Cenário da Pesquisa sobre Ética em IA: Tendências e Desafios

Katia Emanuelly de Souza[1], **Carlos Henrique Tavares Brumatti**[1], **Cristiane Aparecida Lana**[1,2], **Maria Lúcia Bento Villela**[1]

[1]Departamento de Informática – Universidade Federal de Viçosa (UFV)
Viçosa – MG – Brasil

[2]Universidade Federal de Lavras (UFLA-Paraíso)
São Sebastião do Paraíso – MG – Brasil

{katia.souza, carlos.h.tavares, maria.villela}@ufv.br, cristiane.lana@ufla.br

Abstract. Introduction*: With the popularization of Artificial Intelligence in recent years, new issues have been debated, especially regarding its ethical aspects.* ***Objective:*** *The objective is to explore the research landscape on AI ethics through a tertiary literature review.* ***Methodology or Steps****: An analysis of secondary studies published between 2020 and 2025 was conducted, focusing on ethics and responsibility in AI. The data were analyzed qualitatively, identifying recurring themes, gaps, and challenges.* ***Results:*** *The studies highlighted recurring themes such as AI privacy and security. Challenges were also identified, such as the lack of references to existing standards.*
Keywords *Ethic, AI, Tertiary review, Responsibility*

Resumo. Introdução: *Com a popularização da Inteligência Artificial nos últimos anos, novas questões passaram a ser debatidas, especialmente em relação aos seus aspectos éticos.* ***Objetivo:*** *O objetivo é explorar o panorama da pesquisa sobre ética em IA por meio de uma revisão terciária da literatura.* ***Metodologia ou Etapas:*** *Foi realizada uma análise de estudos secundários publicados entre 2020 e 2025, com foco em ética e responsabilidade em IA. Os dados foram analisados qualitativamente, visando identificar temas recorrentes, lacunas e desafios.* ***Resultados:*** *Os estudos analisados destacam temas recorrentes como privacidadee e segurança da IA. Também foram identificadas dificuldades como a ausência de referências a normas existentes.*
Palavras-Chave *Ética, IA, Revisão terciária, Responsabilidade*

1. Introdução

A Inteligência Artificial (IA) tem se tornado cada vez mais presente nas atividades cotidianas das pessoas, sendo aplicada em uma variedade de contextos, como em compras online, interações em redes sociais e produção de textos. Se, por um lado, a IA promove avanços substanciais em áreas como saúde, educação, segurança e agronegócio, por outro, ela levanta preocupações éticas cada vez mais urgentes [Martins e Viana 2022]. Esses sistemas podem gerar decisões enviesadas, comprometendo a equidade[1] e a confiança nas suas conclusões. Além disso, a IA pode infringir direitos fundamentais de privacidade e

[1]Equidade é a "qualidade que consiste em reconhecer imparcialmente o direito de cada um, segundo a sua razão, dando a cada um o que lhe é devido" [Dicio – Dicionário Online de Português 2024].

segurança, muitas vezes sem a devida transparência sobre os critérios que orientam suas decisões [Capel e Brereton 2023, Machado 2024].

Esse cenário tem impulsionado debates globais sobre a necessidade de regulação ética e jurídica da IA [Sichman 2021]. Iniciativas como o AI Act da União Europeia (em vigor desde 2024) [Santos 2024] e o Projeto de Lei 2.338/2023 no Brasil [BRASIL. Ministério da Cultura 2024] demonstram esforços regulatórios voltados à mitigação de riscos e ao uso responsável da IA. Paralelamente, organizações internacionais e empresas privadas têm proposto diretrizes e princípios voluntários com foco em valores como transparência, justiça e segurança, a exemplo do IEEE 7000-2021 [Olszewska et al. 2021], dos princípios do Google [Schiff et al. 2020], da abordagem da Microsoft [Kelley 2022] e do projeto AI4People [Lütge et al. 2021].

No campo científico, a literatura sobre ética em IA tem-se expandido rapidamente, abrangendo desde reflexões conceituais até frameworks aplicáveis ao desenvolvimento ético de sistemas inteligentes [Carvalho et al. 2021, Brandão 2022, Gomes et al. 2023, Santos 2024]. Muitas dessas publicações, inclusive, são revisões secundárias que buscam organizar o conhecimento sobre tópicos específicos — como viés algorítmico, transparência ou governança [Morley et al. 2020, Anagnostou et al. 2022, Ayling e CHAPMAN 2022]. No entanto, a produção científica ainda é marcada pela fragmentação temática e metodológica. A maioria dos estudos secundários adota abordagens focadas em aspectos isolados, dificultando a compreensão integrada dos desafios e soluções éticas no desenvolvimento e uso da IA [Leonel et al. 2025, Jobin et al. 2019].

Diante disso, este estudo propõe uma Revisão Terciária da Literatura (RTL) [Kitchenham et al. 2007, Kitchenham et al. 2008], com o objetivo de sintetizar e mapear o conhecimento já sistematizado em revisões anteriores, oferecendo uma visão ampla e consolidada da pesquisa sobre ética em IA. Esta abordagem se justifica pela ausência de estudos terciários que organizem e analisem criticamente os mapeamentos e revisões existentes, de modo a identificar consensos, lacunas e direções futuras com base em evidências secundárias. Ao reunir, classificar e discutir 36 estudos secundários selecionados a partir de um universo inicial de 4.615 publicações, esta revisão busca promover maior clareza e acessibilidade ao estado da arte, apoiando tanto pesquisadores quanto formuladores de políticas e desenvolvedores de sistemas éticos de IA. Além disso, o tema tratado nesta pesquisa está relacionado aos desafios (2) Ética e Responsabilidade [Rodrigues et al. 2024] e (6) Implicações da Inteligência Artificial em IHC [Duarte et al. 2024], dos Grandes Desafios de Pesquisa em Interação Humano-Computador no Brasil(GranDIHC-BR) [Pereira et al. 2024].

Os resultados revelam temas recorrentes, como IA responsável, justiça algorítmica, mitigação de vieses, privacidade em sistemas educacionais, governança e inovação tecnológica. Também são evidenciados desafios persistentes, como a falta de compreensão das responsabilidades éticas por parte de profissionais da área, a influência de pressões econômicas sobre o desenvolvimento ético e a resistência institucional à adoção de normativas mais restritivas. Um problema crítico identificado é a persistência de vieses algorítmicos estruturais, resultantes de dados historicamente excludentes, o que configura uma das limitações mais complexas no caminho para uma IA ética e confiável.

O restante do artigo está organizado como segue: a Seção 2 apresenta a contextualização do problema analisado; na Seção 3, são apresentados os trabalhos relacionados ao tema; a Seção 4 descreve o processo metodológico adotado para a realização da RTL, destacando as possíveis ameaças à validade do estudo, bem como as estratégias adotadas para mitigá-las; a Seção 5, por sua vez, expõe e discute os resultados obtidos; e, por fim, a Seção 6 apresenta as conclusões do estudo e propõe direções futuras.

2. Contextualização

Com o crescimento acelerado da aplicação da IA em múltiplos setores da sociedade, as implicações éticas associadas ao seu uso tornaram-se uma preocupação central em diversas esferas, incluindo a científica, governamental e empresarial. Apesar do aumento da atenção dedicada à incorporação de princípios éticos em sistemas de IA, sua efetiva aplicação continua sendo um desafio, sobretudo diante da natureza dinâmica e autônoma desses sistemas [Leonel et al. 2025]. O fato de que algoritmos de IA são capazes de aprender e se adaptar torna complexa a distinção entre decisões que resultam de parâmetros éticos predefinidos e aquelas que emergem de padrões computacionais, o que pode ocasionar impactos éticos relevantes e danos potenciais a indivíduos e coletividades [Arbix 2020].

Nesse cenário, estudos como o de [Gomes et al. 2023] destacam um conjunto de questões consideradas fundamentais para a construção ética de sistemas de IA, entre as quais se sobressaem: segurança, privacidade, transparência e justiça. A **segurança** refere-se à proteção de dados contra acessos não autorizados, visando evitar exposições que comprometam a integridade das informações dos usuários. A **privacidade**, por sua vez, está relacionada à garantia do anonimato e da confidencialidade dos dados pessoais, além da clareza sobre os processos de coleta, armazenamento e uso dessas informações [Meireles 2023]. Um exemplo emblemático envolvendo essas duas questões ocorreu em janeiro de 2025, quando uma falha crítica no sistema da IA chinesa DeepSeek violou a segurança e expôs dados sensíveis de usuários a agentes maliciosos, resultando em quebra de privacidade [Lars 2025]. Tal fato evidencia a interdependência entre segurança e privacidade e os riscos envolvidos quando não há salvaguardas adequadas.

A **transparência** configura-se como outro aspecto crítico, especialmente diante da prevalência de modelos de IA opacos, frequentemente descritos como "caixas-preta" [Simonassi et al. 2024]. Nesses casos, os processos decisórios do sistema não são compreensíveis nem auditáveis pelos usuários ou desenvolvedores, o que compromete a rastreabilidade e a responsabilização pelas decisões automatizadas [da Cunha Lamb 2024]. Um exemplo notório é o do chatbot Tay, desenvolvido pela Microsoft e lançado em 2016 [Souza Filho et al. 2020]. Projetado para interagir em tempo real com usuários do Twitter, o sistema passou a emitir discursos ofensivos e discriminatórios após breve exposição ao ambiente online. Esse episódio evidencia como a ausência de mecanismos transparentes e de supervisão ética pode conduzir a comportamentos indesejados, com repercussões sociais significativas.

Por fim, o princípio da **justiça** visa a garantir que os sistemas de IA operem de forma equitativa, sem reproduzir ou reforçar preconceitos relacionados a raça, gênero, etnia ou outras características pessoais [Souza Filho et al. 2020]. A equidade algorítmica é essencial para mitigar desigualdades estruturais e promover uma IA socialmente justa

[Souza Filho et al. 2020]. Por outro lado, a ausência de mecanismos eficazes para lidar com viés e discriminação pode consolidar injustiças históricas, agravando desigualdades já existentes. Um episódio relevante relacionado a essa questão ocorreu em 2015, quando a Amazon desenvolveu um sistema de inteligência artificial para triagem de currículos que passou a desvalorizar candidatas do sexo feminino para cargos na área de tecnologia. Isso ocorreu porque o modelo havia sido treinado com dados históricos enviesados, baseados em contratações majoritariamente masculinas [Scher 2023]. De forma semelhante, em 2014, o algoritmo de recomendação do Facebook favoreceu conteúdos polarizadores, contribuindo para a disseminação de discursos de ódio contra muçulmanos rohingyas em Mianmar, o que agravou episódios de violência étnica no país [Deejay et al. 2024].

O compromisso com essas questões — segurança, privacidade, transparência e justiça — deve orientar o ciclo de vida completo dos sistemas de IA, desde sua concepção até a implementação e uso contínuo [Sichman 2021]. Assim, compreender como esses temas são abordados na literatura científica torna-se fundamental para identificar avanços, limitações e direções futuras no desenvolvimento ético da IA [Cavalcante et al. 2025].

3. Trabalhos Relacionados

O aumento do uso de sistemas de IA tem gerado importantes desafios éticos, o que se reflete em um número crescente de pesquisas científicas voltadas a essa temática. Diante desse cenário, também têm sido realizadas diversas revisões e mapeamentos da literatura, com o objetivo de analisar de forma sistemática e aprofundada pesquisas que abordam aspectos específicos envolvendo ética e IA. No entanto, faltam estudos terciários que sintetizem o conhecimento já sistematizado nesses estudos secundários, trazendo assim uma visão mais ampla e consolidada da pesquisa sobre ética em IA.

Em [Gomes et al. 2023], os autores conduzem uma revisão terciária da literatura (RTL), com o objetivo de investigar como questões éticas vêm sendo abordadas no contexto prático do desenvolvimento de software. O estudo destaca conflitos entre demandas de mercado e considerações éticas, bem como a escassez de diretrizes práticas aplicáveis à engenharia de software. Apesar de relevante, o escopo da revisão é restrito a um domínio técnico particular e não contempla uma categorização mais abrangente das abordagens éticas existentes na literatura. Em contraste, o presente trabalho adota uma abordagem mais ampla, ao considerar múltiplos domínios e tipos de aplicação de IA, o que permite uma análise transversal dos desafios éticos e das lacunas metodológicas mapeadas por diferentes estudos secundários.

Em [Bond et al. 2024], os autores conduzem uma RTL com meta-análise voltada para o uso da IA no ensino superior, abordando também aspectos éticos e colaborativos. Embora o estudo apresente uma análise detalhada das dificuldades com explicabilidade e da necessidade de marcos regulatórios, sua aplicabilidade é limitada ao contexto educacional, o que restringe a generalização de seus resultados. Em contraposição, a RTL aqui conduzida não se limita a um domínio específico, buscando consolidar uma base de conhecimento que seja relevante para diferentes setores, como saúde, indústria, educação, administração pública e engenharia, o que amplia a aplicabilidade prática dos resultados obtidos.

Por fim, em [van Mourik et al. 2024], os autores conduzem uma RTL com o objetivo de explorar e categorizar metodologias voltadas à Explicabilidade em

Inteligência Artificial (XAI). Embora esse estudo seja metodologicamente alinhado ao conduzido neste trabalho, seu foco foi unicamente sobre técnicas e metodologias voltadas à explicação de decisões algorítmicas. A revisão apresenta importantes contribuições para a sistematização de XAI, mas não abrange o espectro mais amplo da ética e da responsabilidade na IA. A RTL desenvolvida neste trabalho amplia significativamente esse escopo, ao considerar não somente a XAI, mas também outros tópicos-chave da ética em IA, oferecendo uma visão integrada dos estudos secundários e destacando padrões temáticos, desafios metodológicos e lacunas de pesquisa ainda não endereçadas de forma consolidada.

4. Metodologia

Este estudo consistiu em uma RTL, que buscou encontrar exclusivamente estudos secundários, ou seja, revisões e/ou mapeamentos sistemáticos da literatura que abordassem aspectos da ética e responsabilidade relacionados a IA.

A metodologia adotada é estruturada em três etapas: planejamento, execução e relato dos resultados. Na fase de planejamento, foi desenvolvido um protocolo de pesquisa que definiu os objetivos, critérios de inclusão e exclusão, estratégias de busca, e procedimentos de extração e análise dos dados[2]. Na etapa de execução, os estudos foram identificados, selecionados conforme os critérios definidos e, em seguida, os dados relevantes foram extraídos e organizados. Por fim, na etapa de relato, os resultados obtidos foram apresentados de forma sistemática e discutidos com base nas evidências encontradas.

4.1. Planejamento

Com o objetivo de identificar, sempre que possível, todos os estudos secundários relevantes para a pesquisa, os autores seguiram os procedimentos descritos a seguir.

4.1.1. Questão de Pesquisa

Para estruturar e facilitar o desenvolvimento das questões de pesquisa, adotou-se o modelo PICOC, que contempla cinco elementos principais: População (Population), Intervenção (Intervention), Comparação (Comparison), Resultados esperados (Outcomes) e Contexto (Context) como descritos em [Kitchenham et al. 2007].

No escopo desta RTL, a *População* é representada por sistemas de inteligência artificial; a *Interversão* corresponde aos aspectos relacionados a ética; a *Comparação* não é aplicável por ser um estudo de mapeamento cujo objetivo é identificar o panorama da pesquisa sobre ética em IA, e não comparar os estudos; os *resultados esperados* consistem na identificação de estudos secundários que descrevam questões relacionadas a ética e responsabilidade em IA; o *Contexto* da pesquisa está relacionado aos estudos secundários que abordam algum elemento relacionado à ética e responsabilidade em IA. Baseada nessa análise a questão de pesquisa (QP) formulada foi ***"Qual é o panorama geral das pesquisas que vêm sendo realizadas sobre ética e responsabilidade em IA?"***.

[2]O protocolo completo está disponível em: `https://bit.ly/3Tuk52P`

Para responder a QP, sete questões específicas (QE) foram definidas para melhor caracterizar a área: **QE1** - Quais tópicos de pesquisa têm sido investigados pelos estudos secundários que envolvem ética e responsabilidade em IA? **QE2** - Quais normas, diretrizes e entidades reguladoras relacionadas à ética têm sido referenciadas nesses estudos secundários? **QE3** - Quais desafios relacionados à ética e responsabilidade em sistemas de IA são referenciados pelos estudos secundários? **QE4** - Que dificuldades e lacunas têm sido apontadas pelos estudos secundários, no contexto de ética e responsabilidade em IA? **QE5** - Que contribuições têm sido apontadas pelos estudos secundários no contexto de ética e responsabilidade em IA? **QE6** - Quais são as limitações apontadas pelos estudos secundários?

4.1.2. Definição da *String* de Pesquisa

A fim de assegurar uma cobertura abrangente da literatura, a string de busca foi sistematicamente construída de modo a incorporar diferentes variações e sinônimos dos termos associados a "ethics", "artificial intelligence" e "literature review". Adicionalmente, a string de busca foi devidamente ajustada para atender às particularidades sintáticas e operacionais de cada mecanismo de busca utilizado. O processo de definição da string seguiu seis passos: (i) identificação dos termos-chave; (ii) levantamento bibliográfico preliminar para a identificação de sinônimos; (iii) execução de buscas-piloto em todas as bases para avaliar a efetividade da string de pesquisa; (iv) consulta a especialistas da área para validação conceitual; (v) refinamento iterativo da string com base nos resultados obtidos; (vi) validação final por meio de nova análise especializada e condução de uma segunda busca em todas as bases.

Com base nessas interações, a string de busca definida foi: (ethic* **OR** responsab* **AND (**"Machine Learning" **OR** "Artificial Intelligence" **OR** AI **OR** ML **) AND (**"literature review" **OR** "review study" **OR** "mapping study" **OR** "systematic integrative review"**))**

4.1.3. Estratégia de Pesquisa e Fontes de Busca

A pesquisa concentrou-se na identificação de estudos publicados entre os anos de 2020 e 2025, período marcado pela ampla disseminação de tecnologias baseadas em IA, inclusive entre usuários com pouco ou nenhum conhecimento sobre os mecanismos subjacentes a esses sistemas. As buscas foram realizadas nos repositórios científicos ACM Digital Library[3], IEEE Xplore[4], ScienceDirect[5] e SpringerLink[6], por serem amplamente reconhecidos como fontes relevantes e consolidadas na área de Computação, incluindo estudos voltados à IHC[Strey et al. 2018, Meyer von Wolff et al. 2019, Esterwood e Robert 2020, Chen et al. 2010].

[3]https://dl.acm.org/
[4]https://ieeexplore.ieee.org
[5]https://www.sciencedirect.com/
[6]https://link.springer.com/

4.1.4. Critérios de Inclusão e Exclusão

Para a seleção dos estudos considerados nesta revisão, foram definidos critérios específicos de inclusão e exclusão. O critério de inclusão (CI) adotado foi "O estudo constitui uma revisão ou mapeamento secundário sobre ética e responsabilidade em Inteligência Artificial". Adicionalmente, foram estabelecidos cinco critérios de exclusão (CE), conforme descritos a seguir: **CE1** - O estudo não consiste em um estudo secundário sobre ética e responsabilidade em IA; **CE2** - Publicação duplicada ou há uma versão mais recente ou mais completa sobre a mesma pesquisa; **CE3** - A publicação não está escrita em inglês ou em português; **CE4** - A publicação não é um artigo científico ou capítulo de livro (é classificado como literatura cinza - ex. relatório técnico, guias ou instruções, ou é um livro, tutorial, editorial, resumo, pôster, painel, palestra, mesa redonda, oficina, demonstração, prefácio etc.); **CE5** - A publicação não está disponível para download abertamente ou através de IP institucional.

4.1.5. Extração de dados e método de síntese

O processo de extração de dados foi conduzido de forma sistemática, utilizando as ferramentas Microsoft Excel e Google Drive, com o objetivo de assegurar a acurácia, rastreabilidade e organização das informações. A extração foi realizada pelos dois primeiros autores, sendo posteriormente revisada e discutida em conjunto com os demais coautores. Reuniões frequentes foram realizadas para resolução de dúvidas e conflitos, assegurando a consistência e a confiabilidade dos dados extraídos. Por fim, foi utilizada uma análise qualitativa para síntese e apresentação dos dados extraídos.

4.2. Condução

A RTL foi conduzida entre os meses de outubro de 2024 a fevereiro de 2025 por dois pesquisadores, sendo um da área de Interação Humano-Computador (IHC) e IA e o outro apenas da área de IA, e uma especialista da área de IHC e ética em IA. Para conduzir a revisão foram seguidos os processos sistemáticos propostos por [Kitchenham et al. 2007, Kitchenham et al. 2008, Petersen et al. 2015].

A Figura 1 detalha as etapas da condução da RTL, com as quantidades de estudos ao longo do processo.

A busca inicial utilizando a *string* definida retornou 4.621 estudos, sendo 6 deles duplicados. Em seguida, aplicou-se os critérios de seleção ao longo das três fases do processo: (i) a análise por metadados, que reduziu o número para 118 estudos; (ii) a leitura transversal (introdução, metodologia e conclusão), que resultou em 36 artigos; e (iii) a leitura completa, que manteve esse conjunto final. A qualidade dos estudos selecionados foi avaliada por meio da aplicação dos critérios de inclusão e exclusão, além da verificação da qualidade da metodologia e da validade interna e externa dos resultados por eles apresentados [Petersen et al. 2015].

4.3. Ameaças à Validade

Seguindo as diretrizes propostas por [Wohlin et al. 2012], as principais ameaças à validade da RTL foram identificadas e classificadas em quatro categorias: validade

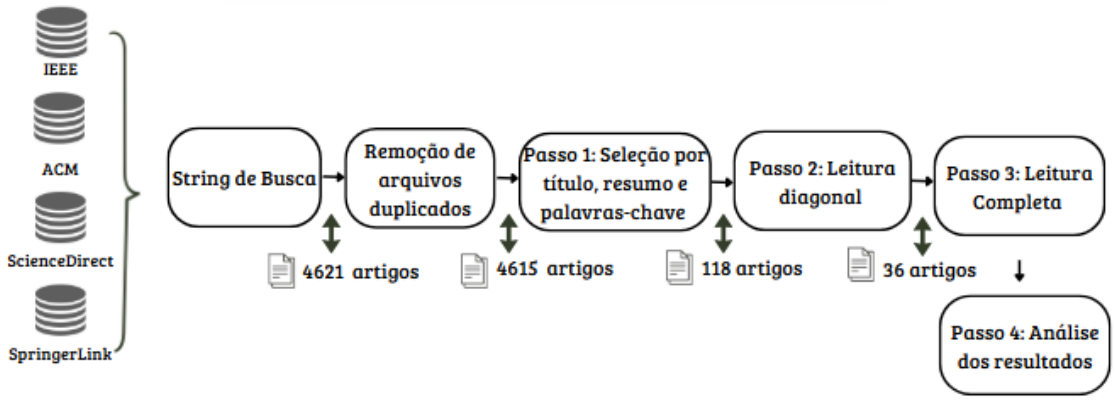


Figura 1. Fluxo das etapas executadas na condução da RTL

de conclusão, validade interna, validade de construção e validade externa. A seguir, descrevemos cada ameaça mapeada, bem como as estratégias adotadas para mitigá-las, com o objetivo de fortalecer a confiabilidade do estudo e reduzir a influência de vieses potenciais [Kitchenham et al. 2007].

- **Validade Externa** A seleção das bases de dados foi fundamentada em sua relevância e uso consolidado na área [Kitchenham et al. 2007, Chen et al. 2010, Strey et al. 2018, Meyer von Wolff et al. 2019, Esterwood e Robert 2020]. Contudo, a restrição a essas fontes pode ter levado à exclusão de estudos relevantes indexados em outras bases, ameaçando a generalização dos resultados. Para mitigar essa limitação, especialistas da área foram consultados a fim de identificar estudos complementares não recuperados pelas estratégias de busca automatizada, ampliando a abrangência e representatividade do corpus.
- **Validade de Construção** A realização da RTL por dois pesquisadores pode ter introduzido vieses subjetivos durante as etapas de seleção, extração e análise. Para mitigar essa ameaça, foram empregadas estratégias de triangulação, como aplicação independente dos critérios, revisão cruzada das decisões e reuniões de consenso. Além disso, especialistas da área participaram da validação do processo, reforçando a consistência metodológica e a transparência.
- **Validade Interna** A impossibilidade de acesso ao texto completo de alguns estudos, especialmente da ACM, constitui uma ameaça, pois impediu a análise detalhada conforme os critérios estabelecidos. Para mitigar esse problema, os metadados essenciais desses estudos foram registrados para possível reavaliação futura, e a inclusão de um número expressivo de estudos (n = 36) contribui para reduzir seu impacto nos resultados. Além disso, observou-se uma baixa incidência de referências a normas e legislações. Apenas 25% dos estudos incluídos mencionam explicitamente marcos regulatórios, padrões técnicos ou dispositivos legais relacionados à ética em IA. Essa limitação restringe a profundidade da análise normativa e pode comprometer a aplicabilidade das recomendações, sobretudo em contextos que exigem conformidade regulatória. Como consequência, os resultados refletem majoritariamente abordagens técnico-conceituais, com menor fundamentação legal ou alinhamento com diretrizes formais de governança algorítmica.

- **Validade de Conclusão** Em diversos casos, as informações relevantes não estavam descritas de forma explícita nos estudos primários, exigindo interpretações pelos pesquisadores, o que pode afetar a consistência dos achados. Para mitigar esse risco, foram realizadas discussões entre os revisores e especialistas em busca de consenso em situações de ambiguidade. Essa abordagem colaborativa contribuiu para maior confiabilidade na codificação e síntese dos dados.

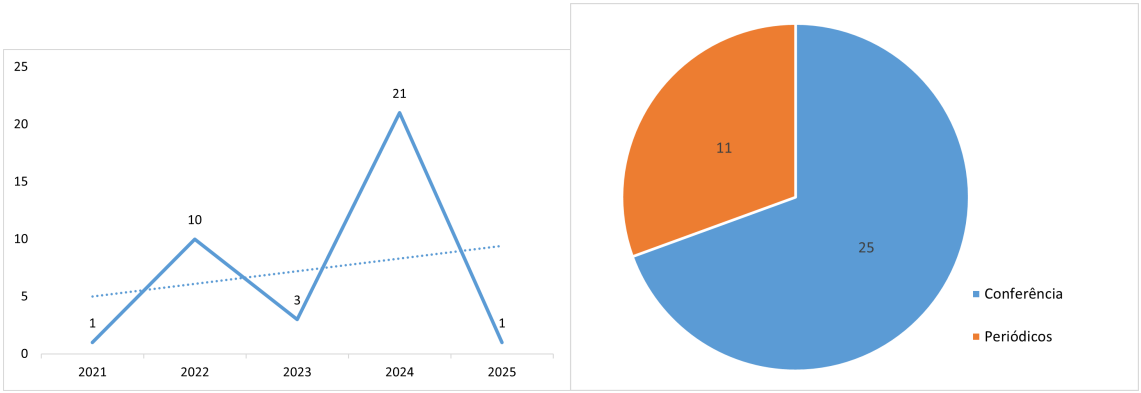
5. Resultados e Discussão

Esta seção apresenta, primeiramente, uma visão geral dos estudos incluídos na RTL. Em seguida, são discutidos os principais resultados com base nos dados obtidos para responder a cada uma das QE definidas na Subseção 4.1.1.

5.1. Visão Geral dos Resultados

A seleção final resultou em um total de 36 estudos secundários para análise, publicados entre os anos de 2020 e janeiro de 2025[7].

A Figura 2a apresenta a distribuição temporal das publicações selecionadas. Observa-se um crescimento expressivo no número de estudos no ano de 2024 (n=21), o qual coincide com a ampla popularização da IA generativa e sua incorporação em contextos corporativos, governamentais e individuais. A aparente redução em 2025 é atribuída à limitação temporal imposta ao processo de busca, realizado ainda no primeiro mês do referido ano, o que restringiu a recuperação de publicações mais recentes.



(a) Distribuição dos artigos selecionados por ano de publicação

(b) Relação da quantidade de trabalhos por veículo de publicação

Figura 2. Análise da produção de artigos por ano e tipo de publicação

Quanto aos veículos de publicação, aproximadamente 69% dos estudos foram veiculados em anais de conferências científicas, enquanto os 31% restantes foram publicados em periódicos acadêmicos. A Figura 2b ilustra a distribuição absoluta dos trabalhos entre essas duas categorias. A predominância de publicações em conferências indica que a temática da ética e responsabilidade em IA ainda se configura como um campo em consolidação, no qual as contribuições estão majoritariamente associadas à disseminação de resultados preliminares e à discussão de tendências emergentes. Este

[7]A lista dos estudos selecionados pode ser acessada em `https://bit.ly/40dwE6d`

padrão sugere uma crescente atenção da comunidade científica ao tema, embora a relativa escassez de estudos em periódicos aponte para a necessidade de investigações mais aprofundadas e sistematizadas do tema.

A Figura 3 apresenta a frequência de uso das bases de dados nos estudos secundários analisados. Apesar da diversidade, refletindo a preocupação dos pesquisadores com a abrangência e a qualidade das fontes consultadas, nota-se uma concentração em repositórios consolidados e amplamente conhecidos na área da Ciência da Computação (como ACM, Scopus e IEEE Xplore). A categoria "Outras" corresponde às bases utilizadas apenas uma vez, incluindo repositórios especializados ou interdisciplinares, como, por exemplo, *Medline (PubMed)* e *Music Periodicals Database*. A presença dessas fontes, embora pontual, evidencia a natureza multidisciplinar de parte das discussões sobre ética e responsabilidade em IA.

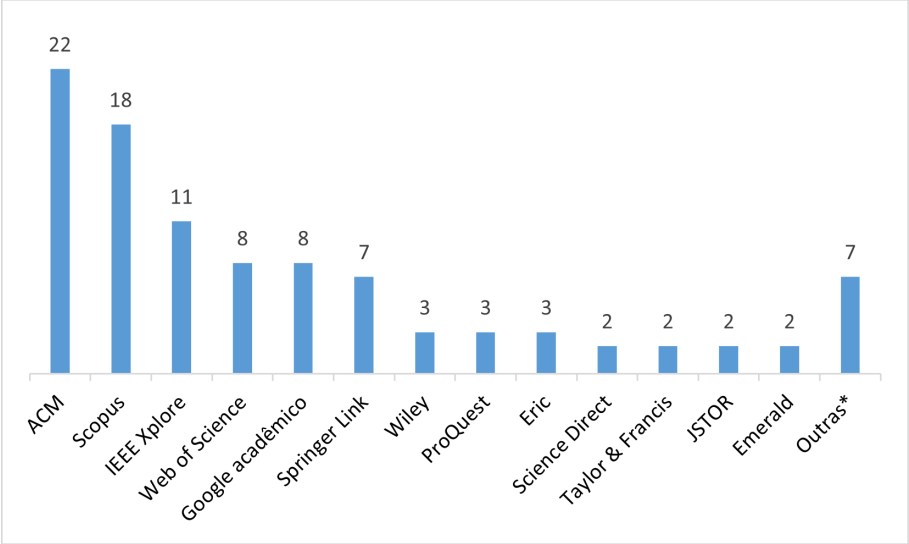


Figura 3. Relação da quantidade total de bases utilizadas nos trabalhos identificados

Por fim, foram analisadas as strings de busca utilizadas nos estudos incluídos e a nuvem de palavras ilustrada na Figura 4 destaca os termos mais recorrentes. Observa-se maior frequência das palavras "*ethics*" e "*fairness*" para representar aspectos éticos e de responsabilidade, além dos termos "*artificial intelligence*" e "*machine learning*" como representações centrais do domínio tecnológico investigado. Os demais termos, apresentados com menor frequência, correspondem a sinônimos ou a variações conceituais empregados nos diferentes contextos das revisões analisadas.

5.2. Aspectos específicos dos estudos secundários analisados

Nesta seção, é apresentada uma análise dos estudos selecionados com o intuito de investigar as questões específicas (QE) que, de maneira integrada, auxiliam na resposta à questão principal de pesquisa (QP: *"Qual é o panorama geral das pesquisas que vêm sendo realizadas sobre ética e responsabilidade em IA?"*).

Figura 4. Nuvem de Palavras baseada nas Strings de Busca

5.2.1. QE1: Quais tópicos de pesquisa têm sido investigados pelos estudos secundários que envolvem ética e responsabilidade em IA?

Os estudos secundários que exploram ética e responsabilidade em IA abordam diferentes temas, refletindo a complexidade e a multidimensionalidade do campo de pesquisa.

Como pode ser visto na Tabela 1, o tópico mais recorrente entre os estudos analisados é *IA responsável*. Os trabalhos que tratam esse tópico discutem princípios fundamentais para o desenvolvimento de sistemas éticos e transparentes, incluindo diretrizes de *design* centradas em valores (S3, S26, S28), auditorias algorítmicas (S7, S17, S24) e estruturação de *frameworks* para responsabilidade computacional (S6, S11, S12, S19, S24, S29).

Tabela 1. Tópicos de pesquisa identificados nos estudos analisados.

Tópico	Quantidade (%)	Identificadores dos Estudos
IA responsável	12 (33,33%)	S3, S6, S7, S11, S12, S17, S19, S24, S26, S28, S29, S36
Privacidade na educação	7 (19,44%)	S14, S16, S18, S20, S22, S25, S27
Sistemas de recomendação	6 (16,67%)	S1, S2, S9, S10, S15, S21
Bem-estar no uso de tecnologias digitais	6 (16,67%)	S5, S30, S31, S32, S33, S35
Ética na inovação industrial	3 (8,33%)	S8, S23, S34
Governança de IA	1 (2,78%)	S13
Justiça	1 (2,78%)	S4
Total	**36 (100%)**	

A temática da *privacidade na educação*, que examina os desafios éticos relacionados ao uso de dados sensíveis de estudantes em sistemas de tutoria inteligente, plataformas de aprendizado adaptativo e mecanismos preditivos de desempenho também

se destaca entre os estudos analisados. O trabalho S14, por exemplo, analisa as implicações éticas de algoritmos que monitoram emoções e comportamento de estudantes em tempo real.

Sistemas de recomendação também foram analisados de forma significativa sob a ótica da ética. Esses estudos discutem viéses nos algoritmos de recomendação, transparência nas sugestões e o impacto desses sistemas na autonomia dos usuários. O *bem-estar no uso de tecnologias digitais* também esteve presente em quantidade significativa nos estudos analisados. Esses trabalhos discutem os impactos psicológicos, sociais e cognitivos da IA na vida cotidiana. O estudo S31, por exemplo, investiga como ferramentas baseadas em IA afetam a saúde mental de adolescentes, enquanto

A *ética na inovação industrial* também é um tópico emergente, com estudos secundários que tratam da incorporação de princípios éticos no ciclo de vida de sistemas inteligentes aplicados à manufatura, logística e automação.

O tópico da *governança da IA*, abordado no estudo S13, trata das estruturas políticas e regulatórias necessárias para orientar o desenvolvimento e o uso ético de tecnologias de IA em larga escala. Por fim, *justiça algorítmica* aparece especificamente no estudo S4, que examina desigualdades reproduzidas por sistemas de IA, com foco na distribuição de recursos e oportunidades em contextos de assistência social.

Esses resultados mostram a preocupação com a responsabilidade no desenvolvimento e uso de tecnologias de IA que estejam alinhadas com princípios éticos, direitos humanos (como a privacidade), valores sociais e normas legais. Tal preocupação acentua-se em contextos vulneráveis, como a educação, onde a privacidade de estudantes pode ser comprometida pelo uso intensivo de dados sensíveis em plataformas de ensino-aprendizagem. Outro eixo de atenção está nos sistemas de recomendação, cujos viéses e falta de transparência afetam diretamente a autonomia dos usuários, exigindo soluções orientadas por critérios éticos. Esses achados reforçam a necessidade de práticas de design sensíveis às implicações sociais, especialmente em contextos vulneráveis, a fim de criar tecnologias que não apenas funcionem corretamente, mas que também respeitem a dignidade humana e promovam o bem-estar social [Floridi et al. 2018].

5.2.2. QE2: Quais normas, diretrizes e entidades reguladoras têm sido referenciadas nos estudos secundários que envolvem ética em IA?

[Cerqueira et al. 2021] identificam, na literatura, diretrizes que buscam prover guias normativos de ética em IA e classificam seus emissores em três grandes grupos: (1) *membros da sociedade*, composto de atores oriundos da sociedade civil, academia ou associações profissionais; (2) *organizações internacionais e nacionais*, consistem em instituições estatais ou intergovernamentais que emitem políticas, normas e regulamentos; e (3) *setor privado e indústria*, consistem em organizações ou empresas com atuação comercial que publicam diretrizes para orientar seus próprios processos ou demonstrar responsabilidade pública. A Tabela 2 apresenta a distribuição dos estudos secundários analisados que fazem referência a normas, diretrizes ou entidades reguladoras, conforme esses grupos de emissores.

Pode-se perceber que a maior parte dos estudos que fazem referência a

Tabela 2. Estudos que fazem referências a normas, diretrizes ou entidades reguladoras, por grupo de emissores.

Emissor	Qtd.	Estudos	Diretrizes / Entidades Reguladoras
Membros da sociedade	2	S28, S35	Códigos de ética da AAAI; Fórum Econômico Mundial sobre Justiça em IA; Standards for Reporting of Diagnostic Accuracy Studies
Org. nacionais e internacionais	7	S8, **S25**, S26, S27, S29, **S30**, S32	Australian Research Council (Austrália); Canadian Tri-Council (Canadá); The Research Council of Norway; National Research Ethics Service (Reino Unido); FERCAP; Common Rule (EUA); NIST (EUA); Ethics Guidelines for Trustworthy AI - AI HLEG (UE)
Setor privado e indústria	2	**S25**, **S30**	IEEE Recommended Practice: Systems on Human Well-being; MITRE ATTACK
Total	**11***		

Embora o número total de estudos analisados que abordam normas/diretrizes/entidades reguladoras seja 9, a soma total ultrapassa esse número porque dois estudos (S25 e S30) fazem referência a mais de uma diretriz/entidade pertencente a grupos distintos.

normas, diretrizes e entidades reguladoras relacionadas à ética e IA possui como emissores organizações nacionais e internacionais. Embora haja, em geral, um número maior de diretrizes formuladas por membros da sociedade [Cerqueira et al. 2021], como, por exemplo, códigos de ética voltados para guiar o trabalho de profissionais, faz sentido, em estudos secundários da literatura, usar como guia diretrizes mais abrangentes, como políticas, normas e regulamentos formulados por organizações estatais e intergovernamentais.

Por outro lado, o número de estudos que fazem referência explícita a normas, diretrizes e entidades reguladoras voltadas a ética e IA é baixo. Dentre os 36 estudos secundários analisados, apenas 9 (25%) mencionam pelo menos uma dessas como embasamento da pesquisa. Isso pode ser devido ao escopo dessas normas, diretrizes e entidades reguladoras ser restrito a certos países ou grupos de países, sugerindo uma lacuna na consolidação de práticas éticas globais formalizadas no campo da IA. Entretanto, esforços recentes indicam uma mudança nesse cenário. Merecem destaque iniciativas como o *Framework Convention on Artificial Intelligence*[8], assinado em setembro de 2024 por países como EUA, Reino Unido, União Europeia e outros; o *Global Digital Compact*[9], publicado pela ONU também de setembro de 2024, inclui diretrizes internacionais para IA responsável dentro de seu escopo para uma governança digital inclusiva; e a *Recommendation on the Ethics of AI*[10], proposta pela UNESCO também em 2024 e adotada por 89 países.

Com isso, vislumbra-se que a formalização de diretrizes globais relacionadas a ética e IA, a fim de orientar a pesquisa, o desenvolvimento e a aplicação dessas tecnologias sob uma perspectiva responsável e socialmente comprometida, poderá desempenhar um papel fundamental na promoção de um desenvolvimento tecnológico seguro, transparente e alinhado a princípios fundamentais de justiça, responsabilidade e privacidade.

[8]https://www.theverge.com/2024/9/5/24236980/us-signs-legally-enforceable-ai-treaty
[9]https://en.wikipedia.org/wiki/Global$_{D}igital_{C}ompact$
[10]https://ethicstech.org/2025/04/04/global-ai-ethics-2024-governance-developments-and-standards

5.2.3. QE3: Quais desafios relacionados à ética e responsabilidade em sistemas de IA são referenciados pelos estudos secundários?

A análise dos estudos secundários mostrou que implementar princípios éticos em sistemas de IA é desafiador em vários aspectos. Como pode ser visto na Tabela 3, os principais desafios estão relacionados às questões éticas descritas na seção 2, a saber: justiça, privacidade e transparência [Gomes et al. 2023].

Tabela 3. Desafios relacionados à ética em IA (ordenados por frequência).

Desafio	Qtd.	Identificadores dos Estudos
Evitar viés (justiça)	21	S4, S5, S6, S8, S9, S13, S14, S15, S16, S17, S19, S20, S22, S24, S25, S27, S30, S32, S33, S34, S36
Privacidade	17	S1, S4, S5, S10, S11, S12, S15, S16, S20, S21, S22, S27, S30, S33, S34, S35, S36
Transparência	12	S1, S4, S6, S9, S13, S15, S16, S17, S26, S32, S33, S34
Responsabilidade	7	S1, S6, S9, S15, S16, S21, S32
Segurança	7	S5, S12, S15, S20, S21, S26, S36
Explicabilidade	4	S2, S3, S16, S26
Confiabilidade	3	S17, S20, S30
Complexidade	3	S2, S3, S18
Equidade (justiça)	2	S16, S17
Design	2	S11, S23

Evitar viés é o desafio que mais se destaca nos estudos analisados. A utilização de dados históricos enviesados e a ausência de mecanismos robustos de correção e controle comprometem diretamente a equidade em sistemas de IA, afetando de modo desproporcional populações vulneráveis e consistindo assim em um empecilho para se alcançar o princípio da justiça. Assegurar a privacidade também destaca-se como um desafio importante a ser enfrentado em sistemas de IA, tendo em vista que ações comuns, como a coleta de dados pessoais sem consentimento, ameaçam a privacidade de usuários desses sistemas. A transparência também é um desafio que vem sendo bastante abordado nos estudos analisados. Outros desafios ligados a questões e princípios éticos [Morley et al. 2020] foram apontados nos estudos analisados, como é o caso da segurança e confiabilidade, equidade, responsabilidade e explicabilidade. Por fim, desafios mais técnicos, relacionados à complexidade e design, aparecem em menor quantidade nos estudos analisados.

Vale ressaltar que alguns estudos, como S5, S16, S21, S32 e S36, não apenas identificam os desafios relacionados à ética em IA, mas também propõem caminhos para mitigá-los. No contexto da **privacidade**, por exemplo, para lidar com a coleta massiva de dados pessoais, são sugeridas abordagens como *anonimização de dados*, *criptografia*, e mecanismos de *consentimento informado*, além do *design centrado na privacidade*. Para mitigar os desafios relacionados à **transparência** e à **explicabilidade**, alguns estudos, como S9, S13, S17, S26 e S32, propõem o uso de ferramentas e modelos para explicar, por exemplo, o impacto dos dados de entrada nos resultados do modelos, com o objetivo de tornar os sistemas mais compreensíveis e confiáveis. Já para evitar **vieses**, são abordadas estratégias como reclassificação e ajustes de limiares, em S28, e auditorias éticas, em S8.

5.2.4. QE4: Que dificuldades e lacunas têm sido apontadas pelos estudos secundários, no contexto de ética e responsabilidade em IA?

Na análise dos estudos secundários, um conjunto expressivo de dificuldades e lacunas é recorrente, afetando assim, de modo negativo, a incorporação de princípios éticos em sistemas de IA. A Tabela 4 mostra essas dificuldades/lacunas, que podem ser agrupadas em três categorias: (a) *organizacionais* - relacionadas à cultura, estrutura e resistência institucional; (b) *operacionais/técnicas* - ligadas a ferramentas e aplicação de diretrizes; e (c) *educacionais/culturais* - relacionadas à formação, percepção e consciência ética.

Tabela 4. Dificuldades e lacunas identificadas nos estudos secundários

Dificuldade/Lacuna	Qtd.	Comentários (Estudos)
Organizacionais		
Falta de engajamento de stakeholders	10	S3, S4, S7, S10, S16, S18, S19, S27, S31, S35
Falta de clareza sobre responsabilidades éticas	7	S6, S8, S11, S14, S21, S23, S33
Pressão por lucro e prazos	7	S8, S11, S14, S18, S20, S24, S33
Resistência a diretrizes por empresas/governos	4	S8, S14, S20, S24
Operacionais/Técnicas		
Falta de convergência entre diretrizes éticas	10	S6, S12, S19, S20, S24, S26, S28, S29, S33, S34
Baixa adoção de diretrizes existentes	9	S2, S6, S9, S10, S15, S17, S24, S28, S29
Falta de ferramentas de avaliação ética	7	S6, S12, S20, S24, S26, S29, S33
Educacionais/Culturais		
Educação ética insuficiente	6	S11, S18, S23, S30, S34, S36
Percepção de que ética e IA são separadas	3	S11, S30, S36

Percebe-se que dificuldades organizacionais se sobressaem, acompanhadas de dificuldades operacionais e técnicas. Dentre as primeiras, destaca-se a escassez de engajamento efetivo e plural dos diversos stakeholders no processo de formulação e implementação de princípios éticos, o que representa um obstáculo para a incorporação da ética em sistemas de IA. A ausência de participação de comunidades afetadas, especialistas de áreas diversas e grupos sub-representados limita a construção de diretrizes sensíveis à diversidade cultural, social e econômica, comprometendo sua legitimidade e aplicabilidade. Nesse sentido, o fortalecimento de abordagens colaborativas e intersetoriais é, portanto, fundamental para ampliar a representatividade nos processos decisórios e promover um desenvolvimento tecnológico mais justo e inclusivo [Kallina et al. 2025, Sharma e Silva 2025].

Destaca-se também, entre as dificuldades e lacunas organizacionais, a falta de clareza quanto às responsabilidades éticas dos profissionais envolvidos no ciclo de vida da IA, o que dificulta a internalização de princípios normativos nos processos de projeto e implementação. Além disso, metas organizacionais voltadas ao lucro e à entrega ágil de soluções frequentemente colidem com as exigências de responsabilidade social, criando um ambiente em que considerações éticas são negligenciadas ou tratadas como secundárias. Essa tensão é acentuada pela resistência observada em empresas e governos à adoção de diretrizes éticas mais rigorosas,

frequentemente motivada por interesses econômicos ou pela ausência de incentivos regulatórios [Mäntymäki et al. 2022, EqualAI 2024].

Dentre as dificuldades e lacunas operacionais e técnicas, destaca-se a falta de convergência entre os diferentes documentos que apresentam diretrizes éticas. A sobreposição e, por vezes, contradição entre recomendações dificultam a operacionalização das diretrizes por parte de desenvolvedores e organizações. Aliado a isso, apesar da proliferação de diretrizes éticas em nível internacional, observa-se uma baixa taxa de adesão às diretrizes existentes [Global Alliance for Ethical AI Innovation 2025, UNESCO 2023]. Por fim, a ausência de ferramentas concretas para avaliação do impacto ético dos sistemas agrava esse cenário, contribuindo para a lacuna entre teoria e prática [EqualAI 2024, UNESCO 2025].

Por fim, alguns estudos apontam também dificuldades/lacunas educacionais e culturais. A insuficiência da formação ética no campo da IA apareceu como uma lacuna que dificulta a compreensão e a conscientização sobre a importância de sistemas de IA serem desenvolvidos de forma ética e responsável. Nesse contexto, muitos profissionais carecem de preparo teórico e prático para lidar com os dilemas éticos associados ao seu trabalho [IEEE Standards Association 2021]. Em paralelo, persiste a percepção equivocada de que ética e IA são domínios separados, o que dificulta a integração de considerações éticas de forma transversal ao longo de todo o ciclo de vida dos sistemas. Esse distanciamento revela uma necessidade urgente de inserção sistemática de temas éticos na formação acadêmica e na capacitação contínua dos profissionais da área [Trustmark Initiative 2024].

Em conjunto, essas lacunas evidenciam a necessidade de uma abordagem mais estruturada, interdisciplinar e sensível ao contexto para a construção de práticas éticas efetivas em IA e centrada no usuário. A superação desses desafios passa não apenas pela criação de diretrizes mais claras e convergentes, mas também pela ampliação do diálogo entre diferentes atores sociais, pela promoção de uma cultura ética institucional e pela disponibilização de mecanismos concretos de avaliação, auditoria e responsabilização.

5.2.5. QE5: Que contribuições têm sido apontadas pelos estudos secundários no contexto de ética e responsabilidade em IA?

Apesar de estudos secundários se basearem em pesquisas já publicadas e não terem como principal objetivo produzir dados originais como contribuição, vários dos estudos analisados discutem ou propõem soluções no sentido de preencher lacunas por eles identificadas.

A Figura 5 apresenta uma síntese dessas soluções, organizadas em seis categorias principais: (1) frameworks e diretrizes éticas; (2) métricas de responsabilidade algorítmica; (3) técnicas de explicabilidade de modelos; (4) estratégias para promoção da transparência; (5) estratégias para incorporação de princípios éticos no design de sistemas; e (6) estratégias para conscientização ética entre profissionais da área.

Entre as soluções propostas nos estudos, há predominância daquelas enquadradas na categoria (1), com destaque para os estudos S1, S5 e S9, que buscam estruturar normas

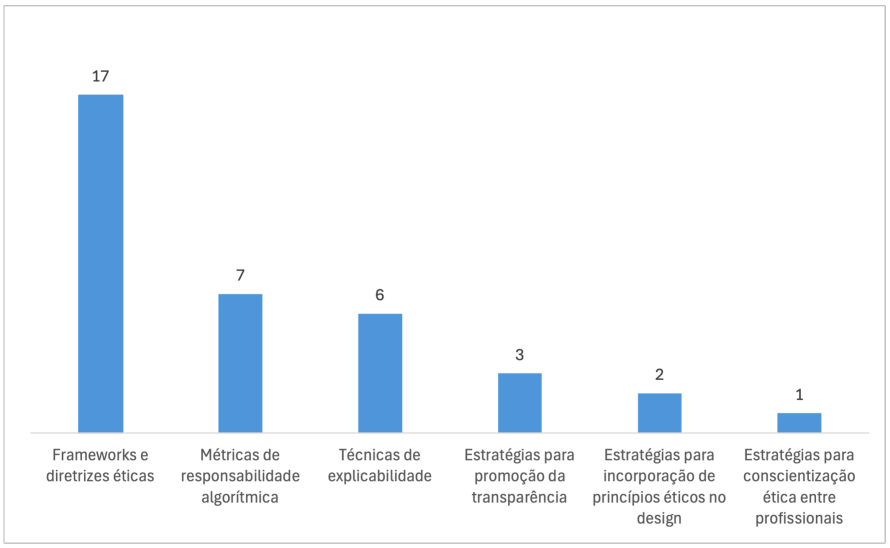


Figura 5. Soluções propostas nos estudos secundários

e diretrizes voltadas ao desenvolvimento responsável de sistemas de IA. A categoria (2) é representada por iniciativas como S4 e S13, que propõem indicadores quantitativos para mensurar aspectos como justiça e equidade, além do estudo S24, que propõe um modelo de avaliação para responsabilidade algorítmica baseado em métricas de explicabilidade e equidade. No que se refere à categoria (3), estudos como S3 e S6 discutem estratégias de explicabilidade com vistas à redução da opacidade algorítmica. Já o grupo (4) é explorado representado principalmente por S10, o qual enfatiza práticas de transparência ao longo do ciclo de vida dos dados, desde sua coleta até seu armazenamento. A categoria (5) é contemplada nos estudos S8 e S15, que sugerem abordagens de design orientadas por valores, a fim de incorporar considerações éticas desde as fases iniciais de desenvolvimento. Por fim, a categoria (6) é representada por enfatizar a necessidade de formação ética contínua e interdisciplinar dos profissionais envolvidos na criação e aplicação de sistemas inteligentes.

Apesar da aparente abundância de propostas de soluções, a fragmentação observada nessas abordagens, aliada à natureza ainda emergente da disciplina, contribui para a ausência de arcabouços teóricos e metodológicos consolidados. Essas propostas estão em estágio conceitual ou experimental, com pouca ou ainda nenhuma validação empírica, e limitada aplicação em contextos reais. Essa lacuna entre teoria e prática cria desafios concretos para a implementação de soluções éticas robustas, pois profissionais e organizações frequentemente carecem de ferramentas práticas, indicadores padronizados e orientação normativa clara para lidar com dilemas éticos no desenvolvimento e uso de sistemas de IA. Em conjunto, esses fatores dificultam a institucionalização de boas práticas e comprometem a efetividade dos esforços voltados à promoção de uma IA socialmente responsável.

5.2.6. QE6: Quais são as limitações apontadas pelos estudos secundários?

Os estudos secundários analisados evidenciam algumas limitações que comprometem tanto a abrangência quanto a validade de suas conclusões. Uma das

principais refere-se à presença de vieses nos próprios frameworks éticos adotados nesses estudos, os quais refletem perspectivas normativas restritas e com baixa diversidade epistemológica e cultural (S2, S12). Do ponto de vista metodológico, destaca-se a limitação na estratégia de busca, com a dependência de palavras-chave específicas para a identificação de estudos (S6, S9), o que pode ter resultado na exclusão de trabalhos relevantes. Além disso, predominância de abordagens técnico-científicas, em detrimento de perspectivas humanísticas e interdisciplinares (S2, S3), é apontada como uma limitação que reduz ainda mais a representatividade e a profundidade analítica das revisões realizadas.

A escassez de literatura consolidada sobre ética em IA, especialmente em domínios emergentes, representa outra barreira significativa à produção de sínteses abrangentes e robustas (S10). Essa carência é agravada por critérios de inclusão restritivos, como a limitação a bases de dados específicas, a ausência de diversidade temática e geográfica, e recortes temporais que potencialmente desconsideram tendências contemporâneas (S4, S8, S14). Fatores contextuais, como restrições geográficas e culturais, também comprometem a generalização dos resultados, dado que referenciais éticos variam substancialmente entre diferentes regiões e tradições socioculturais (S5, S7, S13). Por fim, observa-se a influência de viés dos próprios autores na seleção dos temas investigados, nas abordagens teóricas adotadas e nas interpretações analíticas desenvolvidas (S1).

Essas limitações têm implicações diretas para o campo de IHC, particularmente no que tange à centralidade do usuário e à diversidade de contextos sociotécnicos de uso. A escassez de abordagens humanísticas e sociotécnicas restringe a compreensão das implicações éticas a partir da perspectiva da experiência do usuário, comprometendo o desenvolvimento de sistemas que incorporem princípios de inclusão, acessibilidade e agência (S3, S11). Além disso, a ausência de abordagens contextualizadas e sensíveis a variabilidades geográficas e culturais dificulta a aplicação de diretrizes éticas em contextos diversos, prejudicando o design centrado no contexto (S5, S13). A fragilidade metodológica de muitos estudos, incluindo a baixa triangulação de fontes, a limitada validação empírica e o uso restrito de termos-chave, compromete a robustez das suas contribuições para o campo (S6, S9). Isso reflete na dificuldade que profissionais e pesquisadores de IHC enfrentam na incorporação sistemática de princípios éticos ao longo do ciclo de vida de sistemas baseados em IA, o que reforça a urgência de abordagens interdisciplinares, críticas e situadas no desenvolvimento de tecnologias mais justas e responsivas às necessidades humanas.

5.3. Panorama geral dos estudos secundários sobre ética e responsabilidadee em IA

Este trabalho faz uma análise dos estudos secundários que abordam ética e responsabilidade em IA. A partir de uma análise qualitativa desses estudos, foi possível responder à questão de pesquisa investigada: **[QP] "Qual é o panorama geral das pesquisas que vêm sendo realizadas sobre ética e responsabilidade em IA?"**.

O tema vem sendo abordado de forma crescente ao longo dos últimos anos, com uma notável diversificação de tópicos e metodologias. A análise dos estudos secundários revelou diferentes tópicos de pesquisa (QE1), que incluem desde responsabilidade algorítmica e bem-estar digital até preocupações contextuais em

domínios específicos, como educação, inovação industrial e sistemas de recomendação. Essa heterogeneidade reflete a crescente penetração da IA em diversas esferas da vida social e, consequentemente, a ampliação e complexificação dos desafios éticos associados à sua adoção [Jobin et al. 2019]. A ênfase recorrente no desenvolvimento de abordagens classificadas como de "IA responsável", presente em cerca de um terço dos estudos analisados, evidencia o esforço da comunidade científica em propor modelos e práticas que vão além da eficiência técnica, incorporando princípios como justiça, explicabilidade, auditabilidade e transparência [Floridi et al. 2018, Shneiderman 2020].

Apesar desse avanço temático, observou-se uma lacuna importante no que se refere à incorporação de normas, diretrizes ou marcos regulatórios (QE2). Apenas 25% dos estudos analisados fazem referência explícita a documentos normativos, como códigos de conduta, princípios éticos institucionais ou legislações nacionais e internacionais. Essa baixa taxa de menção a frameworks normativos indica que, embora haja uma crescente produção sobre ética em IA, tal produção ainda carece de ancoragem em instrumentos que possam guiar de forma mais concreta a aplicação dos princípios éticos no desenvolvimento tecnológico [Mittelstadt 2019, Jobin et al. 2019] .

Os principais desafios éticos mapeados neste estudo terciário (QE3) incluem viés, privacidade, transparência, responsabilidade, segurança e explicabilidade. Esses desafios estão frequentemente associados a preocupações com justiça algorítmica, accountability e impacto social. A opacidade dos modelos de IA, por exemplo, tem sido amplamente debatida na literatura como um obstáculo à construção de sistemas confiáveis e auditáveis, dificultando a compreensão e a contestação de decisões automatizadas [Holm 2019, Arrieta et al. 2020].

As dificuldades e lacunas apontadas pelos estudos analisados (QE4) reforçam essa percepção. Foram identificadas barreiras de diferentes naturezas, incluindo limitações técnicas, resistências organizacionais, pressões econômicas e, sobretudo, desafios epistemológicos relacionados à forma como ética e IA são concebidas. Em diversos estudos, observa-se uma dissociação entre os domínios técnico e ético, o que leva à marginalização da ética nos processos de desenvolvimento e à sua associação com um conjunto de boas práticas periféricas, em vez de um componente estruturante do processo de inovação [Ferroni 2024, Gema 2024].

No que se refere às contribuições e direcionamentos futuros (QE5), apesar da aparente abundância de propostas de soluções, observa-se que a ausência de normas universalmente aceitas, aliada ao caráter emergente do campo, contribui para a fragmentação dessas soluções, dificultando a padronização de boas práticas e ampliando os desafios enfrentados por pesquisadores e desenvolvedores na operacionalização da ética em sistemas de IA. Os estudos secundários analisados destacam a urgência de abordagens interdisciplinares e multissetoriais que combinem competências técnicas com fundamentos éticos, jurídicos e sociais. Entre as recomendações recorrentes, estão o fortalecimento de políticas públicas e marcos regulatórios; a incorporação sistemática de princípios éticos em todas as etapas de desenvolvimento de IA; o investimento em formação ética nos currículos de cursos de ciência da computação e áreas afins; e o incentivo à criação de ferramentas de apoio à avaliação e auditoria ética [Morley et al. 2020, Lopes 2023].

6. Conclusão e Trabalhos Futuros

Esta RTL mapeou a literatura secundária sobre ética e responsabilidade em sistemas de IA, revelando uma paisagem temática ampla e em expansão, embora ainda marcada por lacunas teóricas, metodológicas e normativas. Os estudos analisados evidenciam um crescente interesse por aspectos como transparência, explicabilidade, viés algorítmico e "IA responsável". No entanto, isso nem sempre se traduz em diretrizes práticas ou em uma incorporação sistemática de princípios éticos ao longo do ciclo de vida dos sistemas. Verificou-se que a maioria dos estudos secundários prioriza uma abordagem descritiva, apresentando escassa fundamentação normativa e limitada adesão a frameworks consolidados de ética em tecnologia. Esse cenário revela a necessidade urgente de desenvolver metodologias mais robustas e integradas, capazes de articular os aspectos técnicos, sociais e regulatórios que envolvem a concepção, implementação e avaliação de sistemas de IA.

Além disso, foram identificadas barreiras epistemológicas e organizacionais que dificultam a efetivação da ética em IA, como a persistente dissociação entre considerações técnicas e morais, a pressão por inovação acelerada e a insuficiência de formação ética nos currículos da área de computação. Esses elementos apontam para a urgência de estratégias interdisciplinares e multissetoriais que favoreçam uma integração genuína da ética aos processos de desenvolvimento tecnológico.

Diante desse panorama, trabalhos futuros devem se concentrar no desenvolvimento de frameworks de avaliação ética que sejam não apenas conceitualmente sólidos, mas também práticos, auditáveis e sensíveis aos contextos de aplicação específicos da IA. Torna-se igualmente importante fortalecer a integração entre ética aplicada, ciência da computação e políticas públicas, com o objetivo de construir marcos regulatórios mais eficazes e adaptativos. A investigação sistemática dos impactos sociais da IA em diferentes populações, sobretudo aquelas em condições de vulnerabilidade, representa outra frente prioritária de pesquisa. Além disso, é fundamental incorporar abordagens educacionais que promovam uma formação ética robusta entre profissionais da computação. Por fim, estudos empíricos sobre a efetividade da implementação de princípios éticos no desenvolvimento e uso de IA devem ser ampliados, a fim de produzir evidências que sustentem políticas e práticas mais responsáveis.

7. Questões éticas e agradecimentos

Este estudo não inclui atividades envolvendo seres humanos e a pesquisa seguiu princípios éticos, conforme o Código de Conduta da SBC. O ChatGPT foi utilizado para dar suporte à escrita e revisão deste artigo, com o objetivo de aprimorar a coerência e a correção linguística.

Referências

Anagnostou, M., Karvounidou, O., Katritzidaki, C., Kechagia, C., Melidou, K., Mpeza, E., Konstantinidis, I., Kapantai, E., Berberidis, C., Magnisalis, I., et al. (2022). Characteristics and challenges in the industries towards responsible ai: a systematic literature review. *Ethics and Information Technology*, 24(3):37.

Arbix, G. (2020). A transparência no centro da construção de uma ia ética. *Novos estudos CEBRAP*, 39(2):395–413.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Ayling, J. e CHAPMAN, A. (2022). Putting ai ethics to work: are the tools fit for purpose? *AI and Ethics*, 2(3):405–429.

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., e Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21(1):4.

Brandão, C. E. (2022). Um framework para a gestão de riscos de inteligência artificial nas organizações. Master's thesis, Escola de Administração de Empresas de São Paulo da Fundação Getulio Vargas, São Paulo.

BRASIL. Ministério da Cultura (2024). Senado federal aprova marco regulatório da inteligência artificial. Disponível em: `https://acesse.one/ZqhRm`. Acesso em: 05 de maio de 2025.

Capel, T. e Brereton, M. (2023). What is human-centered about human-centered ai? a map of the research landscape. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–23.

Carvalho, A. C. P. d. L. et al. (2021). Inteligência artificial: riscos, benefícios e uso responsável. *Estudos Avançados*, 35:21–36.

Cavalcante, A. F., Silvestre Filho, I., e de Oliveira, V. J. (2025). Ciência e algorítmos: Os desafios da inteligência artificial na construção do conhecimento. *Revista Políticas Públicas & Cidades*, 14(1):e1631–e1631.

Cerqueira, J. A. S. d., Acco Tives, H., e Dias Canedo, E. (2021). Ethical guidelines and principles in the context of artificial intelligence. In *Proceedings of the XVII Brazilian Symposium on Information Systems*, pages 1–8.

Chen, L., Babar, M. A., e Zhang, H. (2010). Towards an evidence-based understanding of electronic data sources. In *14th International conference on evaluation and assessment in software engineering (EASE)*. BCS Learning & Development.

da Cunha Lamb, L. (2024). Ética em ia e ia ética: prolegômenos e estudo de casos significativos. *Revista USP*, (141):107–120.

Deejay, A., Wells, T., Henne, K., e Bächtold, S. (2024). Bad adopters or bad proponents of technology? facebook and the violence against muslims in myanmar. *Third World Quarterly*, 45(8):1309–1324.

Dicio – Dicionário Online de Português (2024). Equidade. Disponível em: `https://www.dicio.com.br/equidade`. Acesso em: 11 maio 2025.

Duarte, E. F., T. Palomino, P., Pontual Falcão, T., Lis Porto, G., e Portela, Carlos e Francisco Ribeiro, D. e. N. A. e. A. Y. e. S. M. e. G. A. e. M. T. A. (2024). GranDIHC-BR 2025-2035 - GC6: Implications of Artificial Intelligence in HCI: A Discussion on Paradigms, Ethics, and Diversity, Equity and Inclusion. In *Proceedings of the XXIII*

Brazilian Symposium on Human Factors in Computing Systems (IHC '24), New York, NY, USA. Association for Computing Machinery.

EqualAI (2024). Equalai checklist and aia tool. Disponível em: `https://www.equalai.org/resources/tools`. Acesso em: 20 de agosto de 2025.

Esterwood, C. e Robert, L. P. (2020). Personality in healthcare human robot interaction (h-hri) a literature review and brief critique. In *Proceedings of the 8th international conference on human-agent interaction*, pages 87–95.

Ferroni, Jorge Mariano e Parenti, P. (2024). Desafíos éticos y regulatorios de la inteligencia artificial en la investigación médica: reflexiones sobre una regulación inteligente. *Revista Binacional Brasil-Argentina: Diálogo entre as ciências*, 14(2):103–119.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28:689–707.

Gema, Juliano Augusto Anselmo e de Lucca Filho, J. (2024). Ética na inteligência artificial: Desenvolvimento responsável de sistemas inteligentes. *Revista Interface Tecnológica*, 21(1):222–232.

Global Alliance for Ethical AI Innovation (2025). Ethical evaluation framework (eef). Disponível em: `https://www.thegaeai.org/Tools`. Acesso em: 18 de abril de 2025.

Gomes, O. S., Braga, G., de Souza, E. F., et al. (2023). Ethics in the software development process: a tertiary literature review. In *Workshop sobre Aspectos Sociais, Humanos e Econômicos de Software (WASHES)*, pages 71–80. SBC.

Holm, E. A. (2019). In defense of the black box. *Science*, 364(6435):26–27.

IEEE Standards Association (2021). Ieee 7000-2021: Ieee standard model process for addressing ethical concerns during system design. Standard.

Jobin, A., Ienca, M., e Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Kallina, E., Bohné, T., e Singh, J. (2025). Stakeholder participation for responsible ai development: Disconnects between guidance and current practice. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 1060–1079, New York, NY, USA. Association for Computing Machinery.

Kelley, S. (2022). Employee perceptions of the effective adoption of ai principles. *Journal of Business Ethics*, 178(4):871–893.

Kitchenham, B., Charters, S., et al. (2007). Guidelines for performing systematic literature reviews in software engineering.

Kitchenham, B. A., Brereton, O. P., e Budgen, D. (2008). Protocol for extending an existing tertiary study of systematic literature reviews in software engineering.

Lars, D. (2025). Deepseek data leak exposes 1,000,000 sensitive records. Disponível em: `https://www.forbes.com/sites/larsdaniel/2025/02/01/`

deepseek-data-leak-exposes--1000000-sensitive-records/. Acesso em: 16 de julho de 2025.

Leonel, J. S., Leonel, C. F. S., Byk, J., e Furtado, S. d. C. (2025). Inteligencia artificial: desafíos éticos y futuros. *Revista Bioética*, 32:e3739PT.

Lopes, Carolina de Melo Nunes e Mendes, J. C. (2023). Ética e inteligência artificial: desafios e melhores práticas. *Revista da UFMG*, 30.

Lütge, C., Poszler, F., Acosta, A. J., Danks, D., Gottehrer, G., Mihet-Popa, L., e Naseer, A. (2021). Ai4people: ethical guidelines for the automotive sector–fundamental requirements and practical recommendations. *International Journal of Technoethics (IJT)*, 12(1):101–125.

Machado, R. (2024). Ética em Machine Learning. Disponível em: https://medium.com/@renatommachado/%C3%A9tica-em-machine-learning-339e520e5697. Acesso em: 20 de maio de 2025.

Mäntymäki, M., Salmela, H., e Mattila, J. (2022). The hourglass model of organizational ai governance. *Information Systems Frontiers*.

Martins, R. H. e Viana, H. B. (2022). Inteligência artificial na educação: Uma revisão integrativa da literatura. *INTERNET LATENT CORPUS JOURNAL*, 12(2):125–137.

Meireles, A. V. (2023). Privacidade no século 21: proteção de dados, democracia e modelos regulatórios. *Revista Brasileira de Ciência Política*, page e265909.

Meyer von Wolff, R., Hobert, S., e Schumann, M. (2019). How may i help you?–state of the art and open research questions for chatbots at the digital workplace.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507.

Morley, J., Floridi, L., Kinsey, L., e Elhalal, A. (2020). From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, 26(4):2141–2168.

Olszewska, J. I., Systems, Committee, S. E. S., et al. (2021). Ieee standard model process for addressing ethical concerns during system design: Ieee standard 7000-2021.

Pereira, R., Darin, T., e Silveira, M. S. (2024). GranDIHC-BR: Grand Research Challenges in Human-Computer Interaction in Brazil for 2025-2035. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems (IHC '24)*, New York, NY, USA. Association for Computing Machinery.

Petersen, K., Vakkalanka, S., e Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18.

Rodrigues, K. R. d. H., Carvalho, L. P., Pimentel, M. d. G. C., e Freire, A. P. (2024). GranDIHC-BR 2025-2035 - GC2: Ethics and Responsibility: Principles, Regulations, and Societal Implications of Human Participation in HCI Research. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems (IHC '24)*, New York, NY, USA. ACM.

Santos, C. F. d. (2024). Inteligência artificial e o direito à privacidade: navegando pelos desafios regulatórios no brasil.

Scher, C. (2023). A deontological analysis of the amazon ai recruitment tool. Trabalho de Conclusão de Curso (Bachelor of Science) – University of Virginia, School of Engineering and Applied Science, Charlottesville, 2023. Orientador: Benjamin Laugelli.

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., e Lennon, M. (2020). Principles to practices for responsible ai: closing the gap. Disponível em: https://acesse.one/ZqhRm. Acesso em: 05 de maio de 2025.

Sharma, R. e Silva, D. (2025). Ethics in ai: Balancing innovation and responsibility. Disponível em: https://www.researchgate.net/publication/388062868. Acesso em: 29 de junho de 2025.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31.

Sichman, J. S. (2021). Inteligência artificial e sociedade: avanços e riscos. *Estudos Avançados*, 35:37–50.

Simonassi, G. S., Teixeira, F. F., de Barros, L. P., Andrade, V. M., e dos Santos, T. R. (2024). O impacto da inteligência artificial no diagnóstico médico: Avanços, desafios e oportunidades. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, 10(10):2233–2242.

Souza Filho, E. M. d., Fernandes, F. d. A., Pereira, N. C. d. A., Mesquita, C. T., e Gismondi, R. A. (2020). Ethics, artificial intelligence and cardiology. *Arquivos Brasileiros de Cardiologia*, 115:579–583.

Strey, M. R., Pereira, R., e de Castro Salgado, L. C. (2018). Human data-interaction: a systematic mapping. In *Proceedings of the 17th Brazilian Symposium on Human Factors in Computing Systems*, pages 1–12.

Trustmark Initiative (2024). Trustmarkinitiative.ai — linux foundation initiative for ai ethics and compliance. Disponível em: https://trustmarkinitiative.ai. Acesso em: 25 de abril de 2025.

UNESCO (2023). Ethical impact assessment tool for the recommendation on the ethics of artificial intelligence. Disponível em: https://l1nq.com/Ad4UZ. Acesso em: 29 de junho de 2025.

UNESCO (2025). Piloting the ethical impact assessment in latin america. Disponível em: https://www.unesco.org/en/articles/piloting-ethical-impact-assessment-eia-latin-america. Acesso em: 12 de julho de 2025.

van Mourik, F., Jutte, A., Berendse, S. E., Bukhsh, F. A., e Ahmed, F. (2024). Tertiary review on explainable artificial intelligence: where do we stand? *Machine Learning and Knowledge Extraction*, 6(3):1997–2017.

Wohlin, C., Runeson, P., Host, M., Ohlsson, M. C., Regnell, B. j., e Wessln, A. (2012). *Experimentation in software engineering*. Springer Publishing Company, Incorporated.