadêmico sobre o tema.

- Mapear os principais países e regiões de origem das publicações, analisando a distribuição geográfica da produção científica e identificando os focos de pesquisa internacionais.

- Identificar os focos temáticos e as frentes de pesquisa, analisando a frequência e a evolução das palavras-chave para diagnosticar mudanças de enfoque no debate sobre vieses em IA.

- Classificar os tipos de estudos predominantes, como artigos de periódicos e de conferência, a fim de determinar os principais canais de disseminação do conhecimento e a maturidade do campo de pesquisa.

- Identificar os trabalhos mais citados e influentes, analisando suas contribuições teóricas e conceituais para compreender como moldaram o debate sobre vieses e ética em IA.

A investigação desses aspectos é fundamental para avançar no entendimento do

cenário atual de pesquisa em torno da IA, oferecendo *insights* que podem orientar políticas públicas, regulamentações e práticas de desenvolvimento tecnológico. Dessa forma, este estudo contribui para o debate acadêmico.

3.3.3 Definição da *string* de busca

A construção da *string* de busca é uma etapa importante da metodologia, projetada para garantir uma recuperação de estudos que fosse ao mesmo tempo abrangente e precisa. A estratégia adotada baseou-se na combinação de três blocos conceituais distintos por meio de operadores booleanos, conforme detalhado na Tabela 1.

Bloco	**ID**	***String* de Busca**
Domínio	SB1	(("Inteligência Artificial" OR "IA" OR "*Artificial Intelligence*" OR "AI") AND
Problema	SB2	("Desafios Éticos" OR "*Ethical challenges*" OR "Viés" OR "*Bias*") AND
Foco	SB3	(("Agregação" OR "*Aggregation*") OR
	SB4	("Interação com o Usuário" OR "*User Interaction*") OR
	SB5	("Comportamental" OR "*Behavioral*")))

Tabela 1 – *Strings* de busca utilizadas na pesquisa.

- **Primeiro Bloco (Domínio):** Estabelece o campo geral da IA (SB1). O uso de sinônimos e siglas em português e inglês ("Inteligência Artificial", "AI") teve como objetivo maximizar a cobertura e incluir publicações em ambos os idiomas.
- **Segundo Bloco (Problema):** Refina a busca para o contexto do problema (SB2), exigindo a presença de termos relacionados a "vieses" ou "desafios éticos". Este passo garante a relevância temática dos artigos recuperados.
- **Terceiro Bloco (Foco Específico):** Direciona a pesquisa para os tipos de vieses focado neste estudo (SB3, SB4, SB5). Os termos "Agregação", "Interação com o Usuário" e "Comportamental" foram conectados pelo operador *OR* para assegurar que qualquer artigo que abordasse ao menos um dos vieses de interesse fosse incluído na seleção inicial.

Essa estrutura (IA) *AND* (Viés) *AND* (Tipo de Viés Específico), é uma forma de fechar o escopo para que os resultados sejam pertinentes, pois dessa forma cada estudo recuperado deve, obrigatoriamente, conter um termo de cada um dos três pilares conceituais.

3.3.4 Definição dos critérios de seleção

Os critérios de inclusão e exclusão são utilizados para guiar a seleção de estudos nesta pesquisa. Eles definem as características necessárias para a incorporação de tra-

balhos e as razões para sua exclusão, assegurando que os estudos selecionados estejam alinhados com os objetivos da pesquisa. Os critérios desta pesquisa são:

3.3.4.1 Critério de Inclusão

Foi definido um único e abrangente critério de inclusão para estabelecer o foco temático central da pesquisa. Este critério garante que todas as publicações selecionadas estejam diretamente alinhadas com o escopo do trabalho, que é a investigação dos vieses abordados neste trabalho.

Código	**Critério de Inclusão**
CI-1	O trabalho aborda ao menos um dos vieses (comportamental, de agregação e de interação com o usuário).

Tabela 2 – Critério de Inclusão.

3.3.4.2 Critérios de Exclusão

Os critérios de exclusão foram aplicados de forma sequencial para refinar a amostra, removendo estudos que, embora recuperados pela *string* de busca, não eram adequados para a análise aprofundada. Sendo eles:

Código	**Critérios de Exclusão**
CE-1	O trabalho não aborda sobre os vieses.
CE-2	O trabalho não está no idioma de inglês, português ou espanhol.
CE-3	O trabalho está duplicado.
CE-4	O trabalho foi publicado há mais de quatro anos.
CE-5	O trabalho não está disponível na *Scopus*.
CE-6	O trabalho não tem a versão completa disponível.
CE-7	O trabalho é um conjunto de estudos.

Tabela 3 – Critérios de Exclusão.

Cada critério desempenhou uma função específica na filtragem:

- **Relevância Temática (CE-1):** Este é o filtro mais importante, agindo como o inverso do critério de inclusão. Ele exclui artigos que, apesar de conterem as palavras-chave, não possuem como foco principal a análise de vieses.
- **Escopo Temporal e Linguístico (CE-2 e CE-4):** Para garantir a contemporaneidade da análise, foi estabelecido um limite temporal, excluindo trabalhos publicados há mais de quatro anos (ou seja, antes de 2021). Já o critério de idioma assegurou que os estudos estivessem em línguas acessíveis para a análise.
- **Acessibilidade e Formato (CE-3, CE-5, CE-6 e CE-7):** Estes critérios garantem a viabilidade da análise. Foram descartados trabalhos cujo texto completo

não pôde ser obtido ou que não estavam disponíveis na base de dados *Scopus* após a verificação, bem como aqueles que não eram publicações primárias completas ou eram algum tipo de compilado de estudos.

3.4 Fase de Execução

3.4.1 Identificação dos estudos utilizando a *string* de busca

A primeira etapa da fase de execução consistiu na identificação dos estudos por meio da aplicação da *string* de busca definida nas etapas anteriores. Para garantir a abrangência da pesquisa, a busca foi realizada na base de estudos *Scopus*. Ao rodar a *string* de busca no *Scopus*, retornaram 814 estudos iniciais, que foram exportados para a ferramenta *parsifal*, para facilitar a etapa de seleção.

3.4.2 Seleção dos estudos

Após a identificação inicial dos estudos, estes foram submetidos a um processo de seleção criterioso e sistemático, baseado nos critérios de inclusão e exclusão previamente estabelecidos. O objetivo desta fase foi refinar o conjunto de artigos, garantindo a máxima relevância e alinhamento com os objetivos da pesquisa. A seleção foi estruturada da seguinte forma:

- **Triagem por Título e Resumo:** Nesta fase inicial, o título e o resumo (*abstract*) de cada artigo foram examinados para uma primeira avaliação de elegibilidade. O estudo era imediatamente aprovado para a próxima etapa se o resumo indicasse claramente o seu alinhamento, rejeitado se indicasse o contrário ou caso caísse em algum outro critério de exclusão.
- **Critério de Desempate:** Nos casos em que a análise do resumo se mostrou ambígua ou insuficiente para tomar uma decisão segura sobre a inclusão ou exclusão, recorreu-se à leitura da introdução do artigo.

Os artigos que não atenderam aos critérios em qualquer uma das etapas foram descartados, e o motivo da exclusão foi devidamente registrado para garantir a transparência do processo metodológico.

3.4.3 Coleta dos dados através do formulário

A coleta dos dados foi realizada por meio de um formulário de extração. Este formulário foi criado como um instrumento metodológico projetado para capturar as informações necessárias para responder a cada um dos objetivos de pesquisa definidos no

início deste estudo. Cada campo do formulário foi pensado para fornecer os dados brutos para possíveis análises subsequentes.

Para o mapeamento da produção científica, por exemplo, o formulário incluiu campos como autor, afiliação, ano e país de publicação, tipo e fonte da publicação. Para a análise de impacto, o campo "Métrica de citação do estudo", já para a análise temática e conceitual, que constitui o cerne desta pesquisa, foram extraídos os campos "Palavras-chave", "Áreas de aplicação" e, de forma central, o "Viés abordado". A lista completa dos campos coletados foi a seguinte:

- Autor do estudo;
- Ano de publicação;
- País de publicação;
- Fonte de publicação;
- Afiliação;
- Palavras-chave;
- Tipo de publicação;
- Métrica de citação do estudo;
- Área de aplicação;
- Viés abordado.

A execução desta coleta foi conduzida utilizando a ferramenta *Parsifal* para centralizar as informações de cada estudo selecionado, por meio do preenchimento da aba de formulário de dados disponível na plataforma.

3.5 Fase de Síntese

Esta etapa foi realizada a partir da exportação dos artigos previamente selecionados e com os dados extraídos no *Parsifal* para uma planilha, que serviu de base para a visualização e análise. A sistematização dessas informações foi essencial para possibilitar a identificação de padrões e recorrências similares, bem como para destacar lacunas temáticas. Em seguida, a base foi integrada aos softwares *Looker Studio* e *VOSviewer*, por meio dos quais foram geradas visualizações interativas e análises bibliométricas que facilitaram a visualização dos resultados e o aprofundamento das interpretações.

A partir da categorização dos dados extraídos, foi possível construir um entendimento mais fundamentado sobre o panorama atual das pesquisas relacionadas aos vieses em algoritmos de IA. As abordagens mais recorrentes foram identificadas com base na frequência de ocorrência dos diferentes tipos de viés, nos principais autores, nas áreas de aplicação predominantes (como saúde, segurança pública e sistemas de recomendação) e na distribuição temporal das publicações. Os dados foram organizados de forma estruturada, permitindo análises estatísticas e visuais que revelaram padrões de coocorrência entre palavras-chave e tendências na evolução do debate ético ao longo dos anos.

A partir das visualizações produzidas, como tabelas, gráfico de bolha, mapas de calor e gráficos de distribuição, foi possível mapear as contribuições consolidadas na literatura, além de evidenciar áreas ainda pouco exploradas ou com baixa representatividade, sinalizando lacunas relevantes para investigações futuras.

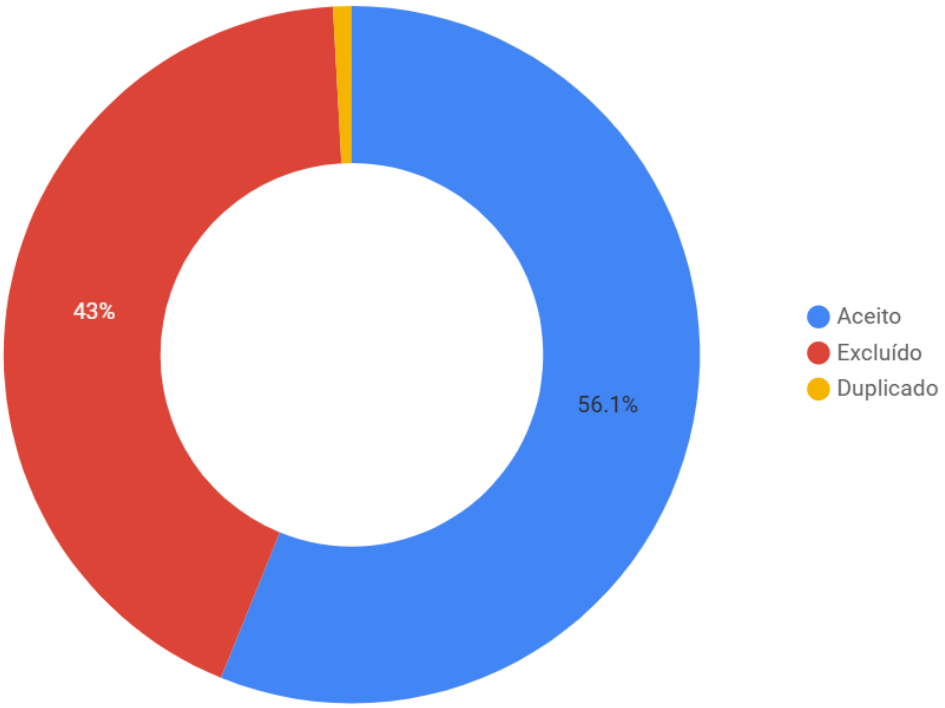
4 Resultados

O objetivo deste capítulo é organizar e sintetizar as informações coletadas, correlacionando os dados obtidos na etapa de extração com as questões éticas e os tipos de vieses presentes em algoritmos de IA. A análise foi estruturada a partir dos artigos selecionados, sendo complementada por visualizações gráficas que possibilitaram a identificação de padrões, tendências e lacunas na literatura. A seguir, são descritos os resultados alcançados na fase de síntese do MSL.

4.1 Resultado da Seleção dos Estudos

A partir da *strings* de busca definidas na Tabela 1, foi realizada a seleção dos estudos com base nos critérios de inclusão e exclusão estabelecidos. Essa busca resultou na identificação de 814 artigos provenientes da utilização dessa *string* de busca. Os documentos coletados passaram, então, por uma etapa de triagem, na qual foram avaliados quanto à elegibilidade.

Figura 2 – Porcentagem de publicações pós seleção

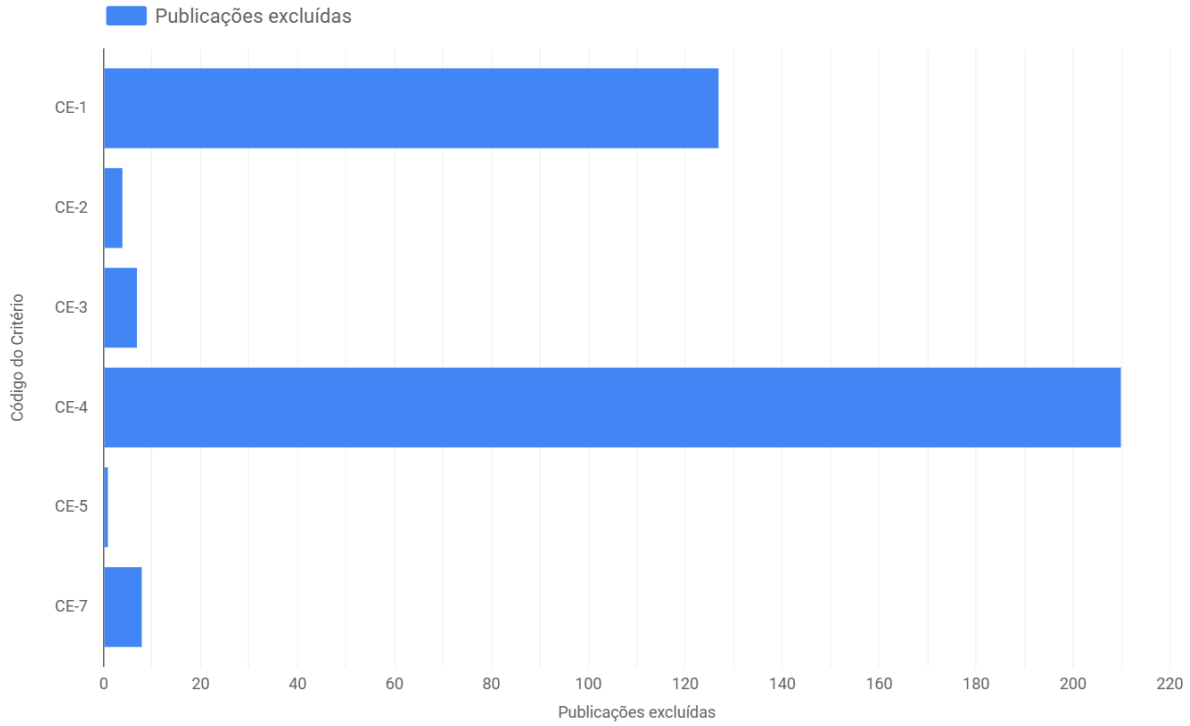


Fonte: Autor

Conforme apresentado na Figura 2, do total de artigos coletados, 457 foram aceitos para as etapas seguintes da análise, representando 56,17% do conjunto inicial de estudos.

Por outro lado, 350 publicações (42,99%) foram rejeitadas por atenderem aos critérios de exclusão, enquanto 7 artigos (0,86%) foram descartados por se tratarem de trabalhos duplicados. Esses números demonstram a abrangência do processo de filtragem e asseguram que a amostra analisada nesta pesquisa é composta por trabalhos relevantes e não redundantes.

Figura 3 – Publicações excluídas por código do critério



Fonte: Autor

Para compreender os fatores que levaram a essa taxa de exclusão, a Figura 3 detalha a quantidade de vezes em que cada critério de exclusão foi aplicado. A análise revela que dois critérios foram predominantes na filtragem dos trabalhos. O principal motivo para a exclusão foi o critério CE-4 (O trabalho foi publicado há mais de quatro anos), respondendo por mais de 200 artigos e demonstrando um foco deliberado em contribuições recentes. Em seguida, o critério CE-1 (O trabalho não aborda sobre os vieses) destacou-se como o segundo filtro que mais cortou publicações de serem incluídas no mapeamento, sendo responsável pela exclusão de aproximadamente 125 estudos que, embora pudessem tangenciar o tema de IA, não se aprofundavam na questão dos vieses. Os demais critérios tiveram um impacto consideravelmente menor no processo. Entretanto, a aplicação dos critérios de escopo temporal (CE-4) e de relevância temática (CE-1) foi decisiva para refinar o conjunto inicial de 814 artigos, garantindo que a amostra final de 457 trabalhos seja não apenas relevante dentro do tema de vieses e atual, mas também alinhada com os objetivos desta pesquisa.

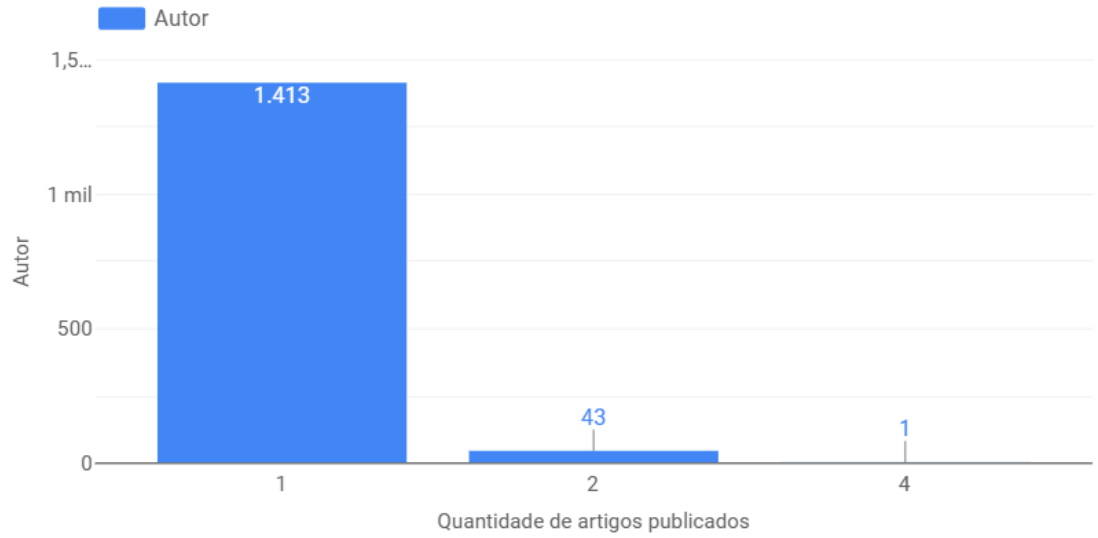
4.2 Resultado dos Indicadores Quantitativos

4.2.1 Principais autores

Com o intuito de responder ao que foi proposto no objetivo 1, realizou-se uma análise da atuação dos autores no campo dos vieses algorítmicos, buscando destacar aqueles que vêm contribuindo de forma recorrente para o avanço do debate. A investigação considerou tanto a recorrência de publicações por autor quanto os temas abordados em suas produções, permitindo uma visão mais qualificada do engajamento dos pesquisadores com a temática.

Conforme apresentado na Figura 4, observa-se uma concentração de autores com apenas uma publicação, o que pode indicar a natureza ampla e interdisciplinar do campo. Ainda assim, a presença de um núcleo reduzido de autores com produções recorrentes pode sugerir a existência de grupos especializados ou linhas de pesquisa mais consolidadas em torno da temática.

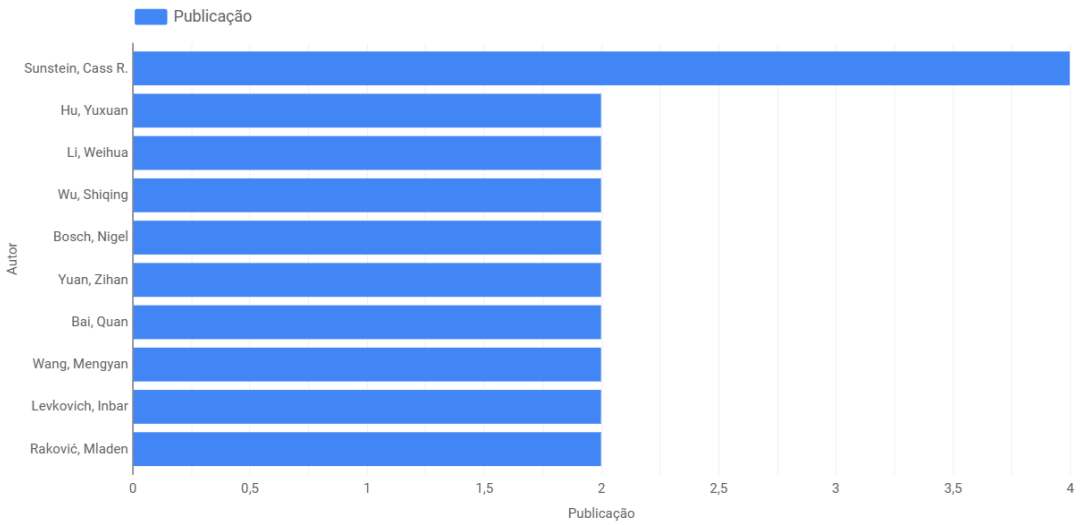
Figura 4 – Distribuição da quantidade de artigos por autor



Fonte: Autor

A Figura 5 detalha os autores com maior número de publicações na base analisada. Cass R. Sunstein aparece como o nome mais recorrente, seguido por autores como Yuxuan Hu, Weihua Li e Shiqing Wu, entre outros. A frequência com que esses autores aparecem sugere um envolvimento mais marcante com a temática dos vieses, principalmente pelo Cass R. Sunstein, que tem 4 publicações. Esses autores com 2 publicações possivelmente indicam a existência de redes de colaboração ou linhas institucionais de pesquisa que tratam dos impactos éticos da inteligência artificial.

Figura 5 – Autores com maior número de publicações



Fonte: Autor

Para além da frequência de publicações, analisou-se também o recorte temático dessas produções, com foco nos tipos de viés abordados. A Figura 6 apresenta essa relação, permitindo identificar padrões de especialização entre os autores. Observa-se que muitos dos nomes mais recorrentes concentram suas investigações no viés comportamental, como é o caso de Cass R. Sunstein, enquanto outros, como Yuxuan Hu e Mengyan Wang, distribuem suas contribuições entre diferentes categorias de viés.

Figura 6 – Tabela de publicação de autores por viés

			Vies / Publicação
Autor	Viés comportamental	Viés de agregação	Viés de interação com usuário
Sunstein, Cass R.	4	-	-
Raković, Mladen	2	-	-
Levkovich, Inbar	2	-	-
Bosch, Nigel	2	1	-
Wang, Mengyan	2	-	2
Hu, Yuxuan	2	-	2
Wu, Shiqing	2	-	2
Li, Weihua	2	-	2
Bai, Quan	2	-	2
Yuan, Zihan	2	-	2

Fonte: Autor

A predominância do viés comportamental entre os autores mais engajados evidencia uma atenção particular da comunidade científica aos efeitos que os sistemas de IA podem ter sobre padrões sociais e decisões humanas. Essa ênfase temática pode estar

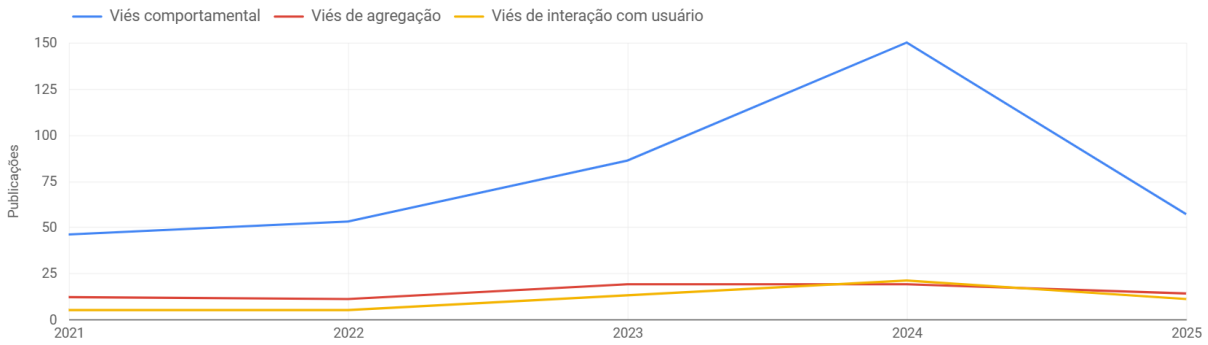
presente em áreas, como: justiça algorítmica, políticas públicas e plataformas de recomendação. Ainda que este trabalho não tenha explorado diretamente as redes formais de colaboração, a recorrência de determinados autores e a semelhança temática de suas publicações indicam possíveis vínculos acadêmicos ou atuação em grupos organizados de pesquisa.

Em síntese, os dados sugerem um campo em processo de avanço, no qual a contribuição de autores recorrentes tem papel central na estruturação e aprofundamento do debate ético sobre vieses algorítmicos. A análise aqui proposta reforça a importância de se considerar não apenas a quantidade de publicações, mas também a continuidade temática e a diversidade de abordagens dos pesquisadores envolvidos.

4.2.2 Tendências temporais das publicações

Para atender ao objetivo 2, realizou-se uma análise da distribuição anual das publicações identificadas, considerando também as áreas de aplicação abordadas nos estudos. Essa abordagem visa oferecer uma visão panorâmica da evolução do debate sobre os vieses algorítmicos ao longo do tempo, bem como dos contextos nos quais ele tem sido explorado. Levando em consideração que a *string* de busca foi usada para recuperar os artigos no início de 2025, dessa forma, em alguns gráficos o número de artigos pode parecer menor, e de fato é, porém faz sentido pelos poucos meses de 2025 quando a busca foi aplicada.

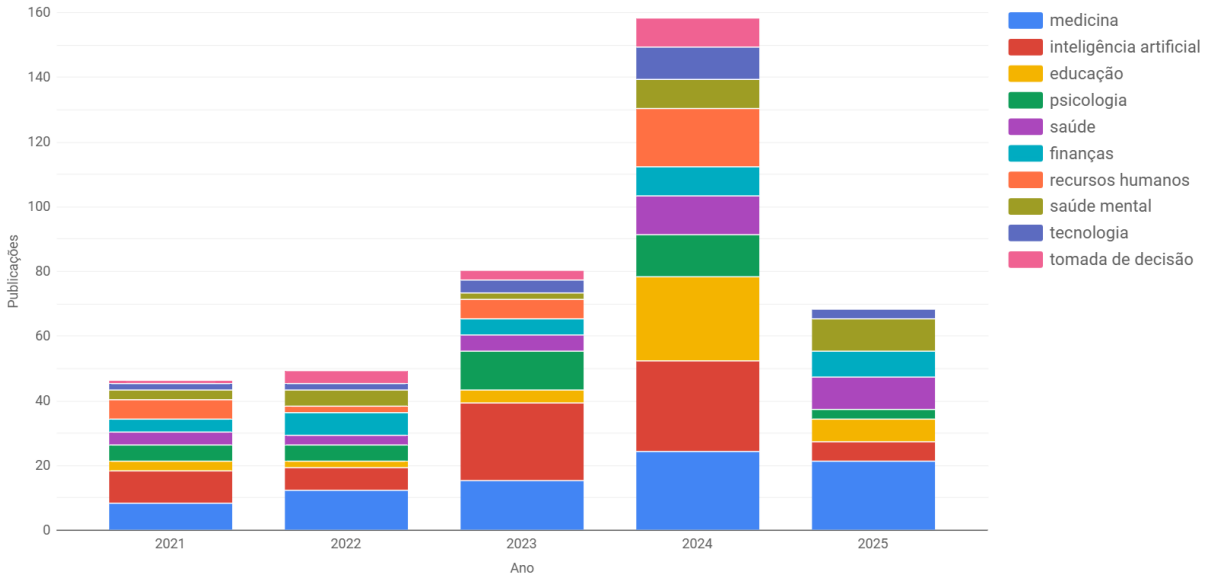
Figura 7 – Publicações pelo tipo de viés por ano



Fonte: Autor

Conforme ilustrado na Figura 7, observa-se um crescimento gradual no número de publicações a partir de 2021, com um aumento mais acentuado nos anos de 2023 e, especialmente 2024, que concentrou o maior volume de estudos no período analisado. Reafirmando que, embora a busca tenha sido realizada no início de 2025, o número já registrado sugere continuidade no interesse pela temática. Essa trajetória ascendente pode estar relacionada à consolidação gradual do interesse acadêmico pelo tema, bem como o reconhecimento crescente da importância dos vieses em sistemas baseados em inteligência artificial.

Figura 8 – Publicações por área de aplicação



Fonte: Autor

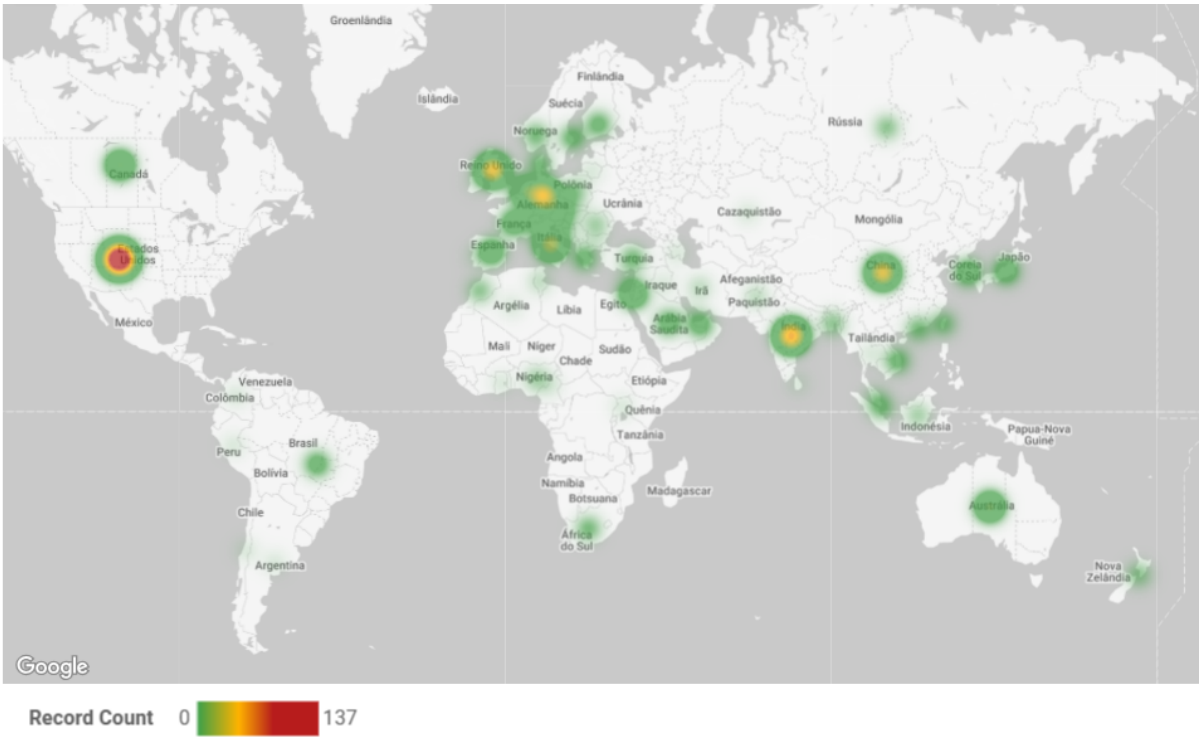
A Figura 8 complementa esse panorama ao demonstrar como as publicações se distribuíram entre diferentes áreas de aplicação. Observa-se que, ao longo dos anos, os estudos sobre vieses passaram a abranger um espectro mais amplo de domínios, com destaque para áreas como medicina, inteligência artificial, saúde mental, educação e sistemas de decisão automatizada. O ano de 2024, em particular, apresentou não apenas um pico no volume total de publicações, mas também uma diversificação mais expressiva nas áreas contempladas, o que pode indicar a expansão do debate para contextos mais variados.

Esses dados podem sugerir um movimento gradual de consolidação temática, no qual a discussão sobre vieses algorítmicos começa a ocupar espaços interdisciplinares. Por fim, apesar de oscilações pontuais ao longo dos anos, a tendência geral identificada é de fortalecimento do tema na literatura recente. Tal movimento pode estar associado à intensificação do uso de tecnologias baseadas em IA em setores sensíveis e à consequente demanda por mecanismos que promovam maior transparência, equidade e responsabilidade algorítmica (MITTELSTADT et al., 2016; MEHRABI et al., 2021).

4.2.3 Principais países e regiões responsáveis pelas publicações

Com o objetivo de aprofundar a análise proposta no objetivo 3, realizou-se um mapeamento geográfico com base nos países de publicação dos artigos presentes na base de dados. Essa etapa buscou identificar a distribuição espacial da produção científica sobre os desafios éticos associados aos algoritmos, permitindo visualizar em quais países o debate tem-se concentrado de forma mais significativa.

Figura 9 – Intensidade de publicações por países



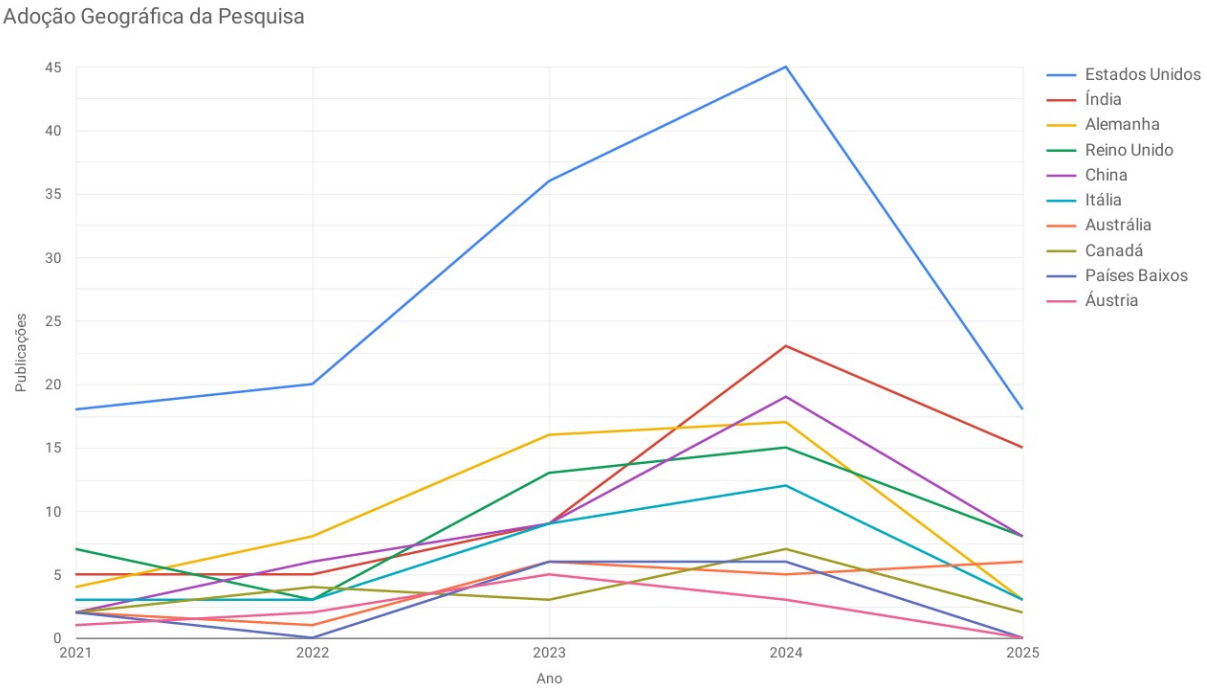
Fonte: Autor

A Figura 9 mostra uma maior densidade de publicações concentrada nos Estados Unidos, porém, embora o mapa de calor não permita inferências causais, a densidade observada em algumas regiões pode estar relacionada a fatores como maior volume de produção científica registrada ou presença de centros de pesquisa ativos na temática. Países europeus como Reino Unido, Alemanha, França, Países Baixos e Itália também apresentam participação significativa, o que pode refletir o envolvimento de ecossistemas de pesquisa já consolidados em debates sobre regulação e governança algorítmica (Council of Europe, 2021). Entretanto, é importante explicitar que a limitação de linguagem imposta nos critérios de exclusão tem relação direta com a formação dos gráficos geográficos, uma vez que pode direcionar a pesquisa. Explicando em parte o motivo de ter bastantes publicações em países anglófonos.

A distribuição temporal apresentada na Figura 10 complementa esse panorama ao mostrar como a contribuição de diferentes países evoluiu ao longo dos anos. Observa-se que os Estados Unidos lideraram a produção no período analisado, com um pico em 2024. No entanto, países como Índia e China também demonstraram crescimento expressivo, especialmente a partir de 2023. A ascensão da Índia em 2024, superando países tradicionalmente mais presentes, como Alemanha e Reino Unido, sugere o fortalecimento da participação asiática nesse campo de discussão (UNESCO, 2021). Contudo, é importante notar que a vasta densidade populacional da China e da Índia, que representam uma parcela significativa da população mundial, é um fator crucial que não pode ser ignorado

ao analisar o crescimento de suas contribuições no gráfico.

Figura 10 – Publicações de países por ano



Fonte: Autor

No caso da China, por exemplo, é possível que o aumento no volume de publicações esteja relacionado a estratégias nacionais de desenvolvimento tecnológico e inovação, como o plano "*Made in China* 2025" (WüBBECKE et al., 2016). Já o Japão, ainda que com menor visibilidade na amostra, tem historicamente contribuído para discussões sobre confiabilidade e segurança de sistemas inteligentes, o que pode indicar uma atuação mais especializada em vertentes técnicas da IA (OECD, 2019).

Além das regiões tradicionalmente mais representadas, a Figura 9 revela a presença de publicações provenientes da América Latina, Sudeste Asiático, Oceania e partes da África. Países como Brasil, Indonésia, África do Sul e Nova Zelândia aparecem com contribuições pontuais, demonstrando que o interesse pelo tema dos vieses algorítmicos não está restrito exclusivamente ao Norte Global. No entanto, ainda se observa uma assimetria expressiva na distribuição da produção científica, especialmente em países africanos, cuja presença nos dados é bastante limitada.

Esse panorama indica que, apesar da predominância de países mais frequentemente representados na base analisada, há sinais de participação mais distribuída entre diferentes regiões. A presença de múltiplos países contribui para uma visão mais abrangente sobre a temática, ainda que em níveis desiguais de envolvimento. Essa diversidade geográfica que começa a emergir é fundamental para enriquecer o debate, pois traz à tona diferentes

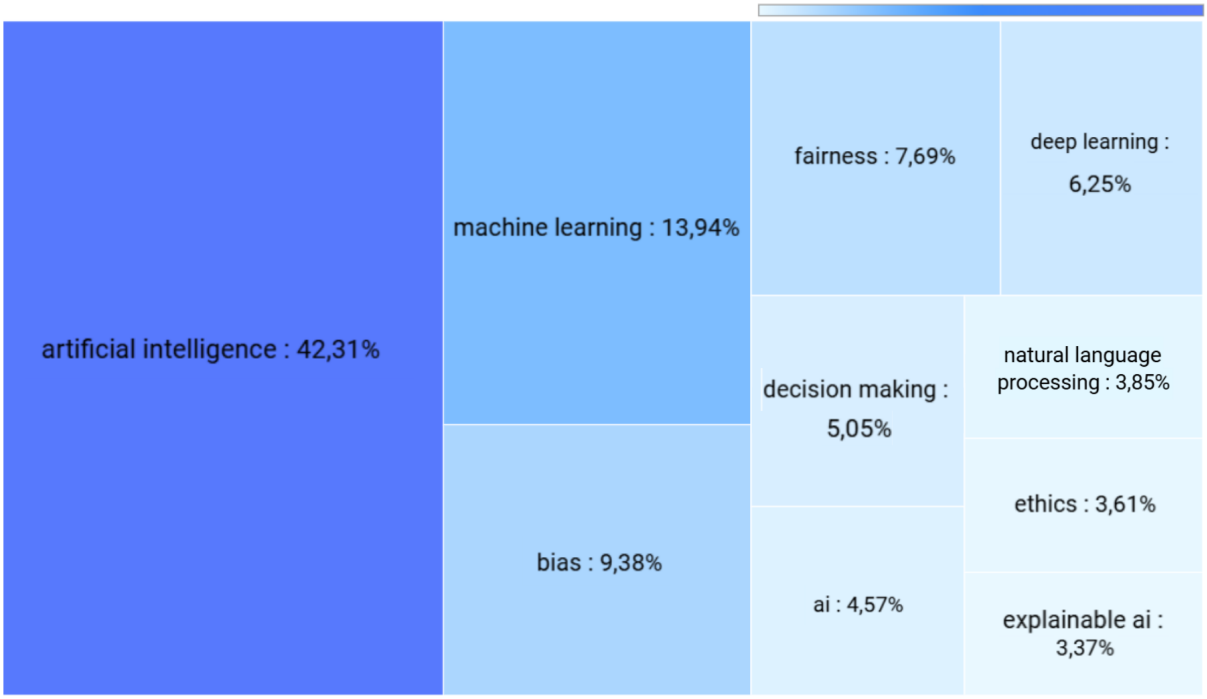
realidades sociotécnicas e amplia as perspectivas culturais, econômicas e políticas sobre os desafios éticos associados à inteligência artificial.

4.2.4 Frequência e coocorrência de palavras-chave

O presente resultado tem como foco principal o objetivo 4. Para isso, a análise de palavras-chave funciona como uma ferramenta de diagnóstico, revelando o núcleo do debate acadêmico. A frequência e a coocorrência desses termos mostram os tópicos mais estudados e como evoluíram ao longo do tempo, sendo possível observar mudanças de paradigma, que indicam como a comunidade científica passa da simples identificação de um problema para uma possível busca de soluções.

A primeira etapa para atingir essa meta é entender o panorama geral dos conceitos que dominam a área. O mapa de palavras-chave apresentado na Figura 11 oferece essa visão macro, exibindo as dez palavras-chave mais frequentes nos estudos selecionados. Na Figura 11, o tamanho e cor de cada retângulo é diretamente proporcional à frequência do termo, permitindo uma identificação visual imediata dos pilares conceituais do campo.

Figura 11 – Mapa de palavras-chave



Fonte: Autor

A interpretação imediata é a respeito do peso esmagador de termos mais técnicos. A palavra-chave *artificial intelligence* (42,31%) sozinha representa quase metade do universo conceitual. Somada a *machine learning* (13,94%), *deep learning* (6,25%), *natural language processing* (3,85%) e a abreviação *ai* (4,57%), os termos que descrevem a tecnologia em si compõem cerca de 70% das palavras-chave mais frequentes. Isso indica que a discussão

sobre vieses está ancorada no domínio técnico, ou seja, aponta que os problemas e as soluções são, em grande parte, enquadrados e investigados a partir de uma perspectiva computacional e de engenharia.

A métrica da palavra-chave *bias* (9,38%) confirma que o ponto central deste estudo é um dos focos nas publicações selecionadas, apesar de não ser o principal. No entanto, o que é mais revelador é a presença de *fairness* (justiça/equidade), com 7,69%. A frequência quase equivalente entre o problema (*bias*) e uma de suas principais soluções (*fairness*) sugere um amadurecimento no campo, que possivelmente olha para além da simples identificação de falhas, engajando-se também na busca por resultados desejáveis. Termos como *ethics* (3,61%), *decision making* (tomada de decisão, 5,05%) e *explainable ai* (IA explicável, 3,37%) complementam este quadro, apontando para frentes de pesquisa focadas nas consequências práticas, na governança e na transparência como mecanismos de mitigação.

Para ir além da frequência geral dos termos e compreender seu contexto de aplicação, a Figura 12 correlaciona as palavras-chave principais com cada tipo de viés estudado.

Figura 12 – Ocorrência de palavras-chave por viés

Palavra-chave	Viés comportamental	Viés de agregação	Viés de interação com usuário
artificial intelligence	152	26	27
machine learning	53	17	3
bias	38	11	7
fairness	30	11	1
deep learning	18	8	4
decision making	20	2	1
ai	16	1	3
natural language processing	15	-	1
ethics	12	3	4
explainable ai	10	4	-
behavioral research	13	1	-
cognitive bias	12	-	-
generative ai	10	2	3
behavioral economics	10	-	1
algorithmic bias	10	-	1

Fonte: Autor

Esta abordagem desvenda as afinidades entre os conceitos técnicos, éticos e os diferentes tipos de problemas, revelando as frentes de pesquisa com maior clareza. A análise desta correlação aponta para três pontos:

O primeiro é a clara dominância do viés comportamental, assim como em outros resultados já obtidos. Praticamente todas as palavras-chave, tanto técnicas quanto conceituais, têm sua maior frequência nesta categoria. Isso sugere fortemente que este viés é tratado como o problema central ou o campo de estudo mais prevalente na literatura selecionada, servindo como o principal catalisador para as discussões sobre vieses em IA.

Em seguida, é possível inferir que a tabela mostra que os conceitos éticos não são aplicados uniformemente. A palavra-chave *fairness*, por exemplo, é proeminente nas discussões sobre viés comportamental (30 ocorrências) e viés de agregação (11 ocorrências), mas quase desaparece quando o tema é viés de interação com o usuário (1 ocorrência). Isso indica que, provavelmente a busca por "justiça" algorítmica está mais focada em corrigir os modelos de decisão e a representação dos dados, e um pouco menos na interface direta com o usuário. De forma semelhante, *decision making* está quase que exclusivamente associado ao viés comportamental, reforçando a ideia de que o foco está em como as decisões automatizadas podem refletir ou induzir vieses de comportamento.

A análise também revela afinidades entre certas tecnologias e certos problemas. *Natural language processing*, por exemplo, tem uma presença relevante em estudos de viés de interação com o usuário, o que é lógico, já que a linguagem é o principal meio de interação. Em contraste, a palavra-chave XAI (*explainable ai*) aparece associada aos vieses comportamental e de agregação, mas está ausente na interação com o usuário. Isso pode sugerir que a necessidade de "explicar" o modelo é fundamental em sistemas complexos (como os de agregação de dados) e menos em interações mais diretas, onde outros fatores podem estar sendo mais priorizados.

4.2.5 Caracterização dos veículos de publicação

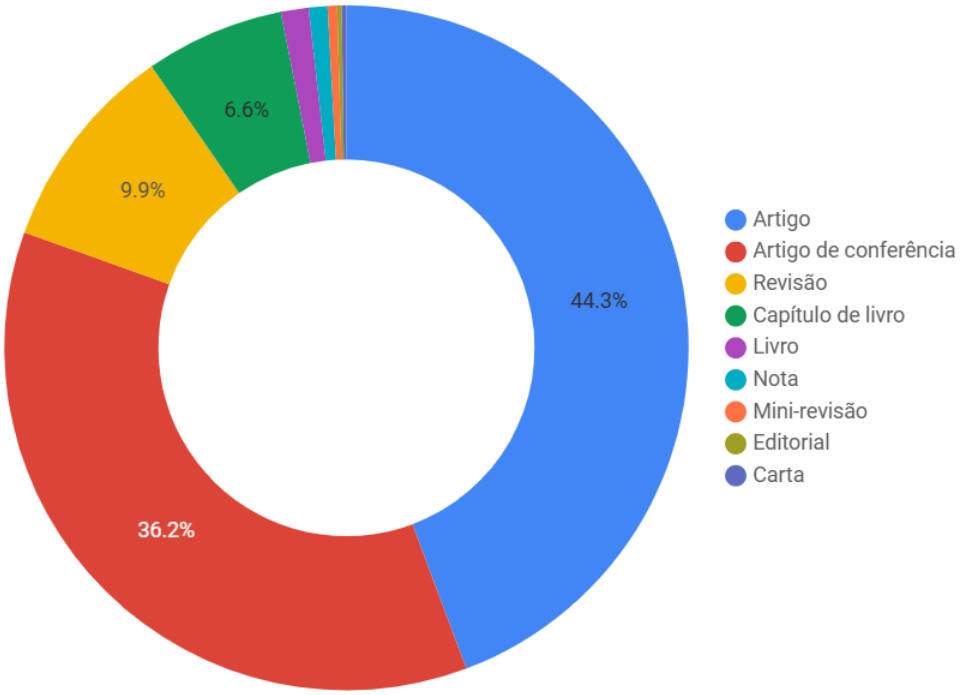
Para atingir o objetivo 5, foi analisada a distribuição dos tipos de publicação da amostra selecionada. A Figura 13 apresenta a distribuição percentual desses tipos, oferecendo uma visão clara dos formatos mais utilizados pela comunidade científica para comunicar suas pesquisas.

A análise deste gráfico revela que a pesquisa sobre vieses em IA é disseminada primariamente através de dois canais principais: artigos (presumivelmente de periódicos científicos) e artigos de conferência. Juntos, esses dois formatos compõem a esmagadora maioria da literatura, somando um total de 80.5% de todas as publicações desde 2021.

A participação de artigos como sendo a categoria mais prevalente, com 44,3% das publicações, sugere uma busca por validação rigorosa através da revisão por pares (*peer review*), o que é um indicador da consolidação e maturidade do campo. Artigos de conferência (36,2%) têm forte presença, representando mais de um terço das publicações, isso aponta para a natureza dinâmica e de rápida evolução da área. As conferências permitem uma disseminação mais ágil de novas descobertas, métodos e algoritmos, fomentando um debate acelerado entre os pesquisadores.

A coexistência e o peso quase equilibrado entre estes dois tipos de publicação pintam o retrato de um campo de pesquisa híbrido: maduro o suficiente para valorizar a profundidade e a consistência dos periódicos, porém ainda sendo dinâmico, necessitando

Figura 13 – Tipo de publicação

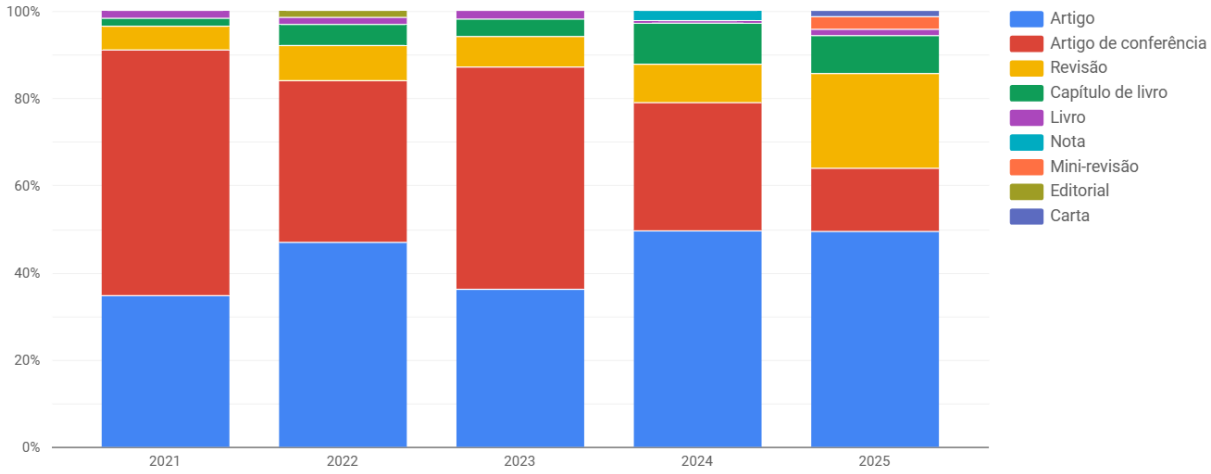


Fonte: Autor

dos canais ágeis das conferências.

Em um outro patamar, encontram-se as revisões (9,9%). A presença significativa desses artigos de revisão sinaliza que o campo pode ter acumulado um corpo de conhecimento substancial que demanda trabalhos de síntese e organização. Os demais formatos, como livros, notas e editoriais, representam uma fração menor do total, servindo provavelmente mais como canais de nicho para a disseminação do conhecimento na área.

Figura 14 – Tipo de publicação por ano

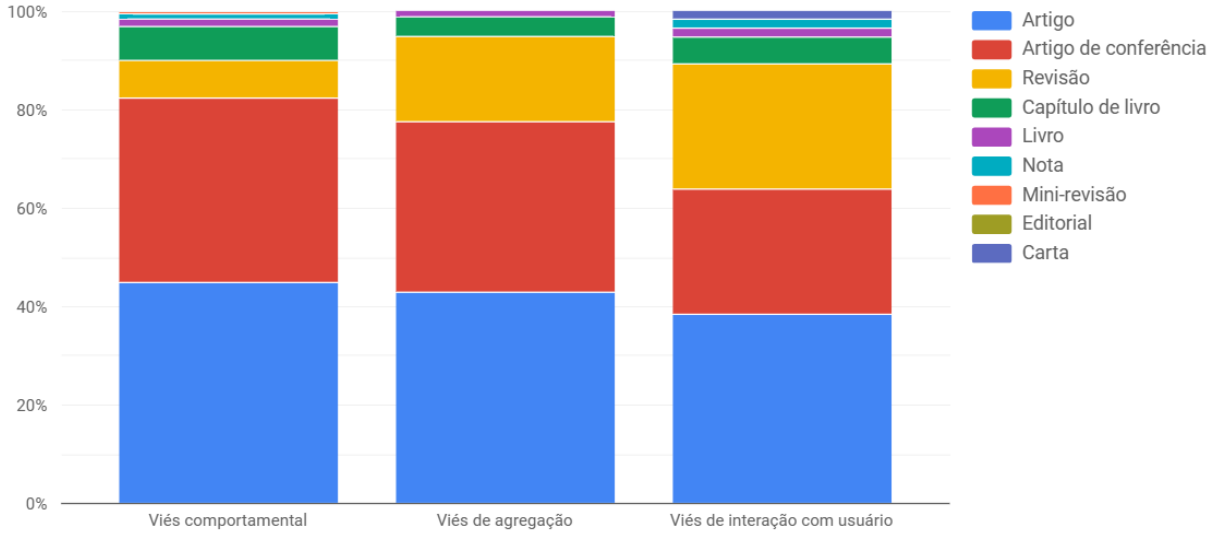


Fonte: Autor

A análise da distribuição dos tipos de publicação ao longo dos anos, apresentada na Figura 14, revela uma trajetória de evolução do campo. É notável a tendência de diminuição dos artigos de conferência e um crescimento correspondente na fatia dos artigos de periódicos, especialmente. Este padrão é um indicativo clássico de um campo de pesquisa, que está em transição de uma fase de disseminação rápida de ideias para uma fase que prioriza a precisão da revisão por pares e a profundidade analítica dos periódicos.

O *insight* mais perceptível, no entanto, é o crescimento expressivo da proporção de artigos de revisão nos anos mais recentes, com um salto proeminente em 2024 e 2025. Como mencionado anteriormente, esse fenômeno sugere fortemente que a área atingiu um momento em que o acúmulo de um volume de pesquisas primárias criou uma demanda e uma necessidade por trabalhos que sintetizem o conhecimento existente, mapeiem o território intelectual, identifiquem consensos e apontem as lacunas para guiar pesquisas futuras. O aumento no interesse por revisões é, portanto, um sintoma bastante forte de que o campo sobre vieses em IA está entrando em uma fase de maior reflexão e consolidação teórica.

Figura 15 – Tipo de publicação por viés



Fonte: Autor

Analisando como os canais de publicação se distribuem entre os diferentes tipos de vieses na Figura 15, percebe-se que a natureza do problema pode influenciar a forma como é comunicado. Embora o perfil seja relativamente semelhante entre as categorias, algumas nuances se destacam.

O debate em torno dos vieses de agregação e de interação com o usuário, por exemplo, apresenta uma maior proporção de artigos de revisão. À primeira vista, isso poderia indicar que esses vieses são os mais estudados ou os mais bem definidos do ponto de vista técnico. No entanto, essa impressão não se confirma ao analisarmos as métricas

relativas ao viés comportamental. Observa-se, assim, uma lacuna importante: embora a maior parte dos estudos estejam focados no viés comportamental, ainda são raras as revisões de literatura voltadas especificamente para esse tema.

O viés de interação com o usuário exibe uma fatia proporcionalmente menor de artigos de conferência e uma presença mais significativa de capítulos de livros. Esta distribuição sugere que as pesquisas sobre a interação com o usuário, que frequentemente envolvem fatores humanos, *design* e estudos qualitativos, podem encontrar um espaço mais adequado para seu desenvolvimento detalhado em formatos como artigos de periódicos. O viés comportamental, por sua vez, apresenta o perfil mais equilibrado dentre os três, refletindo uma natureza ampla que abrange tanto discussões teóricas quanto implementações técnicas.

4.2.6 Métricas de citações dos estudos

Para elencar os resultados do objetivo 6, foi realizada uma análise com base nas métricas de citação dos estudos selecionados.

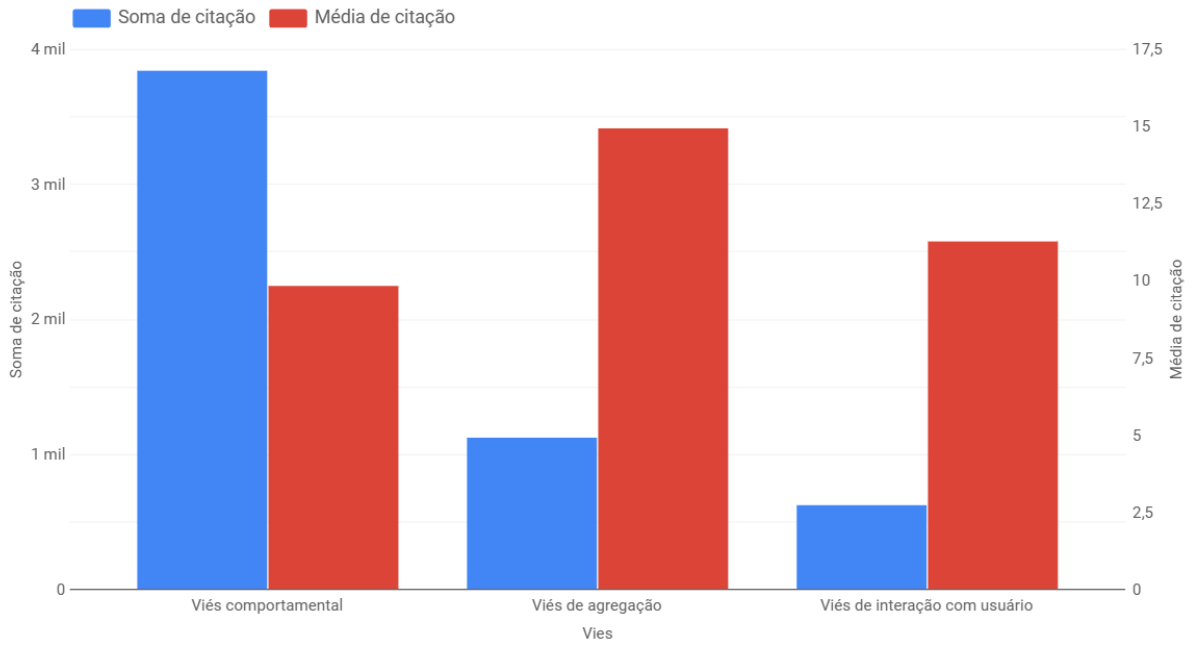
Figura 16 – Métricas de citação por publicação

Publicação	Métrica de citação ▾
Algorithmic bias: review, synthesis, and future research directions	299
On the genealogy of machine learning datasets: A critical history of ImageNet	143
Socially responsible AI algorithms: Issues, purposes, and challenges	125
Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identifi...	114
Explanations Can Reduce Overreliance on AI Systems During Decision-Making	103
AI and the transformation of social science research	102
Digital Mental Health for Young People: A Scoping Review of Ethical Promises and Challenges	101
A comprehensive review and analysis of supervised-learning and soft computing techniques for stress diagnosis in hum...	94
How cognitive biases affect XAI-Assisted decision-making: A systematic review	93
FairFed: Enabling Group Fairness in Federated Learning	92

Fonte: Autor

A análise inicial, focada nos trabalhos de maior impacto individual, conforme a Figura 16, revela as fundações conceituais do campo. É notável que os artigos mais citados são trabalhos de síntese e revisão, como "*Algorithmic bias: review, synthesis, and future research directions*" (MILANO; TADDEO; FLORIDI, 2020), que obteve 299 citações, e "*Bias and Unfairness in Machine Learning Models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods*" (PAGANO et al., 2023) com 114. Isso indica que a comunidade científica valoriza grandemente as contribuições que organizam, definem e mapeiam o território de pesquisa.

Figura 17 – Soma e média de citações por viés



Fonte: Autor

Ao aprofundar a análise para a distribuição de citações por viés, observam-se padrões mais específicos. O gráfico de citações por tipo de viés Figura 17 demonstra uma distinção clara entre o volume de debate e o impacto relativo de cada tema. O viés comportamental acumula o maior volume de citações (soma próxima de quatro mil), indicando que é o tópico mais extensivamente discutido na literatura. Contudo, é no estudo do viés de agregação que os trabalhos possuem, em média, o maior impacto individual, com uma média de quinze citações por artigo, superando a média de dez do viés comportamental.

Este achado sugere que, enquanto o viés comportamental constitui uma área de pesquisa ampla e com grande produção, o debate sobre viés de agregação, embora menor em volume, pode conter um núcleo de pesquisa mais denso, com artigos marcantes de maior influência relativa. A alta média de citações pode indicar que os trabalhos sobre este tema são mais técnicos ou fundacionais, servindo como base para um número maior de outras pesquisas.

4.3 Resultados Finais

Os resultados deste mapeamento sistemático revelam uma produção científica em amadurecimento, com crescente preocupação em compreender os riscos éticos associados aos vieses algorítmicos sob múltiplas dimensões. A análise da autoria reforça essa percepção, ao identificar um ecossistema de pesquisa de cauda longa, com uma base ampla de autores com contribuições únicas e um núcleo pequeno e restrito de autores com mais

publicações no tema. Ao entrelaçar autores, áreas de aplicação, regiões, palavras-chave, tipos de publicação e impacto acadêmico, observa-se uma convergência de esforços voltada à consolidação de um campo que busca não apenas diagnosticar os vieses (especialmente o comportamental), mas também formular respostas conceituais e metodológicas mais consistentes.

A prevalência de termos como *fairness*, *explainable AI* e *decision making*, aliada ao crescimento de revisões de literatura, indica um movimento de reflexão e busca por soluções mais estruturadas, que não se limitam à identificação de falhas, mas almejam maior transparência e justiça algorítmica. Nesse contexto, o Brasil está seguindo os primeiros passos para emergir como um ator relevante entre os países do Sul Global, contribuindo com uma produção qualificada que, embora ainda modesta em volume, ampliaria as perspectivas do debate ao inserir realidades sociotécnicas diversas, historicamente ausentes na literatura dominante.

Essa inserção representa um passo importante na direção de um ecossistema de pesquisa mais plural e sensível às desigualdades. Como contribuição, os resultados oferecem subsídios teóricos e empíricos para pesquisadores, formuladores de políticas públicas e desenvolvedores interessados em fundamentar práticas mais responsáveis no uso de IA. Ao mapear padrões e lacunas, este estudo não apenas sistematiza o conhecimento existente, como também orienta caminhos futuros para uma agenda ética e inclusiva, mais atenta às interseções entre tecnologia, justiça social e diversidade global.

5 Considerações finais

Este trabalho teve como propósito compreender os desafios éticos associados aos vieses em algoritmos de IA, por meio de um MSL. A partir da delimitação de três tipos de viés (comportamental, de agregação e de interação com o usuário), buscou-se identificar como esses vieses têm sido abordados na produção acadêmica recente. Para isso, foram analisadas diversas variáveis, como a área de aplicação da IA, o ano e o tipo de publicação (artigo, revisão, conferência, etc.), o número de citações, as palavras-chave atribuídas e o país de publicação dos estudos, em consonância com os objetivos específicos desta pesquisa.

Principais Resultados da Pesquisa

A análise revelou que o viés comportamental é o mais frequentemente abordado, com presença expressiva em diferentes áreas de aplicação, especialmente na saúde, na educação e nas ciências sociais. Esse viés está frequentemente relacionado a palavras-chave como *ethics*, *decision-making*, *cognitive biases* e *transparency*, o que demonstra uma preocupação com os impactos subjetivos e morais dos sistemas algorítmicos.

Por outro lado, os vieses de agregação e de interação com o usuário, embora com produção crescente, ainda são menos explorados na literatura, sendo o primeiro, recorrentemente ligado a discussões sobre qualidade de dados e estruturação de bases, enquanto o viés de interação surge em contextos mais recentes, como a popularização de assistentes virtuais e modelos generativos, apontando para um campo em expansão.

A análise da autoria revelou um cenário de pesquisa geograficamente concentrado nos Estados Unidos, Reino Unido, China e Índia. Tendo uma grande base de pesquisadores, porém a grande maioria possuindo apenas uma publicação na área, enquanto um núcleo restrito de autores detém duas ou mais publicações. Este pequeno núcleo de especialistas contêm grande parte de suas contribuições voltadas ao viés comportamental, o que evidencia o foco de especialização dos pesquisadores mais engajados no tema.

Outro achado relevante foi a identificação de uma intensificação do debate a partir de 2023, possivelmente impulsionada pelo avanço das tecnologias de IA generativa e pelo aumento da vigilância pública e regulatória. Os resultados demonstram uma crescente demanda por estudos relacionados aos vieses em IA nesse período, evidenciando uma fase de consolidação temática e metodológica, marcada por abordagens mais consistentes e aprofundadas por parte da comunidade científica.

Contribuições do Estudo

Como contribuição, este estudo oferece uma visão panorâmica e fundamentada sobre o conhecimento acumulado a respeito dos vieses algorítmicos. Os resultados apresentados podem servir de base para pesquisadores, formuladores de políticas públicas e desenvolvedores que desejam compreender os desafios éticos, bem como identificar as principais frentes de investigação e as lacunas existentes na literatura.

Limitações da Pesquisa

A principal fragilidade reside na restrição linguística dos estudos, que foram limitados aos idiomas inglês, português e espanhol, conforme um dos critérios de exclusão. Essa delimitação, embora necessária, pode ter introduzido um viés na análise geográfica, sub-representando a produção científica de países e regiões que publicam extensivamente em outras línguas. A escolha foi uma contrapartida metodológica para garantir a viabilidade da análise, uma vez que a leitura dos resumos era indispensável para a classificação dos tipos de viés. Dessa forma, os resultados referentes à distribuição de publicações por país devem ser interpretados com cautela, pois podem não refletir a totalidade da produção científica global sobre o tema.

Adicionalmente, outra limitação metodológica refere-se ao escopo dos dados efetivamente analisados. Embora informações relevantes tenham sido obtidas na etapa de extração de dados, variáveis como a afiliação institucional dos autores e a fonte de publicação não foram incorporadas na análise final. A decisão de excluí-las deveu-se à alta complexidade para a padronização e tratamento desses dados, o que demandaria um esforço de normalização léxica. É importante frisar, no entanto, que todas as demais variáveis extraídas foram sistematicamente utilizadas para alcançar os objetivos propostos.

Recomendações para Trabalhos Futuros

Para trabalhos futuros, recomenda-se aprofundar a análise qualitativa dos artigos, a fim de extrair de forma mais detalhada as nuances das estratégias de mitigação propostas. Sugere-se também a ampliação do recorte geográfico para incluir deliberadamente perspectivas de regiões sub-representadas, especialmente da América Latina e da África.

A análise das variáveis de afiliação institucional e fonte de publicação, que não foram exploradas neste estudo, também se apresenta como uma sugestão relevante; um exame aprofundado desses dados poderia revelar redes de colaboração entre instituições e identificar os periódicos e conferências mais influentes na disseminação da pesquisa sobre vieses.

Além disso, futuros estudos podem explorar a evolução dos vieses ao longo do ciclo de vida dos algoritmos, considerando desde o *design* e a coleta de dados até a aplicação

final e o monitoramento. Outra frente promissora é o desenvolvimento de *frameworks* avaliativos que incorporem indicadores éticos e sociais, permitindo mensurar o impacto real dos vieses nos diferentes contextos de aplicação. Por fim, seria relevante ampliar o diálogo entre abordagens técnicas e críticas, promovendo pesquisas interdisciplinares que integrem os campos da ciência da computação, das ciências sociais e da filosofia da tecnologia.

Referências

ABRAMSON, J. et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, Nature Publishing Group UK, 2024. Citado na página 29.

AWARDS, T. C. *'AI for Good' Award*. 2024. Disponível em: <https://www.cloud-awards.com/ai-awards/ai-for-good-award>. Citado na página 23.

BAROCAS, S.; HARDT, M.; NARAYANAN, A. *Fairness and machine learning*. fairmlbook.org, 2019. Disponível em: <https://fairmlbook.org/>. Citado na página 24.

BAROCAS, S.; SELBST, A. D. Big data's disparate impact. *California Law Review*, JSTOR, v. 104, n. 3, p. 671–732, 2016. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899>. Citado na página 31.

BINNS, R. Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, ACM, p. 149–159, 2018. Disponível em: <https://proceedings.mlr.press/v81/binns18a.html>. Citado 3 vezes nas páginas 24, 31 e 32.

BOSTROM, N.; YUDKOWSKY, E. The ethics of artificial intelligence. In: _____. *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, 2011. p. 316–334. Disponível em: <https://nickbostrom.com/ethics/artificial-intelligence.pdf>. Citado 2 vezes nas páginas 27 e 28.

BRERETON, P. et al. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, v. 80, n. 4, p. 571–583, 2007. Disponível em: <https://www.researchgate.net/publication/222555385_Lessons_from_applying_the_systematic_literature_review_process_within_the_software_engineering_domain_J_Syst_Softw>. Citado na página 35.

BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: PMLR. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. 2018. p. 77–91. Disponível em: <https://proceedings.mlr.press/v81/buolamwini18a.html>. Citado 4 vezes nas páginas 24, 30, 32 e 34.

BURRELL, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, v. 3, n. 1, 2016. Citado na página 23.

CATH, C. et al. Artificial intelligence and the 'good society': The us, eu, and uk approach. *Science and Engineering Ethics*, Springer, v. 24, n. 2, p. 505–528, 2018. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/28353045/>. Citado na página 24.

COLUMBIA, T. U. of B. *Microsoft AI for Humanitarian Action*. 2024. Disponível em: <https://research.ubc.ca/microsoft-ai-humantiarian-action#:~:text=Microsoft%20has%20opened%20grant%20proposal,biggest%20challenges%20facing%20society%20today.> Citado na página 23.

COMMISSION, E. *White Paper on Artificial Intelligence - A European approach to excellence and trust*. [S.l.], 2020. Disponível em: <https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en>. Citado 2 vezes nas páginas 30 e 33.

Council of Europe. *Artificial Intelligence, Human Rights, Democracy and the Rule of Law: A Primer*. Strasbourg: [s.n.], 2021. Disponível em: <https://www.coe.int/en/web/artificial-intelligence>. Citado na página 50.

CRAWFORD, K. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven and London: Yale University Press, 2021. ISBN 978-0-300-20957-0. Disponível em: <https://www.essra.org.cn/upload/202105/The%20Atlas%20of%20AI%20Power,%20Politics,%20and%20the%20Planetary%20Costs%20of%20Artificial%20Intelligence.pdf>. Citado na página 34.

CRAWFORD, K. et al. *AI Now 2019 Report*. [S.l.], 2019. Disponível em: <https://ainowinstitute.org/publication/ai-now-2019-report-2>. Citado 2 vezes nas páginas 30 e 32.

DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. Disponível em: <https://arxiv.org/abs/1702.08608>. Citado 2 vezes nas páginas 30 e 33.

FELIZARDO, K. *Revisão Sistemática: a importância da literatura científica na prática baseada em evidências*. São Paulo: Editora Schoba, 2017. ISBN 978-85-8258-319-1. Citado na página 37.

FLORES, M. da C. *ARISTÓTELES. ÉTICA A NICÔMACO*. Principis, 2021. (Clássicos da literatura mundial). ISBN 9786555524222. Disponível em: <https://books.google.com.br/books?id=UJEkEAAAQBAJ>. Citado na página 27.

FLORIDI, L. et al. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, Springer, v. 28, n. 4, p. 689–707, 2018. Disponível em: <https://link.springer.com/article/10.1007/s11023-018-9482-5>. Citado 2 vezes nas páginas 30 e 35.

FLORIDI, L.; SANDERS, J. W. On the morality of artificial agents. *Minds and Machines*, v. 14, n. 3, p. 349–379, 2004. Disponível em: <https://dl.acm.org/doi/abs/10.1023/b%3Amind.0000035461.63578.9d>. Citado 2 vezes nas páginas 27 e 28.

JOBIN, A.; IENCA, M.; VAYENA, E. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, Nature Publishing Group, v. 1, n. 9, p. 389–399, 2019. Disponível em: <https://www.researchgate.net/publication/335579286_The_global_landscape_of_AI_ethics_guidelines>. Citado 4 vezes nas páginas 30, 32, 34 e 36.

JOHNSON, D. G. *Computer Ethics*. Prentice Hall, 2008. Disponível em: <https://mrce.in/ebooks/Computer%20Ethics%204th%20Ed.pdf>. Citado 2 vezes nas páginas 27 e 28.

KANT, I.; QUINTELA, P. *Fundamentação da metafísica dos costumes*. Edições 70, 1997. (Textos filosoficos). ISBN 9789724403069. Disponível em: <https://books.google.com.br/books?id=GoF_PAAACAAJ>. Citado na página 27.

KITCHENHAM, B. *Procedures for Performing Systematic Reviews*. Keele, UK, 2004. Disponível em: <https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf>. Citado na página 35.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing systematic literature reviews in software engineering*. [S.l.], 2007. Disponível em: <http://researchgate.net/publication/302924724_Guidelines_for_performing_Systematic_Literature_Reviews_in_Software_Engineering>. Citado 2 vezes nas páginas 34 e 35.

MASLEJ, N. et al. *Artificial Intelligence Index Report 2024*. 2024. Citado na página 29.

MEHRABI, N. et al. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, ACM, v. 54, n. 6, p. 1–35, 2021. Disponível em: <https://dl.acm.org/doi/10.1145/3457607>. Citado 5 vezes nas páginas 23, 31, 32, 37 e 49.

MILANO, S.; TADDEO, M.; FLORIDI, L. Algorithmic bias: review, synthesis, and future research directions. *Science and engineering ethics*, Springer, v. 26, n. 2, p. 553–583, 2020. Citado na página 57.

MITTELSTADT, B. D. et al. The ethics of algorithms: Mapping the debate. *Big Data & Society*, v. 3, n. 2, p. 1–21, 2016. Disponível em: <https://doi.org/10.1177/2053951716679679>. Citado na página 49.

NEGNEVITSKY, M. *Artificial Intelligence: A Guide to Intelligent Systems*. 2. ed. Harlow: Pearson Education, 2005. Disponível em: <https://www.academia.dk/BiologiskAntropologi/Epidemiologi/DataMining/Artificial_Intelligence-A_Guide_to_Intelligent_Systems.pdf>. Citado na página 29.

NOBLE, S. U. *Algorithms of Oppression: How Search Engines Reinforce Racism*. [S.l.]: NYU Press, 2018. Citado na página 23.

OBERMEYER, Z. et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, American Association for the Advancement of Science, v. 366, n. 6464, p. 447–453, 2019. Disponível em: <https://www.science.org/doi/10.1126/science.aax2342>. Citado 3 vezes nas páginas 23, 31 e 38.

OECD. *Recommendation of the Council on Artificial Intelligence*. Paris, 2019. Disponível em: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>. Citado na página 51.

OKOLI, C. *A Guide to Conducting a Systematic Literature Review of Information Systems Research*. 2015. Cardiff School of Business and Management, Sprouts. Disponível em: <https://www.researchgate.net/publication/228276975_A_Guide_to_Conducting_a_Systematic_Literature_Review_of_Information_Systems_Research>. Citado na página 36.

O'NEIL, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, 2016. Disponível em: <https://www.researchgate.net/publication/314165204_Cathy_O'Neil_Weapons_of_Math_Destruction_How_Big_Data_Increases_Inequality_and_Threatens_Democracy_New_York_Crown_Publishers_2016_272p_Hardcover_26_ISBN_978-0553418811>. Citado na página 31.

PAGANO, T. P. et al. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, v. 7, n. 1, 2023. ISSN 2504-2289. Disponível em: <https://www.mdpi.com/2504-2289/7/1/15>. Citado na página 57.

PARISER, E. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Books, 2011. Disponível em: <https://www.semanticscholar.org/paper/The-Filter-Bubble%3A-How-the-New-Personalized-Web-Is-Pariser/936a6393eb42fba75a228cd339b7c9ba36e3f696>. Citado na página 31.

PETERSEN, K. et al. Systematic mapping studies in software engineering. In: *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*. [s.n.], 2008. p. 68–77. Disponível em: <https://www.researchgate.net/publication/228350426_Systematic_Mapping_Studies_in_Software_Engineering>. Citado na página 35.

RAJI, I. D.; BUOLAMWINI, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In: *Proceedings of the AAAI/ACM Conference on AI, ethics, and society, association for computing machinery, New York, AIES*. [s.n.], 2019. p. 429–435. Disponível em: <https://doi.org/10.1145/3306618.3314244>. Citado na página 30.

RAJI, I. D. et al. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: ACM. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020. p. 33–44. Disponível em: <https://arxiv.org/abs/2001.00973>. Citado 4 vezes nas páginas 24, 31, 32 e 35.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 4. ed. Pearson, 2020. Disponível em: <http://lib.ysu.am/disciplines_bk/efdd4d1d4c2087fe1cbe03d9ced67f34.pdf>. Citado 4 vezes nas páginas 23, 28, 29 e 37.

SELBST, A. D. et al. Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, p. 59–68, 2019. Disponível em: <https://dl.acm.org/doi/10.1145/3287560.3287598>. Citado 2 vezes nas páginas 24 e 38.

SURESH, H.; GUTTAG, J. V. A framework for understanding unintended consequences of machine learning. *Communications of the ACM*, ACM New York, NY, USA, v. 64, n. 10, p. 62–71, 2021. Disponível em: <https://arxiv.org/abs/1901.10002>. Citado 2 vezes nas páginas 24 e 31.

UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. Paris, 2021. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>. Citado na página 50.

WüBBECKE, J. et al. *Made in China 2025: The making of a high-tech superpower and consequences for industrial countries*. Berlim: [s.n.], 2016. Disponível em: <https://merics.org/en/report/made-china-2025>. Citado na página 51.

ZUBOFF, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, 2019. Disponível em: <https://www.hbs.edu/faculty/Pages/item.aspx?num=56791>. Citado na página 30.