Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Do Abstrato ao Concreto: Uma Ferramenta para Implementação de Requisitos Éticos em Inteligência Artificial através de Histórias de Usuário

João Gabriel Rossi de Borba

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientadora
Prof.a Dr.a Edna Dias Canedo

Brasília
2025

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Do Abstrato ao Concreto: Uma Ferramenta para Implementação de Requisitos Éticos em Inteligência Artificial através de Histórias de Usuário

João Gabriel Rossi de Borba

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof.a Dr.a Edna Dias Canedo (Orientadora)
University of Brasília (UnB)

Prof.a Dr.a Sheila dos Santos Reinehr
Pontifícia Universidade Católica do Paraná,
(PUC-PR)

Prof. Dr. Geraldo Pereira Rocha Filho
Universidade do Sudoeste da Bahia

Prof. Dr. Rodrigo Bonifácio de Almeida
Coordenador do Programa de Pós-graduação em Informática

Brasília, 28 de Janeiro de 2025

Dedicatória

Dedico esta tese a todos que foram impedidos de percorrer o caminho do conhecimento devido as mazelas da vida. Que no futuro as portas da oportunidade sejam abertas para todos.

"E eu acredito que um navio perdido,
conduzido por navegantes cansados e mareados,
ainda pode ser guiado para atracar
em casa."
- Assata Shakur

Agradecimentos

Gostaria de iniciar os meus agradecimentos expressando imensa gratidão a todos que estiveram ao meu lado durante este caminho. Cada etapa desta dissertação foi sustentada pelo apoio e encorajamento daqueles que trago no coração, e é com a mais profunda gratidão que dedico este espaço ao reconhecimento de todos os que estiveram ao meu lado nesta jornada.

À Deus, que antes que eu possa ter imaginado, preparou este caminho para mim.

À minha maravilhosa companheira de vida, Maria Cecília, que me acompanha nesta jornada chamada vida. Sei que ela estará comigo até ao fim.

À minha família, Neysa, Ênio, Dejanira, Darcy, Paulo, Rodrigo, Ricardo, Ednilson, Emily e Cecilia, que guiaram meus passos desde a minha infância e hoje caminham ao meu lado. À minha tia Cleide e avó Onélia, que me ensinaram a ser quem eu tenho orgulho de ser hoje, suas memórias sempre estarão comigo.

À minha orientadora, Dra. Edna Dias Canedo e minha co-orientadora Dra. Fabiana Freitas Mendes, cujo apoio durante todo este período tem sido vital para mim. Sem a disponibilidade e disposição nada disso teria sido possível.

Aos meus professores e colegas do PPGI, que contribuíram nesta minha jornada para conhecer um pouco mais a cada dia.

Por último, mas não menos importante, gostaria de agradecer a existência das universidades públicas do Brasil, que mesmo depois de tantos ataques, permitem o acesso indiscriminado ao conhecimento, apoiando aqueles que procuram percorrer o caminho que eu também decidi percorrer.

Resumo

Contexto: Nos últimos anos, o campo da inteligência artificial passou por um processo de expansão notável, tanto na academia como na indústria. Este crescimento pode ser observado de várias formas, incluindo o desenvolvimento de tecnologias mais complexas, o aumento do investimento, uma maior atenção dos meios de comunicação social e a expansão das suas áreas de aplicação. No entanto, este avanço deu origem a questões éticas que estão sendo uma questão de crescente preocupação social, como o enviesamento de sistemas, aplicações danosas, entre outros. **Objetivo:** Este trabalho tem como objetivo identificar uma visão geral do estado atual das soluções práticas para o desenvolvimento ético de sistemas baseados em IA em todas as fases do ciclo de vida de um sistema de software, e com isso, desenvolver uma ferramenta que operacionalize a tradução de requisitos éticos de alto nível em histórias de usuário éticas. **Método:** Para atingir este objetivo, este estudo baseou-se na metodologia *Design Science Research*, que produziu três resultados até ao momento. Inicialmente, na fase de consciência do problema, foi atualizada a revisão sistemática da literatura desenvolvida por Cerqueira. Posteriormente, na fase de sugestão, a ferramenta proposta foi formulada como um sistema que traduz requisitos éticos em histórias de usuário éticas. Finalmente, na fase de desenvolvimento, foi implementada a primeira versão da ferramenta, juntamente com a sua validação. **Resultados:** No total foram identificados 38 estudos primários. Dentre estes estudos, a maior parte (63%) propõe uma solução prática para facilitar a aplicação de ética em IA, mas ainda existe uma lacuna entre teoria e prática no que se diz sobre ética em IA. Além disso, foi também compilada uma lista com 26 dos principais princípios éticos que foram discutidos na literatura. Foi também desenvolvida a ferramenta *Requirements to US*, que utiliza histórias de usuário, princípios éticos em IA e modelos de LLMs para promover a integração da ética em IA durante a fase de Engenharia de Requisitos. **Conclusão:** Os resultados do processo de treino, teste e validação do sistema demonstram o seu potencial como uma ferramenta valiosa para a integração de ética em IA no desenvolvimento de software. Os resultados indicam que a ferramenta apoia a integração de aspectos teóricos e práticos, oferecendo uma solução promissora para uma lacuna existente na atual oferta de aplicações práticas.

Palavras-chave: Ética, Ética em Inteligência Artificial, Engenharia de Requisitos, Requisitos Éticos, Aprendizagem de Máquina, Desenvolvimento de Software.

Abstract

Context: In recent years, the field of artificial intelligence has undergone a remarkable process of expansion, both at the academic and industrial levels. This growth can be observed in various ways, including the development of more complex technologies, increased investment, greater media attention and the expansion of its areas of application. However, this advancement has also given rise to a number of ethical issues that are becoming a matter of growing social concern, including system bias and the potential for harmful applications. **Objective:** This work aims to develop a tool that operationalises the translation of high-level ethical requirements into ethical user stories. **Methods:** In order to achieve this objective, this study was based on the Design Science Research methodology, which has yielded four results. Initially, in the problem awareness phase, a systematic literature review was updated. Subsequently, in the suggestion phase, the proposed tool was formulated as a system that transforms ethical requirements into ethical user stories. Following the second phase, the development stage, the first version of the tool was implemented. Finally, the tool was subjected to evaluation in order to generate empirically based evidence of whether it is fit for purpose. **Results:** A total of 38 primary studies were identified. The majority of these studies (63%) put forward a practical solution to facilitate the application of ethics in AI. Nevertheless, a discrepancy persists between theoretical and practical applications of ethics in AI. Additionally, a list of 26 frequently discussed ethical principles in the literature was compiled. Subsequently, the Requirements to US tool was proposed, developed, and validated. This tool employs ethical user stories, AI ethics principles, and Large Language Models to facilitate the integration of AI ethics during the requirements engineering phase. The validation process yielded positive outcomes, confirming the effectiveness of the tool. **Conclusion:** The outcomes of the model's training process demonstrate its potential as a valuable tool for the early integration of AI ethics into software development. The results indicate that the tool supports the integration of theoretical and practical aspects, offering a promising solution to an existing gap in the current range of practical applications.

Keywords: Ethics, Ethics in Artificial Intelligence, Requirements Engineering, Ethical Requirements, Machine Learning, Software Development.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

DSR *Design Science Research*.

IA Inteligência Artificial.

LLM *Large Language Model*.

RSL Revisão Sistemática de Literatura.

Capítulo 1

Introdução

A área de Inteligência Artificial (IA), seja no contexto de pesquisa ou na indústria, registrou uma expansão significativa nos últimos anos. Este crescimento pode ser observado de várias maneiras, incluindo o desenvolvimento de tecnologias avançadas, o investimento, a atenção dos meios de comunicação social e as suas áreas de aplicação, como a saúde, setor bancário, sistemas de vigilância e vários outros [1, 9, 10]. Porém, esta rápida expansão alertou a indústria e academia para as implicações éticas da utilização de tais sistemas [1].

A popularização de sistemas baseados em IA trouxe consigo uma série de incidentes de grande visibilidade. Um número crescente de pesquisadores, meios de comunicação, bem como de governos, vem pedido o desenvolvimento de sistemas de IA mais éticos devido ao aumento destes casos de fracasso [11]. Entre este conjunto de falhas encontra-se o sistema COMPAS, um software baseado em IA projetado para a avaliação da probabilidade de reincidência de um réu criminal. Angwin et al. [12] relataram que este sistema apresenta preconceitos raciais contra negros e outras minorias. Outro caso relevante é o da IA de recrutamento de funcionários da Amazon, que demonstrou preconceito contra candidatas mulheres [11]. A aprendizagem da IA foi moldada pelo fato de ter observado predominantemente contratações de homens devido ao uso de dados de recrutamento do passado, o que a levou a inferir que os candidatos homens eram os mais adequados para a contratação [11]. Ao citar questões relacionadas a privacidade, o uso de IA para casos de personificação são os mais citados [13], seguido por fraudes e *sockpuppeting*.

Todas essas questões estão relacionadas às consequências éticas de sistemas de IA específicos, especialmente os de *deepfakes*. *Deepfakes* são vídeos e arquivos de áudio que foram sintetizados e gerados por IA, nos quais pessoas reais são representadas em situações que nunca ocorreram [14]. A disseminação de tais vídeos tem o potencial de causar danos à reputação e até mesmo danos legais, principalmente quando usados de forma maliciosa. Além disso, podem diminuir a confiança do público no contexto digital, contribuindo para

uma ampla desconfiança em relação à mídia. Isso, por sua vez, pode dar origem a maiores preocupações com segurança e privacidade.

Todos estes problemas podem ser ligados a falta de operacionalização dos princípios éticos de IA. Os possíveis problemas com o sistema COMPAS e o sistema de recrutamento da Amazon poderiam ter sido evitados se os princípios éticos, como justiça, equidade, remediação, reparação, não maleficência, responsabilidade e beneficência, tivessem sido levados em consideração durante o processo de desenvolvimento. Da mesma forma, os desafios associados aos sistemas de *deepfake* poderiam ter sido mitigados pela incorporação de princípios éticos, como reversibilidade, remediação, contestação, segurança, proteção, não maleficência, responsabilidade, privacidade, consentimento, confiabilidade, entre outros. Portanto, é fundamental que o desenvolvimento de sistemas de IA seja orientado por uma estrutura ética robusta que não apenas antecipe os possíveis impactos negativos, mas também forneça estratégias para sua prevenção e mitigação. A incorporação desses princípios em um estágio inicial do processo pode contribuir para a criação de sistemas mais justos, seguros e confiáveis.

Com isso, a fim de prevenir novos incidentes e garantir a prevenção de potenciais problemas no futuro, diversas diretrizes e princípios éticos para o desenvolvimento de IA foram propostas nos últimos anos [15]. Mas, mesmo com diversas pesquisas na área, ainda há falta de consenso quanto à definição do que é uma "IA ética" e aos requisitos éticos, normas técnicas e melhores práticas necessárias para a sua implementação [10]. De acordo com Cerqueira [16], outra dificuldade encontrada é que as propostas não satisfazem os requisitos do desenvolvimento de sistemas éticos baseados em IA no mundo real, visto que estes princípios éticos são, em sua maioria, abstratos e gerais, não constituindo provas reais de que possam influenciar a tomada de decisões éticas.

A análise feita por Corrêa et al. [9] teve como resultado a identificação de que "existe uma necessidade de regulamentação e que um dos maiores desafios que enfrentamos atualmente neste domínio é o fato de os princípios éticos não poderem ser universalizados, tornando a normalização dos parâmetros éticos contextuais um verdadeiro desafio na procura de regulamentação". Organizações privadas e governamentais buscam solucionar os problemas éticos advindos de sistemas de IA. Assim, as leis e regulações ainda se baseiam em princípios, não cumprindo de maneira correta seu papel de restringir tais sistemas [9].

Outra dificuldade pode ser atribuída ao fato de que uma parte considerável das diretrizes existentes permanece em um nível teórico, não oferecendo soluções práticas que possam ser inseridas no ciclo de vida do desenvolvimento de software. Consequentemente, esta lacuna impede equipes de desenvolvimento de adotarem práticas éticas de forma eficaz. Portanto, para a operacionalização de requisitos éticos de IA, é importante que existam soluções práticas que apoiem este processo. Estas soluções são desenvolvidas em

diferentes formatos, sendo eles:

- Processos: Série de passos, ações ou atividades realizadas com o objetivo de obter um resultado especifico [17].
- Ferramentas: Um sistema que auxilia em alguma fase do processo de desenvolvimento de software pode ser considerado uma ferramenta de software [18].
- Técnicas: Uma técnica é constituída por um conjunto de métodos, abordagens e habilidades utilizados na aplicação de conhecimentos técnicos sistemáticos [19].
- Metodologias: Uma metodologia é composta de métodos, estudando-os e utilizando-os para compor uma estrutura maior [20].
- Métodos: Um método é uma maneira de estruturar o pensamento e ações de uma forma clara e explícita, servindo como prescrições para ações humanas [20].
- *Frameworks*: Um *framework* é uma solução de design genérica para um problema ou domínio específico. Define os vários elementos de design envolvidos na solução e as suas relações [20].

No entanto, as soluções práticas existentes ainda não estão suficientemente desenvolvidas devido a um certo número de dificuldades encontradas [21]. Ayling & Chapman [21] identificaram um conjunto de obstáculos encontrados quando se tenta operacionalizar o conjunto de requisitos éticos. O primeiro obstáculo se dá pela complexidade do domínio, que torna difícil desenvolver ou adquirir a ferramenta mais adequada para os objetivos determinados. O segundo se dá pelo fato de que atualmente não existem regulamentos ou legislação específica para sistemas de IA, dificultando o desenvolvimento de tais ferramentas. Já o terceiro obstáculo é a falta de clarificação das metodologias utilizadas na prática ética da IA.

Para ultrapassar estes obstáculos, é necessária a implementação de processos, ferramentas, metodologias, métodos, *frameworks* ou técnicas que utilizem as práticas existentes dentro do contexto de desenvolvimento de software. Assim, este trabalho propõe uma ferramenta automatizada para apoiar a operacionalização dos princípios éticos de IA por meio de práticas de Engenharia de Software, especificamente utilizando um conjunto de histórias de usuário sob uma perspectiva de ética de IA. A ferramenta se destaca ao transformar um conjunto de requisitos éticos em histórias de usuário, permitindo que os desenvolvedores integrem considerações éticas de forma prática e eficiente desde as fases iniciais do projeto, sob a perspectiva da ética em IA.

1.1 Justificativa

A necessidade de ética em IA decorre dos desafios encontrados em equilibrar a tensão entre o apoio à inovação, visto que é um direito da sociedade se beneficiar da ciência, e limitar os possíveis danos associados a sistemas baseados em IA mal projetados [22]. Portanto, é de grande importância realizar um estudo sistemático para compreender melhor as soluções de éticas para IA que foram propostas tanto na indústria como no meio acadêmico, tal como as lacunas existentes. Neste trabalho foi proposta uma atualização da revisão sistemática de literatura conduzida por Cerqueira [1], visto que o estudo utilizou estudos publicados até maio de 2021. Para uma verificação primária da necessidade desta atualização, foi feita uma pesquisa na base de dados da ACM Digital Library [1]. Esta pesquisa identificou que, desde a publicação da revisão sistemática de literatura [1], o número de estudos publicados tem apresentado um padrão de crescimento, como pode ser observado na Figura 1.1.

A atualização da revisão sistemática de literatura revelou uma discrepância entre os aspectos teóricos e práticos da ética em IA. Estudos como o de Barletta et al. [23] demonstraram que não é possível encontrar um *framework* completo, de fácil utilização, organizado e uniforme para apoiar os *stakeholders* durante todo o ciclo de vida no atual conjunto de soluções práticas, além de, em um futuro próximo, tal solução não será desenvolvida. Tidjon & Khomh [24] realizaram uma análise de 100 princípios éticos e suas implementações, com o objetivo de identificar e abordar as lacunas entre esses dois elementos. Entre as lacunas identificadas, os autores observaram uma falta significativa de ferramentas para apoiar a implementação dos princípios éticos em IA. Foi encontrado também que as ferramentas existentes tendem a focar em um grupo exclusivo de princípios e muitas vezes não abordam todos os aspectos relacionados aos princípios em questão. Além disso, os autores identificaram a falta de padrões efetivos para o desenvolvimento ético em IA [24]. Então visto que existe a necessidade da utilização de diferentes soluções no formato de uma infraestrutura de ética em IA, há uma procura por métodos e ferramentas práticas que utilizam padrões já existentes, facilitando a sua adoção por times de desenvolvimento [24].

Portanto, devido a crescente necessidade de ferramentas que auxiliem a implementação de princípios éticos em *IA*, o desenvolvimento de uma solução que transforme requisitos éticos em histórias éticas de usuário neste contexto específico é justificada por sua capacidade de ser aplicada em uma infraestrutura de ética em *IA*. Também é importante entender que em modelos de ciclo de vida de software, como cascata e modelos ágeis, a elicitação e análise de requisitos é uma das, se não a primeira etapa a ser realizada

[1] https://dl.acm.org

em um projeto [25]. Esta etapa também tem como característica a definição das bases do projeto, identificando as necessidades e expectativas dos *stakeholders*, incluindo questões não-funcionais, que precisam ser abordadas para garantir conformidade com padrões, incluindo o conjunto de princípios éticos de IA [25, 26].

Ao integrar o conjunto de princípios éticos desde a fase inicial, é possível garantir que eles sejam considerados no design e desenvolvimento do sistema de forma estruturada e contínua. Além disso, esta etapa é uma fase de comunicação crítica entre desenvolvedores, usuários e outros *stakeholders*, especialmente durante a elicitação [26]. Portanto ao utilizar uma ferramenta de ética em *IA* aplicada diretamente a análise e desenvolvimento de requisitos, é possível facilitar o entendimento e a aplicação prática dos princípios éticos para com *stakeholders* de diferentes níveis técnicos. Isso não só reduz o risco de falhas éticas no produto final, mas também diminui os custos de correções posteriores, que tendem a ser custosos.

Para a solução foi escolhida a transformação de requisitos éticos de alto nível em histórias éticas de usuário. Esta escolha se deu pela sua simplicidade, visto que uma história de usuário pode ser entendida como a representação mais granular de um requisito que os desenvolvedores usam para criar novos recursos, tal como sua ampla adoção na academia e indústria [27]. Outro fator para a escolha foi a existência do estudo de Halme et al. [8], que contém o *template* para a escrita e formalização de histórias éticas de usuário. Portanto, ao converter requisitos éticos em histórias de usuário, a ferramenta não apenas facilita a compreensão e implementação dos princípios éticos por parte dos desenvolvedores, mas também promove a integração de padrões éticos interconectados com as funcionalidades do sistema, criando uma ponte entre teoria e prática desde as primeiras etapas do projeto.

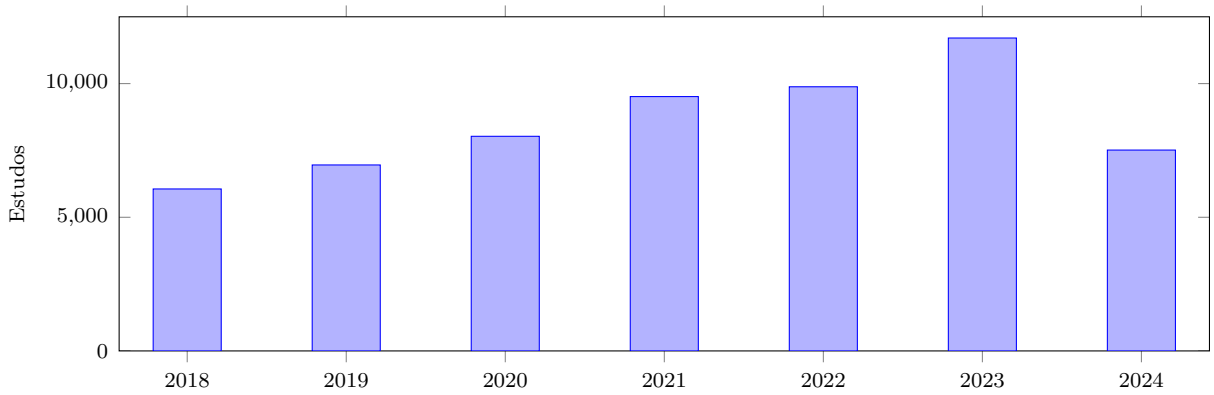


Figura 1.1: Estudos por ano relacionados à ética em IA na base de dados ACM. Fonte: o Autor.

1.2 Objetivos

O objetivo geral deste trabalho é desenvolver uma ferramenta que operacionalize a conversão de requisitos éticos de alto nível em histórias éticas de usuário.

1.2.1 Objetivos Específicos

Para atingir o objetivo geral deste trabalho, foram definidos os seguintes objetivos específicos:

- Conduzir a atualização da revisão sistemática de literatura publicada por Cerqueira [1], identificando novos estudos área de ética em IA.
- Identificar lacunas não preenchidas pelas soluções práticas existentes nas quais a ferramenta possa ser aplicada;
- Definir a arquitetura da ferramenta e de seus componentes para alcançar o resultado desejado;
- Identificar requisitos funcionais e não-funcionais para a ferramenta;
- Desenvolver a ferramenta;
- Avaliar a ferramenta;
- Propor melhorias e definir direções futuras para o avanço da ferramenta com base nos resultados da avaliação.

1.3 Resultados

Este trabalho alcançou os seguintes resultados:

- Identificação de processos, ferramentas, metodologias, métodos, *frameworks* ou técnicas para o desenvolvimento ético de IA em todas as fases do ciclo de vida de desenvolvimento de sistemas baseados em IA;
- Desenvolvimento de uma ferramenta para a conversão de requisitos éticos em histórias de usuário para apoiar a operacionalização de princípios éticos durante o desenvolvimento de sistemas baseados em IA;
- Avaliação da ferramenta desenvolvida para possível disponibilização aos profissionais de desenvolvimento de software.

1.4 Método de Pesquisa

Esta pesquisa foi guiada pelo método *Design Science Research* (DSR) [28]. Este método é baseado em três conceitos: pesquisa, design e ciência do design. O conceito de pesquisa pode ser definido como uma atividade que contribui para o entendimento de um fenômeno [28]. O conceito de design é atrelado a criação de algum artefato que ainda não existe [28]. Por fim, o design é transformado em uma ciência e chamado de ciência do design, sendo definida por Vaishnavi & Kuechler [28] como:

> O conhecimento sob a forma de construções, técnicas e métodos, modelos, teoria bem desenvolvida para efetuar um mapeamento do espaço funcional - um requisito funcional que constitui um ponto neste espaço multidimensional - para o espaço de atributos, onde um artefato que satisfaz o mapeamento constitui um ponto nesse espaço (Vaishnavi & Kuechler, 2015, p. 11, tradução nossa).

Esta combinação de conceitos da origem ao método DSR, que pode ser descrito como "uma pesquisa que cria este tipo de conhecimento em falta por meio do design, análise, reflexão e abstração"[28]. Portanto, o método pode ser resumido pela utilização de design como método de pesquisa, criando conhecimento por meio do desenvolvimento de artefatos [28].

A utilização deste método é justificado pela capacidade de ajudar os pesquisadores a construir e melhorar um artefato por meio de um processo contínuo de avaliação [28]. O método é composto por um modelo de processo composto de cinco etapas principais e suas saídas [28]. A Figura 1.2 contém as etapas e uma visão geral dos resultados desenvolvidos. As características mais detalhadas de cada etapa e sua aplicação neste trabalho são apresentadas a seguir:

- **Consciência do problema**: Esta etapa tem como foco a conscientização do problema de pesquisa, para entender os atores envolvidos, os objetivos, as causas do problema, os efeitos e as contribuições ao propor uma resolução. Portanto, nesta etapa foi realizada a atualização da revisão sistemática de literatura conduzida por Cerqueira [1] para determinar o estado atual de ética em IA e as abordagens prevalecentes para sua operacionalização. Esta etapa tem como saída a definição do estado atual das soluções práticas de ética em IA e suas lacunas por meio da RSL atualizada, possibilitando a definição e elaboração da ferramenta.

- **Sugestão**: Subsequente à etapa anterior, esta etapa é, em essência, a qual uma nova funcionalidade é imaginada com base em uma nova configuração de elementos existentes ou novos. Nesta etapa, a ferramenta foi definida como um meio para a tradução de requisitos éticos em histórias éticas de usuário utilizando *Large Language Models* (LLM). O resultado desta etapa foi a formulação de uma proposta e o planejamento da ferramenta.

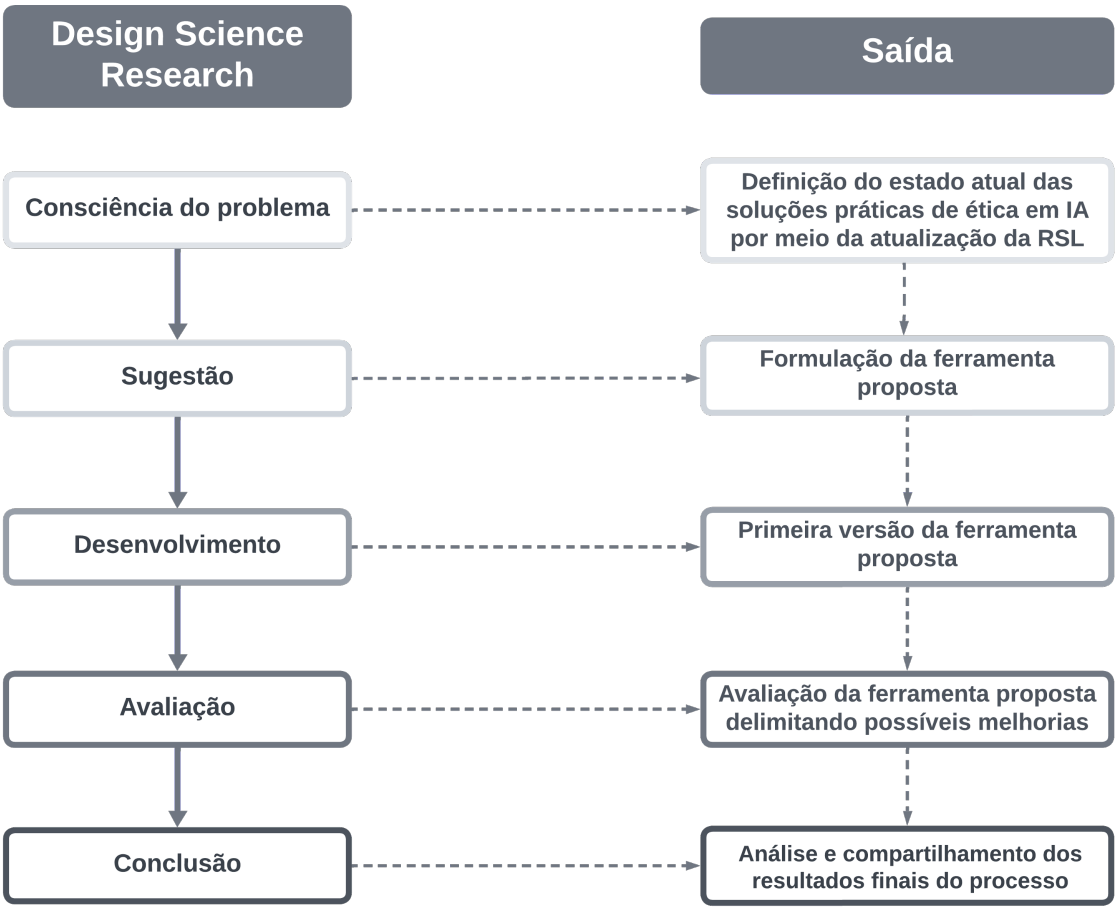


Figura 1.2: Etapas da Design Science Research com suas respectivas saídas. Fonte: o Autor.

- **Desenvolvimento**: Nesta etapa, o artefato planejado é refinado e desenvolvido. No caso deste trabalho, é importante ressaltar que a ferramenta tem como objetivo facilitar a adoção de ética em IA por times de desenvolvedores alterando minimamente o fluxo de desenvolvimento de software. O resultado desta etapa foi a primeira versão da ferramenta.

- **Avaliação**: Nesta etapa, o artefato é avaliado de acordo com critérios variados, normalmente com o objetivo de gerar evidências empíricas. Desvios quantitativos e qualitativos devem ser notados e explicados provisoriamente, gerando medidas de performance. O resultado desta etapa é a avaliação da ferramenta delimitando possíveis melhorias a serem realizadas em novas iterações do desenvolvimento.

- **Conclusão e Comunicação**: Nesta etapa, é realizada a análise dos resultados finais do processo, identificando as contribuições do trabalho, as melhorias potenci-

ais a serem realizadas na ferramenta e trabalhos futuros, tal como a comunicação e compartilhamento dos resultados encontrados. Neste trabalho os resultados serão compartilhados por meio da publicação dos artigos da revisão sistemática de literatura e da ferramenta.

1.5 Estrutura do trabalho

Este trabalho é organizado em quatro capítulos, em adição a introdução, sendo eles:

- **Capítulo 2:** apresenta o referencial teórico relacionado a IA, *Large Language Models* (LLM), ética em IA, utilização de histórias de usuário e as tecnologias definidas para o desenvolvimento da ferramenta.
- **Capítulo 3:** apresenta a Revisão Sistemática de Literatura conduzida neste trabalho, definindo características e respondendo as questões de pesquisa definidas.
- **Capítulo 4:** apresenta a metodologia utilizada no desenvolvimento da ferramenta, os resultados ao fim do processo de desenvolvimento e a validação do produto final.
- **Capítulo 5:** apresenta as principais conclusões e resultados do trabalho, bem como um panorama do trabalho futuro.

Capítulo 2

Referencial Teórico

Este Capítulo apresenta os principais elementos conceituais necessários para uma compreensão mais abrangente deste trabalho. Na Seção 2.1 o conceito de Inteligência Artificial é apresentado juntamente com características e funcionalidades das sub-áreas de Aprendizado de Máquina, *Deep Learning* e Processamento de Linguagem Natural (PNL). A Seção 2.3 define e apresenta as principais características de ética em IA, abordando o conjunto principal de princípios éticos encontrados na literatura, aspectos práticos de ética em IA e a engenharia de requisitos dentro do contexto de ética em IA. Por fim, na Seção 2.5 são apresentados os trabalhos relacionados e de interesse a este estudo.

2.1 Inteligência Artificial

O domínio da inteligência artificial (IA) está passando um período de rápido crescimento nos últimos anos. As suas origens remontam a uma série de disciplinas diferentes, incluindo a ciência da computação, a matemática, a filosofia, a sociologia e as ciências cognitivas [29]. O desenvolvimento de sistemas de IA tem como ponto inicial o fim da segunda guerra mundial, tendo o termo IA sido proposto pela primeira vez em 1956, o que o torna um dos domínios mais recentes da ciência [30]. Em menos de um século, o campo da IA avança a um ritmo notável, com aplicações em domínios tecnologicamente complexos, incluindo a construção de veículos autônomos, jogos, reconhecimento de falas, entre outros [30]. No contexto de Ciênca da Computação, IA é a disciplina dedicada ao desenvolvimento de sistemas artificiais que exibem características associadas à inteligência pelo ser humano [29].

Existem duas abordagens principais para o desenvolvimento de sistemas baseados em IA. O primeiro utiliza uma abordagem *top-down* conhecida como IA simbólica, na qual os desenvolvedores representam o conhecimento e regras de maneira declarativa por meio de regras e fatos [29]. A segunda, e a mais utilizada, utiliza uma abordagem *bottom-up*, na

qual processos cognitivos são modelados por meio de experiência [29]. Essa abordagem não descreve regras de maneira declarativa, mas utiliza modelos matemáticos e dados para gerar conhecimento sem a necessidade de representações explícitas [29]. Nas subseções seguintes alguns dos principais métodos utilizados para o desenvolvimento de IA serão explicados, assim como a abordagem utilizada neste estudo.

2.1.1 Aprendizado de Máquina

Um agente aprende se melhora o seu desempenho em tarefas futuras depois de efetuar observações sobre o mundo [30]. Para entender o conceito de Aprendizado de Máquina é importante entender o conceito de aprendizagem por meio de dados. Para Norvig & Russel [30] este conceito pode ser definido como o aprendizado que "a partir de um conjunto de pares entrada-saída, aprende uma função que prevê a saída para novas entradas". Para este tipo de aprendizagem, o usuário define um conjunto de dados, normalmente separado entre dados de treino e teste, e realiza o treinamento de um modelo matemático por meio deste conjunto de dados com o objetivo de prever dados fora do conjunto de treino [30]. Em suma, aprendizado de máquina se baseia em métodos numéricos para encontrar um procedimento de decisão que se comporta de maneira esperada em prática [29].

Este modelo matemático é treinado por meio do ajuste de seus parâmetros internos em resposta aos dados de entrada. Essencialmente, o sistema recebe os dados de entrada, os resultados previstos e um conjunto de parâmetros de ajuste. O sistema atualiza os seus parâmetros internos de aprendizagem por meio de funções matemáticas e gera resultados, que são depois comparados com os resultados previstos [29]. Finalmente, com base nesta comparação, o modelo efetua os ajustes nos parâmetros internos e realiza uma nova iteração com base na quantidade de iterações definida nos parâmetros de ajuste [29].

Assim, o objetivo principal da Aprendizagem de Máquina é a criação de um modelo que age sem a necessidade da escrita de códigos que ditam as ações ou previsões que com base em uma situação específica, mas que toma decisões adequadas por meio de padrões e semelhanças que reconhece de experiências anteriores [29]. O fluxo de um modelo de Aprendizagem de Máquina é apresentado na Figura 2.1.

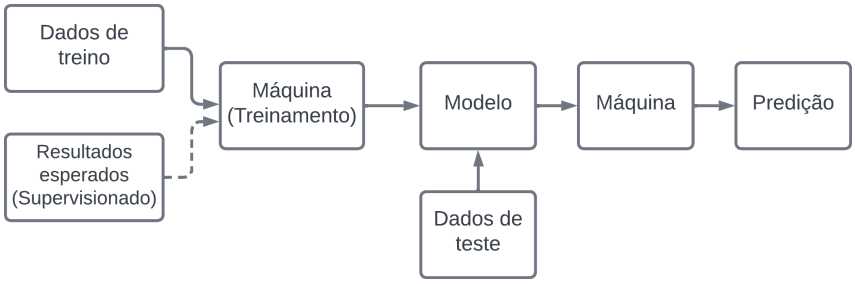


Figura 2.1: O fluxo de um modelo de Aprendizagem de Máquina. Fonte: Adaptado de [1]

No contexto de Aprendizagem de Máquina, os modelos podem aprender de duas maneiras principais: aprendizagem supervisionada e não-supervisionada. A aprendizagem supervisionada tem como objetivo o aprendizado de uma função que, com base nos dados de entrada e resultados esperados, melhor descreva as ligações entre entrada e saída [29]. Formalmente, dado um conjunto de treino de N exemplos de pares entrada-saída

$$(x1, y1), (x2, y2), ...(xn, yn),$$

em que cada yj foi gerado por uma função desconhecida $y = f(x)$, descobrir uma função h que aproxima a verdadeira função f [30]. No caso de aprendizagem não-supervisionada, o conjunto de treino não possui os pares de entrada-saída, apenas os dados de entrada. Neste caso, o objetivo é identificar um conjunto de estruturas ou padrões presentes nos dados de entrada, tirando conclusões sem a comparação com os resultados esperados [29].

2.1.2 *Deep Learning*

Os algoritmos de *Deep Learning* são abordagens à aprendizagem de máquina que utilizam modelos de redes neurais profundas. São particularmente úteis para domínios complexos devido à sua capacidade de aprender informações com mais precisão [29]. Isto é possível graças ao seu extenso conjunto de parâmetros e conexões, bem como à utilização de técnicas matemáticas específicas, como o método do gradiente e álgebra linear [29]. A arquitetura de modelos de *Deep Learning* é baseada no conjunto de neurônios biológicos, conectando camadas de neurônios para obter resultados por meio de passos pequenos, mas com um conjunto grande de camadas, neurônios e conexões [29]. Estes passos contêm operações matemáticas pontuais, como a multiplicação de matrizes, a normalização e aplicação de funções de ativação, que se tornam complexas quando ligadas [1, 29]. Algoritmos de *Deep Learning* podem ser utilizados em diversas áreas e com diferentes tipos de dados, como imagens, vídeos, texto, sequências ou áudio [1].

A arquitetura mais básica de uma rede neural profunda é composta em camadas de entrada, saída e ocultas. A quantidade de camadas ocultas vai de acordo com a modelagem feita pelos desenvolvedores [29]. Dignum [29] apresentou um exemplo de seu funcionamento:

> Em uma aplicação de reconhecimento de imagem, por exemplo, uma primeira camada de unidades pode combinar os dados brutos da imagem para reconhecer padrões simples na imagem; uma segunda camada de unidades pode combinar os resultados da primeira camada para reconhecer padrões de padrões; uma terceira camada pode combinar os resultados da segunda camada; e assim por diante (Dignum, 2020, p. 27, tradução nossa).

Na Figura 2.2 pode ser observada a arquitetura de uma rede neural profunda.

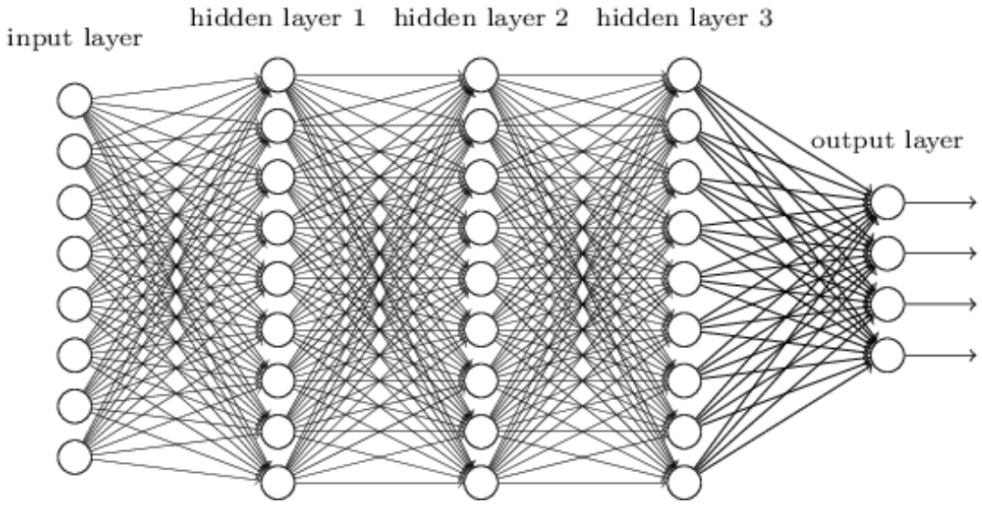


Figura 2.2: Arquitetura de uma rede neural. Fonte: Mazzeschi [2].

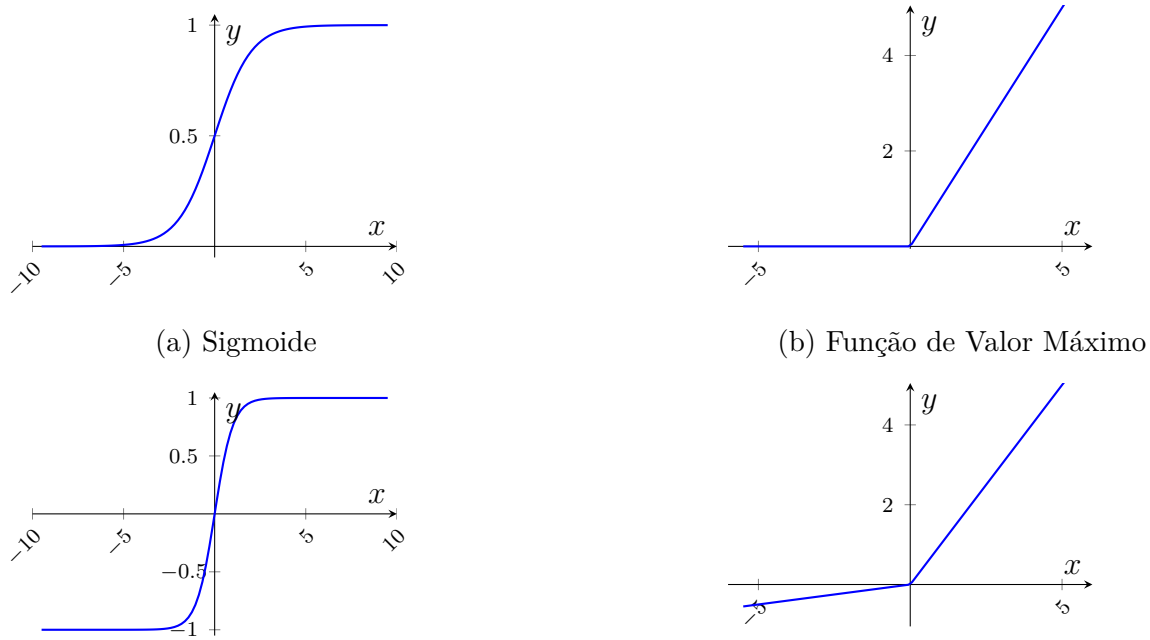
Formalmente o processo de predição, chamado de *forward pass* ou *forward propagation* é representado pela fórmula [31]:

$$z_j^{[i]} = {w_j^{[i]}}^T x + b_j^{[i]} \tag{2.1}$$

No qual z, w e b são respectivamente a saída, matrizes de peso e valor de viés. Além desta fórmula, é aplicada uma função de ativação no final de cada neurônio para introduzir não-linearidade. As quatro funções mais comuns são: sigmoide, tangente hiperbólica, função de valor máximo e unidade linear retificada (relu) [32]. Na Figura 2.3 pode ser visualizado o comportamento destas quatro funções. Por fim, o processo de treinamento é feito por meio do método chamado de *backward propagation*, que atualiza os pesos se baseando na saída real e na saída desejada [31]. O método calcula a derivada em relação a matriz de peso w utilizando a regra da cadeia e possui a seguinte fórmula [31]:

$$\frac{\partial L(z,y)}{\partial w} = \frac{\partial L(z,y)}{\partial a} \times \frac{\partial a}{\partial z} \times \frac{\partial z}{\partial w} \tag{2.2}$$

No qual z, w e b são respectivamente a saída, matrizes de peso e valor de viés, como no *forward pass*, y é a saída esperada, a o resultado da aplicação da função de ativação e $L(z,y)$ a função que compara a distância entre o resultado obtido e o esperado, chamado de função de perda [31].



(a) Sigmoide

(b) Função de Valor Máximo

(c) Tangente Hiperbólica

(d) ReLu

Figura 2.3: Funções de ativação. Fonte: o Autor.

No entanto, o domínio da *Deep Learning* utiliza um conjunto diversificado de modelos de redes neurais profundas para resolver uma série de problemas. No contexto das redes neurais profundas, existem vários tipos de arquitetura propostas. Por exemplo, no domínio de visão computacional, as redes neurais convolucionais são frequentemente utilizadas. Como pode ser observado na Figura 2.4, uma rede neural convolucional é normalmente composta por uma série de camadas de convolução e de agrupamento, seguidas de uma camada totalmente ligada e de uma camada de normalização [33]. Já outras arquiteturas utilizam o conceito de camadas das redes neurais profundas de uma forma distinta, como os *autoencoders*, que as utilizam em conjunto com um algoritmo não supervisionado para aprender a representação dos dados de entrada com o objetivo de redução da dimensionalidade e recriação do conjunto original [33].

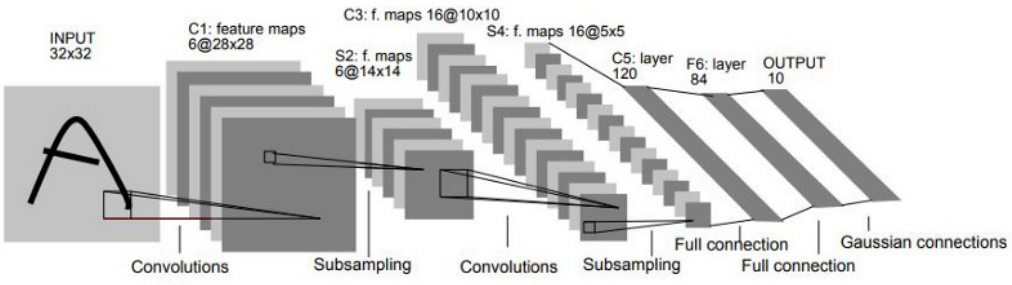


Figura 2.4: Arquitetura de uma rede neural convolucional. Fonte: Lecun et al. [3]

2.2 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PNL) é um domínio interdisciplinar de investigação e desenvolvimento que tem por objetivo melhorar a capacidade dos computadores para processar texto e discurso em linguagem natural, permitindo-lhes executar uma série de tarefas [34]. Ao integrar princípios computacionais, linguísticos e matemáticos, a PNL procura transformar a linguagem em comandos computacionais. As aplicações da PNL abrangem um conjunto diversificado de domínios, incluindo áreas como: tradução automática, o processamento e o resumo de textos em linguagem natural, as interfaces de usuário, a recuperação de informação multilíngue e interligue (CLIR), o reconhecimento de fala, a inteligência artificial e sistemas especializados, entre outros. [34].

2.2.1 *Deep Learning* aplicado a PNL

Antes de 2017, o campo de processamento de texto dentro de *Deep Learning* era composto primariamente por redes neurais recorrentes (RNN). Diferentemente das redes neurais profundas mais básicas, esta arquitetura utiliza ligações recorrentes para possibilitar a modelagem utilizando dados sequenciais para reconhecimento e previsão de sequencias [35]. A arquitetura RNN é composta por estados ocultos com uma dinâmica não linear [35]. Estes estados ocultos funcionam como um tipo de memória da rede, tendo seus estados num determinado momento condicionados ao seu estado anterior. A estrutura geral de RNNs permite que elas armazenem e processem sinais sequenciais complexos [35]. Isto é conseguido por meio do mapeamento de uma sequência de entrada para uma sequência de saída, tendo em conta o passo de tempo atual, permitindo assim a previsão do passo seguinte [35]. A arquitetura de uma RNN pode ser observada na Figura 2.5.

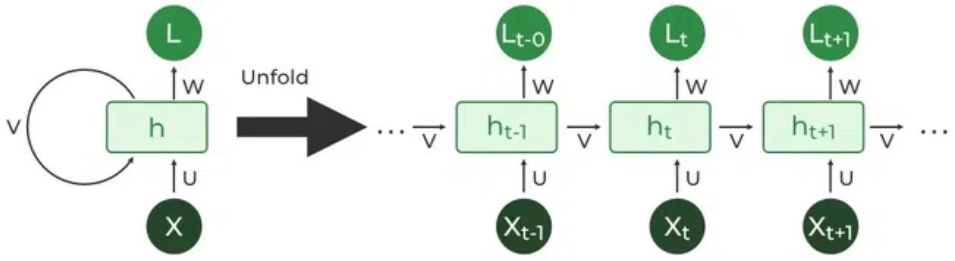


Figura 2.5: Arquitetura de uma rede neural recorrente. Fonte: *Geeks for Geeks* [4]

Mas os modelos de RNN sofrem com dois problemas em relação a informações de longo termo: explosão e desaparecimento de gradientes. O fenômeno de explosão de gradientes se dá pelo crescimento exponencial dos gradientes devido à propagação temporal, aumentando sua variação e gerando instabilidade no treinamento [35]. Já o fenômeno de desaparecimento dos gradientes se dá pelo decaimento exponencialmente, que ocorre à medida que este se propaga no tempo [35]. Para contornar este problema, foi proposta uma arquitetura conhecida como *Long Short Term Memory* (LSTM). Dentre os modelos que contornam estes dois fenômenos, LSTM é o mais popular [35].

A arquitetura de um modelo LSTM se diferencia do modelo básico de RNNs pela alteração nas suas células de memória. Em uma arquitetura LSTM, cada célula de memória está equipada com portas de entrada e saída que regulam a adição de entradas ao valor armazenado e a influência deste valor na saída [36]. De acordo com Le et al. [36]

> Essas portas são unidades logísticas com seus próprios pesos aprendidos nas conexões originadas da entrada e também das células de memória no passo de tempo anterior. Além disso, existe uma porta de esquecimento com pesos aprendidos que controla a taxa a que o valor analógico armazenado na célula de memória decai. Para os períodos em que as portas de entrada e saída estão desligadas e a porta de esquecimento não está causando decaimento, uma célula de memória simplesmente mantém o seu valor ao longo do tempo, de modo que o gradiente do erro em relação ao seu valor armazenado permanece constante quando passa pela etapa de *backpropagation* ao longo desses períodos (Le et al., 2015, p. 1, tradução nossa).

A arquitetura do modelo pode ser observada na Figura 2.6.

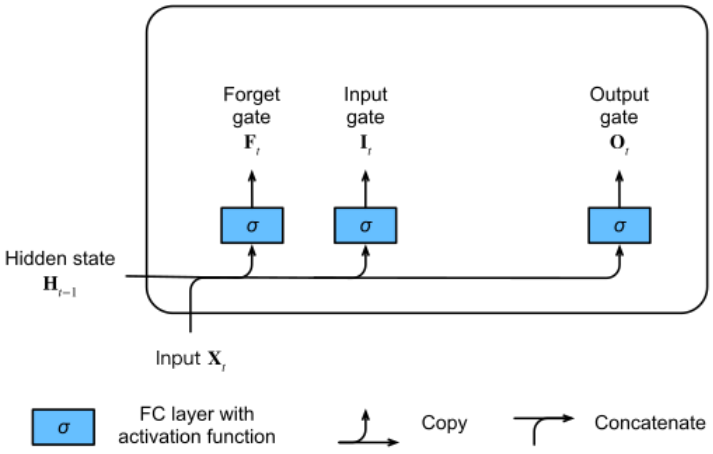


Figura 2.6: Arquitetura de uma LSTM. Fonte: Zhang et al. [5]

2.2.2 *Transformers* e *Large Language Models*

A arquitetura Transformers tem como objetivo contornar um problema inerente de modelos RNN, definido por Vaswani et al. [6]

> A natureza sequencial de modelos RNN impede a paralelização dentro dos exemplos de treino, o que se torna crítico em comprimentos de sequência mais longos, uma vez que as restrições de memória limitam o agrupamento de exemplos (Vaswani et al., 2017, p. 2, tradução nossa).

Esta arquitetura é composta de pilhas *encoder-decoder*, que por sua vez são compostas de normalizações, redes neurais profundas simples (*feed forward*) e o mecanismo principal, chamados de blocos de atenção [6]. Na Figura 2.7 pode ser observada a arquitetura de um *transformer*.

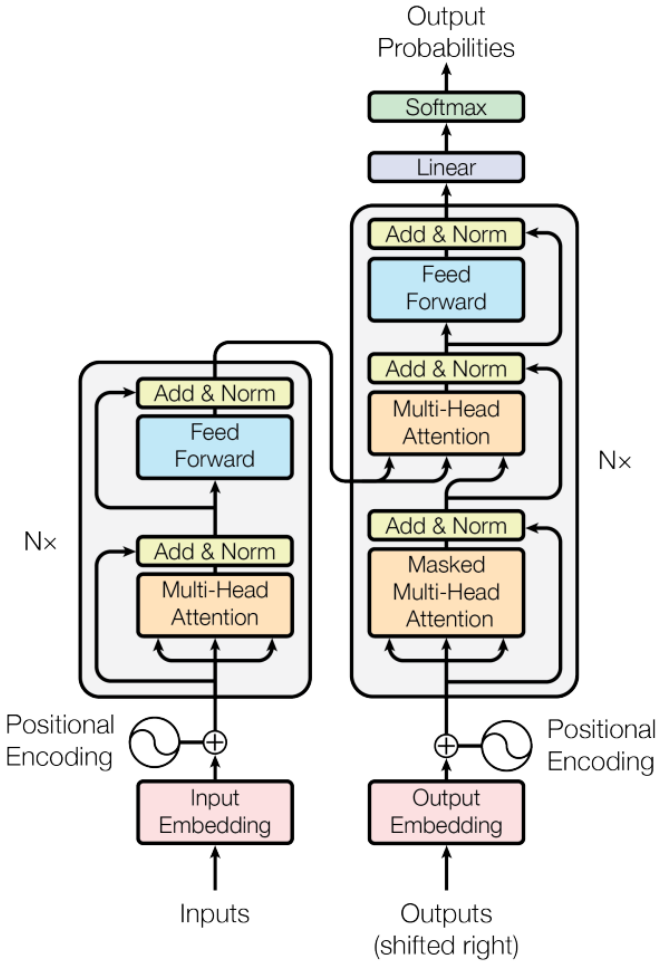


Figura 2.7: Arquitetura de um *transformer*. Fonte: Vaswani et al. [6].

O mecanismo de atenção é, em termos simples, uma função que mapeia uma *query* e um conjunto de pares chave-valor a um resultado, tendo em conta que as *queries*, as chaves e os valores são vetores [6]. O mecanismo de atenção é composto por uma "cabeça de atenção", que executa uma série de passos denominados "*Scaled Dot Product Attention*"[6]. Esta série de passos começa com o produto escalar entre o conjunto de vetores de *query* e

a chave que corresponde ao seu alinhamento. Após este passo, os valores são submetidos a um método de escalonamento, a uma função de mascaramento opcional e são finalmente transformados em probabilidades por meio da aplicação de uma função chamada *softmax*. Finalmente, é feito um produto escalar entre o resultado destes passos e o vetor de valores, resultando no delta da atualização dos valores. Formalmente, uma "cabeças de atenção"é definida por:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.3}$$

Na arquitetura proposta, várias "cabeças de atenção"funcionam em paralelo, em um mecanismo chamado de *multi-headed attention*, para garantir valores mais precisos [6]. Na Figura 2.7 pode ser observada a arquitetura interna das chamadas "cabeças de atenção"e *multi-headed attention*.

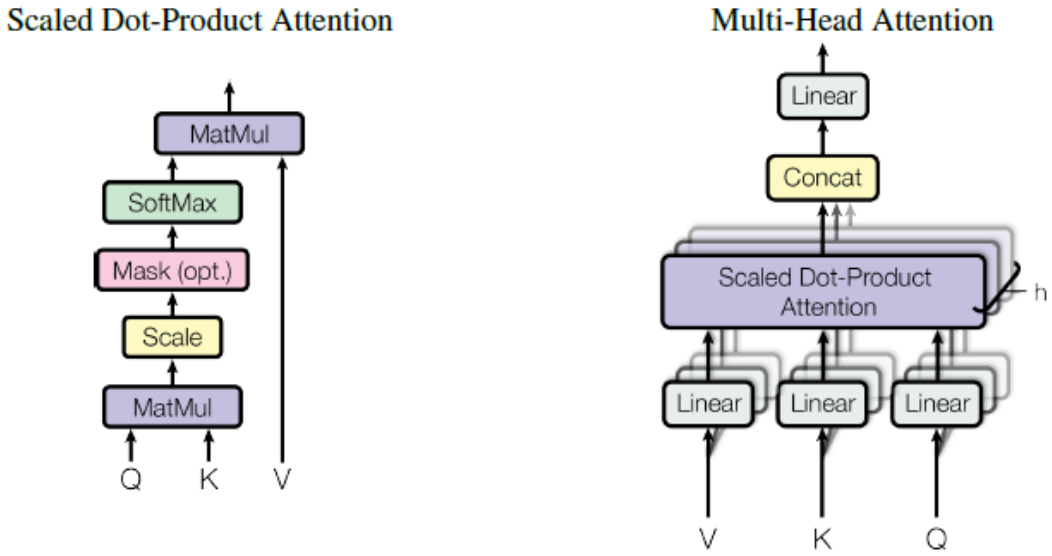


Figura 2.8: Arquitetura de "cabeças de atenção"e *multi-headed attention*. Fonte: Vaswani et al. [6]

Por meio da utilização de *Transformers*, a área de PNL conseguiu realizar avanços consideráveis, sendo o principal a criação de *Large Language Models* (LLMs). A lógica de desenvolvimento de uma LLM consiste em treinar primeiro a predição de palavras num corpus muito grande, com o objetivo de obter um modelo de uso geral [37]. Este seria depois adaptado a tarefas específicas, como responder a perguntas num determinado contexto, por meio de um processo chamado *fine-tuning* [37]. O conceito de *fine-tuning* é, em suma, um segundo treino supervisionado utilizando um conjunto de dados extremamente reduzido para contextualizar as previsões do modelo [37]. Por meio do treinamento inicial e o processo de *fine-tuning*, modelos de LLM podem ser contextualizados para um grande conjunto de tarefas que envolvem o uso de linguagem natural, sendo necessário apenas

um conjunto de dados e acesso a uma máquina potente o suficiente para treino. Neste trabalho será realizado o processo de *fine-tuning* em um modelo já existente.

2.3 Ética em IA

2.3.1 Princípios Éticos

O domínio de ética em IA é abordada por meio de vários princípios. Estes princípios de ética em IA são utilizados para categorizar as problemáticas de ética em IA e são, na sua maioria, separados em diretrizes constituídas por conjuntos de princípios [38]. Isto levou a que fosse proposto um conjunto diferente de princípios para cada diretriz publicada, uma vez que o foco principal da investigação até 2021 se centrou nos aspectos teóricos e conceituais, com propostas de princípios e diretrizes [16]. Jobin et al. [10] identificaram na sua análise de diretrizes que certos princípios estão presentes na maioria dos estudos, sugerindo a existência de uma convergência da ética em IA em torno destes princípios no panorama mundial. No entanto, os autores observaram que, embora se possa identificar quantitativamente um processo de convergência, existem diferenças na forma como estes princípios são interpretados e nos requisitos considerados necessários para a sua concretização[10].

Portanto, mesmo que existam tentativas de sintetização, é necessário definir uma abordagem a se seguir. Este estudo utilizará o conjunto de princípios definidos por Ryan & Stahl [39], que conduziram um estudo rigoroso no qual analisaram as diretrizes existentes e, após uma avaliação detalhada, apresentaram uma lista de onze princípios para o conjunto principal de *stakeholders* relevantes. Outras definições contendo um conjunto diferente de princípios pode ser observada no Capítulo 3.

Transparência

O conceito de transparência pode ser entendido de duas formas distintas: a transparência da própria tecnologia de IA e a transparência das organizações de IA que a desenvolvem e utilizam [39]. Este princípio está ligado à clareza em objetivos, possíveis danos e ganhos pelo uso do sistema de IA, possibilitando que os *stakeholders* principais entendam o funcionamento do sistema e possam realizar escolhas embasadas [39].

1. **Explanabilidade:** Relacionada ao monitoramento e controle ativo do sistema de IA, para garantir exatidão e confiabilidade. Métodos para tomada de decisão feitos pela IA devem ser documentadas pelas organizações para auditorias futuras, garantindo assim que a IA seja explicável a organismos externos de auditoria. Tensões entre desempenho e explicabilidade devem ser identificadas [39].

2. **Explicabilidade:** Organizações que utilizam ou desenvolvem IA devem ser capazes de explicar de forma concisa os dados de entrada, de saída, o funcionamento do algoritmo e o seu objetivo. Isto deve caminhar junto com um forte grau de rastreabilidade e explicabilidade para garantir a segurança do sistema. Decisões tomadas pela IA devem ser reproduzíveis por auditores externos [39].

3. **Compreensibilidade:** Organizações precisam implementar métodos para monitorar os dados, os algoritmos e o conjunto de decisões que serão tomadas por meio destes processos, garantindo que as decisões tomadas pela IA sejam compreensíveis. Além disso, as organizações devem compreender como funcionam os seus sistemas de IA de modo a explicar o funcionamento técnico e as decisões tomadas pela tecnologia [39].

4. **Interpretabilidade:** Organizações de IA devem ser capazes de entender o processo de decisão de seus algoritmos para assegurar um grau de supervisão humana para evitar e remediar possíveis danos. Domínios de alto risco, como saúde, justiça penal e segurança social devem reconsiderar a utilização de sistemas de IA com característica de "caixa-preta" [39].

5. **Comunicação:** Usuários finais devem receber informações exatas para garantir que não são manipulados, enganados ou coagidos pela IA e entender os objetivos e resultados da tecnologia. Além disso, organizações de IA devem explicitar as possíveis falhas e danos advindos de seus sistemas. Por fim, as organizações de IA devem comunicar aos governos os seus avanços e a probabilidade de atingirem determinados objetivos [39].

6. **Divulgação:** Sistemas de IA devem ser desenvolvidos e utilizados com o objetivo de utilizar a menor quantidade possível de dados pessoais. Em caso de utilização, é necessário que estes dados sejam anonimizados, encriptados e processados de maneira segura, existindo a possibilidade de auditorias externas neste processo. É necessário a realização de processos internos e externos de auditoria para garantir que os sistemas de IA estão adequados ao objetivo. Mesmo que sejam realizados processos de auditoria, organizações devem ser capazes de explicar e justificar a utilização de suas IAs [39].

7. **Apresentação:** Organizações devem garantir e apresentar a precisão e atualidade dos seus dados, garantindo transparência e monitoramento de sua qualidade. Desenvolvedores de IA fornecer acesso de seus códigos de ética para autoridades públicas, usuários organizacionais e, em casos possíveis, ao público. Isto pode ser possível por meio de revisões periódicas, mecanismos adequados de revisão e responsabilidade

coletiva. Além disso, é importante que o usuário final tenha consciência de que está interagindo com um sistema de IA e não com um ser humano [39].

Justiça e Equidade

Questões de justiça, igualdade e equidade são discutidas repetidamente em diretrizes de ética. Para além de abordarem simplesmente as questões de danos e injustiça, muitas delas fornecem recomendações sobre a forma de implementar medidas para minimizar esses danos [39]. Portanto os especialistas devem identificar o nível de justiça e equidade que pode ser incorporados no sistema [39]. Embora os desenvolvedores de IA possam ter os seus próprios valores, é imperativo que os algoritmos não sejam desenvolvidos com preconceitos historicamente injustos [39]. Devem também ser adotadas medidas para garantir que os dados utilizados pela IA não sejam injustos ou contenham erros e imprecisões que prejudiquem as decisões tomadas pela IA [39].

1. **Consistência:** Organizações devem garantir a coleta, análise e utilização de dados amostrais precisos e representativos para evitar ações prejudiciais no processo de tomada de decisão. Se torna imprescindível que organizações estabeleçam procedimentos para identificar e prevenir imprecisões nos seus sistemas de IA. Devem também ser realizadas auditorias algorítmicas externas e discussões periódicas entre o conjunto de *stakeholders* [39].

2. **Inclusão:** IAs não devem se tornar outra maneira de exclusão social. Deve ser dada atenção especial para grupos minoritários e vulneráveis, garantindo que dados utilizados sejam o mais inclusivos possíveis. Organizações devem reduzir problemas de exclusão e promover inclusão de grupos minoritários no desenvolvimento de IA [39].

3. **Igualdade:** IAs não devem prejudicar, mas tentar sempre promover igualdade em respeito a direitos, dignidade e liberdade. É necessário adotar mais medidas para combater os danos relacionados a sexismo, misoginia e preconceito de gênero resultantes de sistemas de IA [39].

4. **Equidade:** Objetivos de sistemas de IA devem ser empoderar e beneficiar indivíduos, providenciando oportunidades e distribuindo os benefícios da sua utilização de uma forma justa. IAs devem ser desenvolvidas para serem utilizadas de maneira justa e igualitária entre os membros da sociedade [39].

5. **Não enviesamento:** Organizações devem investir em meios para identificar, abordar e mitigar vieses negativos, examinando-os e eliminando-os durante todas as fases

do processo de desenvolvimento. Deve ser prestada especial atenção aos dados de treino, aos potenciais enviesamentos humanos e aos enviesamentos derivados dos resultados dos processos de algoritmo. Em qualquer identificação de enviesamento, as organizações devem demonstrar a eliminação do problema e informar para as autoridades competentes [39].

6. **Não-discriminação:** Sistemas de IA devem ser desenvolvidos para uso universal, prevenindo discriminação com base em atributos como gênero, raça e idade. Organizações devem implementar mecanismos para prevenir, remediar e reverter resultados discriminatórios tomados por IA, juntamente com a criação de "avaliações de impacto sobre a discriminação"para identificar problemas antes da disponibilização de sistemas de IA [39].

7. **Diversidade:** Organizações devem criar um ambiente de trabalho inclusivo, contratando funcionários de diferentes origens e áreas, realizar regularmente sessões sobre diversidade e integrar os os pontos de vista de um amplo conjunto de *stakeholders* [39].

8. **Pluralidade:** Desenvolvedores de IA devem ter em conta a variedade de pontos de vista sociais e culturais e tentem evitar a homogeneização social dos comportamentos e das práticas. Além disso, é de extrema importância que as organizações não se concentrem apenas nas mudanças de "modelos de pipeline", mas que também assegurem a pluralidade e representatividade de indivíduos, criando uma cultura de inclusão que deve refletir-se na tecnologia de IA. Recomenda-se que seja criado um diálogo entre o conjunto de *stakeholders* e que os pontos de vista das mulheres, dos grupos sub-representados e dos indivíduos marginalizados sejam incorporados em todas as fases das aplicações de IA [39].

9. **Acessibilidade:** Organizações devem garantir todos os direitos dos titulares de dados, dentre eles o direito de ter acesso aos dados que estão a ser armazenados e utilizados a seu respeito e, subsequentemente, de solicitar a sua eliminação. Decisões tomadas sobre indivíduos devem possuir explicações acessíveis, gratuitas e de fácil utilização [39].

10. **Reversibilidade:** Organizações devem definir se decisões tomadas por algoritmos de IA são reversíveis, garantindo que a autonomia de sistemas de IA seja restringida e que existam opções de revisão e correção em caso de resultados danosos [39].

11. **Remediação:** Ao existir possibilidade de dano por meio de algoritmos de IA, as organizações devem garantir que sejam tomadas medidas preventivas para detectar

estas questões e tratá-las de forma rápida e responsável. As organizações devem respeitar a máxima de que quando um sistema deixa de estar sob o controle humano, este deve ser terminado [39].

12. **Reparação:** Em situações de eventos danosos ou injustos advindos de sistemas de IA, indivíduos afetados devem dispor de medidas de reparação adequadas e rápidas, podendo também realizar reclamações e requerer justificação das decisões tomadas [39].

13. **Contestação:** Organizações devem permitir denuncias relacionadas a preocupações éticas. Devem também possuir políticas claras para proteger os opositores, garantindo que possam expressar preocupações éticas e se sentirem protegidos ao realizá-las [39].

14. **Acesso e distribuição:** Organizações devem garantir que suas tecnologias sejam justas e acessíveis, concentrando na população em desvantagem social, como indivíduos com problemas de visão e mobilidade. Sempre que possível, as organizações devem utilizar dados abertos para a sua IA, a fim de garantir o acesso e a transparência [39].

Não maleficência

No domínio da ética da IA, a prevenção de danos aos seres humanos tem sido uma preocupação fundamental. Para categorizar essa prevenção é utilizado o princípio de não maleficência. Este princípio é definido, em sua forma mais básica, como a prevenção de danos a outros. Portanto é necessário que sistemas de IA sejam desenvolvidos com o objetivo de não causar danos previsíveis a seres humanos. Para isso organizações devem testar os seus algoritmos regularmente com o objetivo de identificar possíveis danos. Deve também ser integrado o aconselhamento jurídico e de conselhos de ética para garantir que os dados sejam utilizados de uma forma que não prejudiquem as pessoas [39].

1. **Segurança (*Security*):** Sistemas de IA devem ser robustos e seguros durante seu ciclo de vida. Além disso, é essencial que a segurança seja integrada na arquitetura do sistema e que esta seja testada antes da implementação. Organizações devem garantir que sua IA possua medidas de segurança contra ataques cibernéticos [39].

2. **Proteção (*Safety*):** Organizações devem aplicar medidas rigorosas de segurança, garantindo gestão e controle da IA. Devem também existir procedimentos adequados em caso de violação de segurança. Desenvolvedores devem garantir que a IA não viole direitos humanos, assegurando a segurança da tecnologia. A IA deve passar por processos de garantia de qualidade e ser testada em cenários reais [39].

3. **Danos:** Objetivos e impactos esperados da IA devem ser verificados e documentados antes de ser disponibilizada. A IA não deve causar danos físicos ou psicológicos a nenhuma pessoa. Qualquer IA que substituir atividade humana deve produzir diminuição de danos antes de ser utilizada no mercado.

4. **Proteção:** A IA deve ser segura durante todo seu ciclo de vida, possuindo mecanismos para proteger os usuários. Auditores externos devem ser autorizados autorizados a realizar verificações e a apresentar relatórios sobre os impactos negativos da IA sem receio de danos ou ameaças por parte das organizações. Além disso, é necessário garantir a proteção dos denunciantes nas organizações [39].

5. **Precaução:** Para garantir a funcionalidade contínua de um sistema de IA, é essencial que este seja controlável e administrável, bem como robusto e confiável. Isto protegerá o sistema contra ataques e manipulações [39].

6. **Prevenção:** Desenvolvedores devem garantir que o sistema seja gerenciável, confiável e robusto, evitando ataques, acesso e manipulação. Além disso o sistema de IA deve prevenir a ocorrência de acidentes, sempre que possível, e evitar a ocorrência de situações críticas [39].

7. **Integridade:** Ataques a IA não devem comprometer a integridade física e mental das pessoas, garantindo assim que o sistema seja confiável e robusto. Em casos extremos, a IA deve desligar de maneira segura ou entre em modo de segurança [39].

8. **Não subversão:** Sistemas de IA devem ser utilizados com o objetivo de respeitar e melhorar a vida em sociedade, evitando subverter processos sociais e cívicos [39].

Responsabilidade

O princípio de responsabilidade é de grande interesse dentro do domínio de ética em IA, uma vez que existe a preocupação de que as grandes organizações possam tentar disfarçar a sua responsabilidade, culpando os sistemas autônomos. Além disso, o desenvolvimento irresponsável de IA por meio de modelos "caixa-preta"pode gerar lacunas de responsabilidade, nas quais se torna difícil encontrar o responsável por problemas éticos. Para evitar este tipo de problemas, é essencial definir que os programadores são os principais responsáveis pelo comportamento do sistema, seu design e suas funcionalidades. Já quando o problema é causado pela utilização e implementação da tecnologia, o responsável é o cliente organizacional. É necessário que os responsáveis, principalmente desenvolvedores, tenham a noção clara de que são os responsáveis pelo impacto destes sistemas [39].

1. **Responsabilidade (*Accountability*):** Organizações devem estar conscientes de que as consequências da utilização de dados de má qualidade são da sua responsabilidade. Além disso devem estar disponíveis para auditorias, monitoramento e efetuar avaliações de impacto da IA regularmente [39].

2. **Responsabilidade civil:** Organizações devem definir, de maneira clara, as atribuições de responsabilidade em situações de mau funcionamento, erro e danos. Isto pode ser conseguido por meio de uma manutenção de registos adequada e de documentação [39].

3. **Agir com integridade:** Organizações devem garantir a qualidade e a integridade dos dados, bem como implementar treinamentos de ética em IA, a fim de evitar problemas futuros. Ao identificar erros, violações de segurança ou vazamento de dados, o desenvolvedor deve comunicar esses problemas às autoridades competentes e partes interessadas [39].

Privacidade

Tendo em conta a grande quantidade de dados necessários para que sistemas de IA funcionem de forma correta, é fundamental que a privacidade individual não seja comprometida como consequência. Organizações devem tomar medidas como anonimização de dados armazenados, detecção de anomalias, garantia de consentimento e controle de dados armazenados pelos usuários para garantir privacidade [39].

1. **Informação pessoal ou privada:** O desenvolvimento e a utilização de IA devem ser conduzidos de acordo com os regulamentos locais de privacidade e proteção de dados. Os dados pessoais dos usuários finais, bem como os dados derivados ou criados a seu respeito, devem ser tratados de forma justa, legal e legítima. A coleta e utilização de dados pessoais devem ser reduzidas ao mínimo [39].

Beneficência

No domínio da ética da IA, o princípio da beneficência é frequentemente ignorado. Esta situação deve-se ao fato da suposição de que a IA sempre trará benefícios. O princípio pode ser definido como fazer o bem, realizando alguma ação com a intenção de beneficiar um indivíduo, grupo, ou sociedade como um todo. Para implementar este princípio em seus sistemas de IA, organizações devem utilizar dados para o benefício de seus consumidores e a sociedade. Por fim, a IA deve complementar de maneira positiva a realidade humana [39].

1. **Benefícios:** Organizações devem garantir que a sua IA beneficia os seres humanos, definindo claramente quais são estes benefícios e quais partes são beneficiadas. Estes benefícios devem abranger o maior número possível de pessoas [39].

2. **Bem-estar:** Organizações devem garantir o bem-estar individual e assegurar que sua IA não inibe o crescimento individual e acesso a bens primários e garante bem-estar humano. A IA deve ser utilizada para complementar as pessoas que trabalham no setor da saúde, a fim de melhorar a prestação de cuidados e apoiar o bem-estar dos doentes [39].

3. **Paz:** Organizações devem se esforçar para prevenir uma corrida armamentista relacionada a armas autônomas. Se a IA ameaça a paz, organizações devem colaborar com governos para reduzir conflitos potenciais [39].

4. **Bem-estar social:** Organizações devem garantir que a sua IA tragam oportunidades e melhorias para a sociedade, além de cultivar uma indústria saudável. O uso de IA não deve resultar num conflito com aqueles que não utilizam estas tecnologias [39].

5. **Bem comum:** Sistemas de IA devem ser desenvolvidos para apoiar o bem comum e servir as pessoas. Organizações devem ponderar os benefícios e danos resultantes de sua IA, sempre procurando solucionar os possíveis danos para proporcionar o bem comum.

Liberdade e Autonomia

Sociedades democráticas valorizam liberdade e autonomia e é de grande importância que sistemas de IA não sobrecarreguem nem prejudiquem estes valores. Desenvolvedores devem identificar e tratar circunstâncias nas quais a IA possa trazer danos por meio de rastreamento, censura ou vigilância, cerceando liberdades sociais. Organizações devem garantir que os usuários finais estão informados e não sejam manipulados pela IA, sempre podendo exercer sua autonomia [39].

1. **Consentimento:** A utilização de dados pessoais deve ser explicitamente definida e aceite pelos usuários, e não deve ser processada de formas que estes considerem inadequadas. Caso os dados pessoais sejam reutilizados, os desenvolvedores devem garantir que são compatíveis com os requisitos originais consentidos [39].

2. **Escolha:** A IA deve proteger o poder dos usuários de tomar suas próprias decisões. Ela não deve também comprometer a liberdade e a autonomia humanas [39].

3. **Auto-determinação:** Organizações não devem manipular a auto-determinação individual, principalmente de quem se encontra em situações de vulnerabilidade. Deve haver um equilíbrio entre o poder de decisão concedido pelo usuário a IA a sua retirada pelo sistema [39].

4. ***Liberty*:** É da responsabilidade das organizações garantir que os seus sistemas de IA não infringem as liberdades individuais, tal como definidas na legislação sobre direitos humanos [39].

5. **Empoderamento:** A IA deve ser utilizada para empoderar e fortalecer direitos humanos. No caso de serem tomadas decisões que possam resultar na violação das liberdades de um indivíduo, é fundamental que este tenha o direito de contestar essas decisões [39].

Confiança

A confiança é um princípio fundamental para as interações interpessoais e é um preceito fundamental para o funcionamento da sociedade, portanto é de extrema importância no desenvolvimento de sistemas de IA [39].

1. **Confiabilidade:** Organizações devem provar que são confiáveis e que suas tecnologias são credíveis. Usuários finais devem poder confiar que as organizações de IA cumpram suas promessas e garantam o funcionamento previsto dos sistemas. A confiabilidade pode ser cultivada por meio de demonstrações da segurança de sua IA, tal como pelo armazenamento responsável dos dados obtidos a partir destes sistemas [39].

Sustentabilidade

Sustentabilidade é um tópico universal, impactando todos os domínios, e a IA não é exceção. As organizações devem garantir que são sustentáveis do ponto de vista ambiental e que incorporam os resultados ambientais nas suas decisões. Os sistemas de IA devem aderir a uma utilização eficiente dos recursos, à promoção da energia sustentável e à proteção da biodiversidade [39].

1. **Meio-ambiente:** Organizações devem utilizar IAs desenvolvidas de maneira ambientalmente consciente. Situações em que danos ambientais causados por IA ultrapassem níveis aceitáveis, seu uso deve ser interrompido e medidas devem ser tomadas para identificar maneiras de uso não prejudicial [39].

2. **Energia:** O uso de IA deve respeitar os limites de eficiência energética, mitigar a emissão de gases com efeito estufa, proteger a biodiversidade e manter seu impacto ecológico em níveis mínimos [39].

3. **Recursos:** Sistemas de IA devem ser desenvolvidos de maneira que garanta o consumo eficaz de energia e recursos, promovendo a eficiência de recursos, utilização de materiais renováveis, redução da utilização de materiais escassos e um nível mínimo de resíduos [39].

Dignidade

Sistemas de IA devem respeitar a dignidade humana, definida pelo reconhecimento de que os indivíduos têm um valor inerente e que os seus direitos devem ser respeitados. Desenvolvedores devem garantir que sistemas de IA respeitem , sirvam e protejam integridade mental e física humanos, tal como senso de identidade pessoal e cultural. Deve ser explícito que o usuário está a interagir com uma IA e não um ser humano. Organizações devem garantir que sistemas de IA não violem a dignidade dos usuários finais [39].

Solidariedade

O potencial da IA para divulgar informações falsas e o risco de violação da privacidade geraram preocupações quanto à possibilidade da IA ser utilizada para comprometer as relações sociais e aspectos de solidariedade. Sistemas de IA devem ser desenvolvidos para promover, ou evitar prejudicar seguridade social, segurança e coesão, evitando pôr em risco os laços e as relações sociais. IA deve promover o desenvolvimento humano ao invés de colocá-lo em perigo [39].

1. **Seguridade social:** A IA não deve comprometer valores democráticos, garantindo que cidadãos recebam informações acuradas e não manipuladas com objetivos políticos. Além disto, sistemas de IA não devem ser utilizados para comprometer decisões eleitorais e políticas [39].

2. **Coesão:** Organizações devem promover a distribuição justa de benefícios concebidos por IA. Organizações devem desenvolver ativamente estratégias com parcerias para promover coesão social e intercâmbio de conhecimentos [39].

2.4 Requisitos Éticos para IA

O desenvolvimento de um sistema exige a identificação das suas funcionalidades, características e necessidades. No domínio da engenharia de software, este papel é desempenhado

pela engenharia de requisitos. Para entender o conceito de engenharia de requisitos, é necessário de uma definição formal de um requisito. Bozyiğit et al. [40] definem o conceito de requisito como "as descrições das funcionalidades necessárias a projetar e desenvolver em um determinado sistema de software". Assim, de acordo com esta definição, a engenharia de requisitos pode ser conceituada como o processo de elicitação, especificação e validação de requisitos da totalidade dos *stakeholders* do sistema [41].

Neste trabalho o foco será na etapa de especificação de requisitos, na qual os analistas e engenheiros lidam com a documentação dos requisitos elicitados anteriormente. A fase de especificação pode ser realizada por meio do uso de uma variedade de técnicas, incluindo a modelagem de casos de uso, cenários, protótipos e definições formais [41]. A técnica abordada será a de histórias de usuário. Uma história de usuário é definida como uma caraterística ou funcionalidade que *stakeholders* desejam ver incluída no produto final [42]. O *template* básico utilizado para as histórias de usuário é: "como <*stakeholder*>, eu quero <objetivo>, [para que eu possa <benefício>]" [27].

A identificação de problemas, abordagens e soluções éticas em sistemas de IA pode ser realizada por meio da engenharia de requisitos. Para isso é necessário traduzir as especificações éticas em requisitos específicos chamados de requisitos éticos. De acordo com Guizzardi et al. [43], requisitos éticos são "requisitos para sistemas de IA derivados de princípios éticos ou códigos éticos". A funcionalidade e as qualidades de sistemas de IA são suficientes para cumprir os requisitos éticos, não havendo, portanto, necessidade de os tratar como agentes éticos propriamente ditos [43].

Para Agbese et al. [44], o conceito de requisitos éticos para IA são requisitos para sistemas de IA derivados de diretrizes, princípios ou normas éticas de IA, tal como os requisitos legais são derivados de leis e normas. A partir das definições apresentadas, pode observar-se que os requisitos éticos no contexto de IA podem ser implementados por meio da colaboração entre *stakeholders*, derivado da prática de engenharia de requisitos, e diretrizes, normas ou princípios éticos.

2.5 Trabalhos Relacionados

Vakkuri et al. [45] apresentaram o ECCOLA, uma metodologia interativa que tem como objetivo disponibilizar aos desenvolvedores uma ferramenta prática para a implementação da ética em IA. A metodologia utiliza um conjunto de 21 cartas, reunidas em oito temas. Sete destas categorias estão relacionadas com a análise dos princípios éticos, enquanto a última é dedicada à análise de *stakeholder*. O conjunto de temas é baseados nas diretrizes definidas pelo grupo de especialistas de IA da União Europeia (HLEG AI) e IEEE EADv1. Cada carta é dividida em três partes: motivação (porque é importante), o que fazer (como

solucionar o problema) e exemplo prático do tópico (para facilitar a aplicação prática). A metodologia é aplicada por meio de um processo modular, integrando ética de IA em histórias de usuário.

Halme et al. [8] propuseram a utilização de histórias éticas de usuário como uma ferramenta para promover a integração da ética em IA em práticas convencionais de engenharia de software. Os autores classificam histórias éticas de usuário como "histórias de usuário concebidas para ajudar a abordar e formalizar questões éticas na engenharia de software, do ponto de vista de um determinado contexto ético". O *framework* utilizado para seu embasamento pode ser qualquer tipo de ferramenta ética, independentemente do seu tipo específico, como diretrizes, métodos, metodologias, princípios, ferramentas ou *frameworks* existentes na literatura e relacionados ao contexto do sistema de IA. O processo consiste em identificar as necessidades do usuário, plicar uma ferramenta ética adequada a tais necessidades, criar histórias de usuário relacionadas a ferramenta e, por fim, formalizar as funcionalidades do sistema.

Luitel et al. [46] desenvolveram uma maneira de utilizar um sistema baseado em *Large Language Models* (LLMs) para gerar previsões contextualizadas para o preenchimento de requisitos de software. O modelo BERT foi utilizado na geração de previsões juntamente WikiDoMiner, uma ferramenta que é capaz de extrair automaticamente um corpus específico do domínio para uma entrada. Posteriormente, os dados são submetidos a um processamento de filtragem utilizando um classificador de *machine learning* com o objetivo de determinar se as previsões geradas pelo modelo anterior são relevantes ou não. A utilização dos processos de filtragem e extração de corpus deste estudo resultaram em um sistema que consegue realizar o processo de preenchimento de requisitos sem a necessidade de *fine-tuning*.

Par facilitar a elicitação de requisitos éticos em sistemas de IA, Cerqueira et al. [16] desenvolveram um guia chamado RE4AI baseado na metodologia ECCOLA de Vakkuri et al. [45]. O principal objetivo do guia é facilitar a definição de requisitos pelos usuários e a sua incorporação em sprints como histórias de usuário, aumentando assim os requisitos globais do sistema. Outra característica chave do guia é sua inclusão de participantes, visto que é encorajado que diferentes *stakeholders* da organização façam parte da atividade de elicitação de requisitos éticos, incluindo a participação dos usuários. O guia é composto por 26 cartas divididas em onze princípios, sendo implementado para ser utilizado de maneira online[1].

A Tabela 2.1 apresenta uma análise comparativa entre os estudos relacionados e a proposta desta pesquisa. A comparação é feita por meio de três características principais deste estudo. A primeira destas características diz respeito ao uso de técnicas de processa-

[1] https://josesiqueira.github.io/RE4AIEthicalGuide/index.html

mento de linguagem natural. A segunda diz respeito à implementação de histórias éticas de usuário. Por fim, a terceira aborda a incorporação de requisitos éticos nos sistemas de IA.

Tabela 2.1: Comparação entre os Trabalhos Correlatos

	Características		
	PLN	Histórias Éticas de Usuário	Requisitos Éticos de IA
Vakkuri et al.[45]			Sim
Halme et al.[8]			Sim
Luitel et al.[46]	Sim	Sim	
Cerqueira et al.[16]	Sim		Sim
Este estudo	Sim	Sim	Sim

Fonte: o Autor

Capítulo 3

Revisão Sistemática de Literatura

Neste capítulo foi efetuada uma atualização da Revisão Sistemática da Literatura (RSL) realizada por Cerqueira [1] para elucidar o estado atual da ética em IA. Foram analisadas diretrizes, princípios, métodos, metodologias, *frameworks*, ferramentas, processos e técnicas para identificar o estado da arte das soluções éticas de IA e como utilizá-las para o desenvolvimento de sistemas éticos. A atualização foi conduzida com base nas fases propostas por Kitchenham et al. [47, 48]:

1. Planejamento: essa fase tem como objetivo identificar a necessidade de uma revisão, definir os objetivos e criar o plano da revisão, composto por questões de pesquisa, estratégia de pesquisa, critérios de seleção, *string* de busca, *quality assessment* dos estudos e estratégia de extração e sintetização de dados.

2. Condução: tem como objetivo colocar o protocolo de planejamento em prática, utilizando os artefatos e definições para filtrar e sintetizar os estudos.

3. Publicação de Resultados: objetiva documentar os resultados da RSL, respondendo as questões de pesquisa e avaliando os resultados encontrados.

3.1 Atualização da RSL

Foi identificada a necessidade de realizar esta atualização, uma vez que uma revisão atualizada é essencial para garantir que a revisão sistemática da literatura reflete com precisão os debates, desafios e soluções mais atuais, fornecendo assim uma visão abrangente do estado atual do campo. Além disso, a incorporação de estudos recentes pode realçar lacunas na literatura existente e sugerir novas direções de investigação. Para apoiar este argumento, foi utilizado o *framework* de decisão proposto por Garner et al. [7] para verificar a viabilidade da atualização. A escolha do *framework* foi feita com base no artigo

de Mendes et al. [49], no qual é apresentado um conjunto de propostas e diretrizes sobre quando e como atualizar uma RSL.

O *framework* de decisão de Garner et al. [7] consiste de três etapas principais aplicadas sequencialmente, como pode ser observado na Figura 3.1. Elas são compostas de perguntas que podem ser respondidas de maneira positiva e negativa. As etapas do *framework* são:

(Etapa.1) Entender o quanto a RSL é atual com base na relevância do tópico, além de analisar os métodos utilizados para a sua condução. Esta etapa pode ser respondida apenas com "Sim" ou "Não";

(Etapa.2) Identificar novos métodos, estudos, diretrizes ou artigos publicados após a RSL inicial. Esta etapa pode ser respondida apenas com "Sim" ou "Não" e requer apenas um "Sim" para dar continuidade a aplicação do *framework*;

(Etapa.3) Entender se a adoção de novos métodos, estudos, diretrizes ou artigos publicados afeta diretamente o resultado final ou a credibilidade da RSL. É necessário entender que critérios de inclusão (IC), exclusão (EC) ou questões de pesquisa (RQ) alteram diretamente o resultado final, portanto alterações ou adições resultam em "Sim" nas respostas. Esta etapa pode ser respondida com "Sim/Talvez" ou "Não" e requer apenas um "Sim" para garantir a viabilidade da atualização.

Após a apresentação do *framework* proposto por Garner et al. [7], ele foi aplicado a RSL publicada por Cerqueira [1]. O intuito é a comprovação da relevância do tópico e da necessidade de atualização.

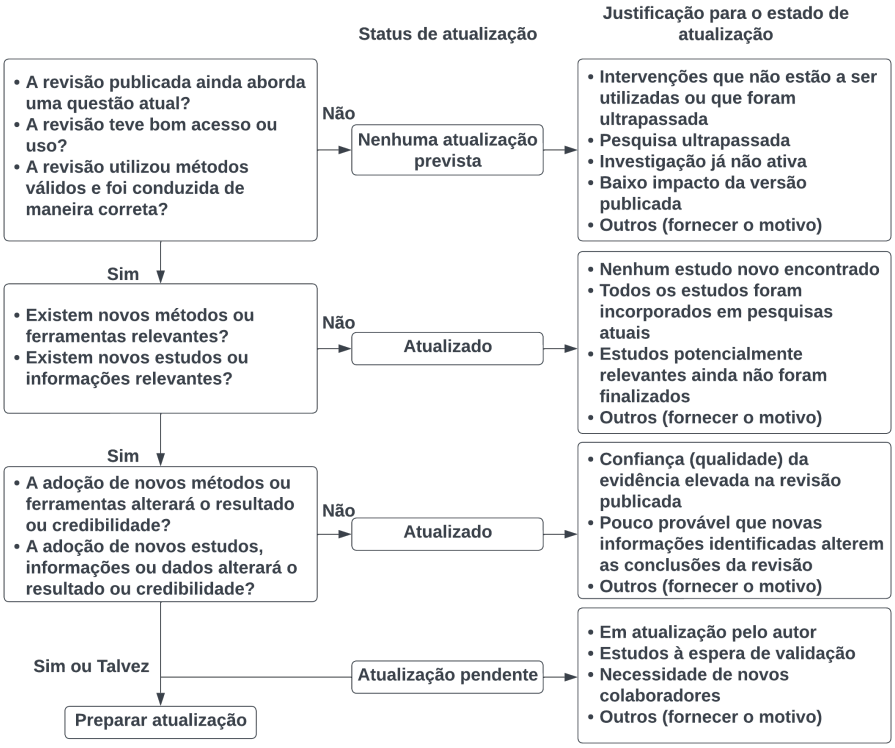


Figura 3.1: *Framework* de decisão. Fonte: Garner et al. [7]

3.1.1 Etapa 1

Para avançar a etapas seguintes, é necessário responder "Sim" para todas as perguntas da etapa inicial. Ela é composta de três perguntas como pode ser verificado na Figura 3.1, sendo elas:

1. A revisão publicada ainda aborda uma questão atual?
2. A revisão teve bom acesso ou uso?
3. A revisão utilizou métodos válidos e foi conduzida de maneira correta?

Garner et al. [7] propuseram uma simples pesquisa bibliográfica como principal meio para responder às duas primeiras questões. É necessário definir se o tema ainda é formalmente debatido e se existem novas informações que possam ser relevantes. Assim, a métrica utilizada foi o número de estudos publicados relacionados com o tema da RSL após a sua publicação. Não é necessário pesquisar em várias fontes, uma vez que o objetivo é apenas confirmar a existência de um número significativo de estudos relacionados.

A RSL de Cerqueira [1] foi publicada em abril de 2021, portanto, apenas estudos publicados entre maio de 2021 e março de 2024 foram considerados. A base de dados escolhida foi a *ACM Digital Library*, por agregar um conjunto maior de estudos. A

pesquisa retornou um total de 1.011 estudos publicados, o que serve para demonstrar a relevância do tema em questão. Consequentemente, ambas as questões foram respondidas positivamente.

Por último, segundo Garner et al. [7], a terceira questão deve ser baseada na leitura da RSL. É possível encontrar problemas como critérios de seleção vagos, métodos inadequados e resolução mal articulada. No caso de ambas as questões anteriores terem resposta positiva, a atualização pode então prosseguir. Isto acontece porque a atualização pode prosseguir com a definição de novos métodos, critérios de seleção e questões de avaliação de qualidade, mesmo que a resposta desta questão seja negativa. Nesta atualização, foram definidos novos critérios de seleção, questões de avaliação de qualidade e métodos. Desta forma, todas as questões foram levadas em conta e, como foram respondidas positivamente, é possível passar para a etapa seguinte.

3.1.2 Etapa 2

A segunda etapa é composta de apenas duas perguntas. Nesta etapa qualquer resposta positiva resulta no avanço da aplicação. As perguntas são:

1. Existem novos métodos ou ferramentas relevantes para a RSL?

2. Existem novos estudos ou informações relevantes para a RSL?

Para Garner et al. [7], a abordagem mais eficaz para responder a ambas as questões consiste em realizar pesquisas em bases de dados e seguir pesquisadores que estejam ativos no campo. Este método foi aplicado, sendo identificado um conjunto de estudos que satisfaziam os critérios para responder a ambas as questões. Para entender melhor o discurso atual no campo da ética em IA, foram selecionados três estudos desse conjunto, que foram realizados com objetivos e abordagens diferentes.

Os estudos selecionados abordaram três áreas principais: a aplicação de requisitos éticos de IA na indústria [44], a utilização de histórias de usuário para a implementação de requisitos éticos [8] e a descrição de uma nova ferramenta estado da arte [50]. Ficou clara a utilização em diversos setores e objetivos variados dentro da temática de ética em IA. Além disso, o campo está passando por um avanço, presenciando o desenvolvimento de novas ferramentas e métodos de estado da arte, bem como estudos destinados a explorar novas conexões com tópicos específicos de outros campos. Isto responde positivamente a ambas as perguntas e permite o avanço a terceira e última fase.

3.1.3 Etapa 3

A terceira e última etapa é composta de duas perguntas, como pode ser observado na Figura 3.1. Se a resposta for positiva ao final desta etapa, a atualização é iniciada. As perguntas são:

1. A adoção de novos métodos ou ferramentas alterará o resultado ou credibilidade da RSL?

2. A adoção de novos estudos, informações ou dados alterará o resultado ou credibilidade da RSL?

Conforme apresentado na definição da terceira etapa 3.1, alterar ou acrescentar critérios de seleção, questões de verificação de qualidade ou questões de pesquisa resultam em resposta positiva. Nesta atualização, os critérios de seleção e as perguntas de avaliação de qualidade foram submetidos a um processo de modificação e aperfeiçoamento, concluindo assim o processo de três fases. As modificações e aperfeiçoamentos foram implementados por meio da reescrita de todos os critérios, que foram depois sujeitos a revisão durante a calibração da *string* de busca. Em seguida, foram incorporados no processo de pesquisa para obter os estudos relevantes. De acordo com o processo de atualização, a seção seguinte define o protocolo para a atualização da RSL.

3.2 Questões de pesquisa

A motivação desta RSL é identificar estudos, princípios, diretrizes, métodos, metodologias, processos, técnicas, *frameworks* e ferramentas do estado-da-arte de ética em IA. Além disso, é necessário entender como as ferramentas podem ser aplicadas na prática durante o processo de desenvolvimento de software. Portanto, na Tabela 3.1 podem ser observadas as questões de pesquisa definidas.

3.3 *String* de Busca

A pesquisa foi realizada de duas maneiras diferentes: uma digital e uma manual. Buscas digitais são realizadas utilizando uma *string* de busca automatizada, sendo definida como uma *string* específica que inclui a divisão da pergunta em elementos individuais e os seus respectivos sinônimos, siglas e escritas alternativas. Uma busca mais refinada pode ser construída utilizando operadores booleanos, especificamente a função AND e OR [51]. Já a busca manual pode ser realizada seguindo autores específicos, listas de referências de

Tabela 3.1: Questões de pesquisa e motivações

ID	Questão de pesquisa	Motivação
RQ.1	Quais princípios e diretrizes existem na literatura e indústria no contexto de ética de IA?	Essa questão de pesquisa tem como objetivo identificar os princípios e as diretrizes existentes no contexto de ética em inteligência artificial para entender o estado atual do tópico na literatura e indústria.
RQ.2	Quais são as técnicas, metodologias, métodos, *frameworks*, ferramentas e processos existentes na literatura para apoiar a operacionalização dos requisitos éticos de IA?	Essa questão de pesquisa tem como objetivo identificar as técnicas, metodologias, métodos, *frameworks*, ferramentas e processos existentes na literatura para apoio a operacionalização de requisitos éticos de IA para entender o estado atual e encontrar lacunas que podem ser preenchidas por propostas futuras.
RQ.3	Como possibilitar a implementação de princípios éticos de IA durante o ciclo de desenvolvimento de software?	Essa pergunta tem como objetivo identificar meios para viabilizar a implementação de princípios éticos de IA durante o ciclo de vida do produto para guiar o desenvolvimento de novas ferramentas.

Fonte: o Autor

estudos primários relevantes e artigos de revisão, revistas, anais de conferências, a internet e registos de pesquisa [51].

A busca manual foi realizada em paralelo com a busca digital, tendo sido seguidos autores com publicações significativas na área, como Abrahamsson[1] e Vakkuri[2]. Adicionalmente, foram realizadas buscas manuais nas bases de dados para identificar estudos recentes. As bases de dados digitais escolhidas foram: ACM Digital Library, IEEE Xplore, e DBLP. Estas bases de dados foram escolhidas por possuírem uma sólida coleção internacional de publicações relacionadas à ciência da computação, indexando um grande número de conferências e periódicos [52].

Como dito anteriormente, é necessário o desenvolvimento de uma *string* de busca automatizada para iniciar a busca digital. A base para este desenvolvimento foi a *string* proposta por Cerqueira [1]. Foi realizada uma série de iterações na base de dados ACM utilizando três estudos de controlo com o objetivo de adaptar a *string* juntamente com os critérios de seleção e avaliação da qualidade. Os estudos utilizados foram Pekka et al. [44], Cerqueira et al. [15] e Vakkuri et al. [53]. Após este passo, a palavra-chave **user stories** foi adicionada à cadeia de pesquisa como distinção de Cerqueira [1] e os critérios de seleção e qualidade foram aprovados. A *string* de busca automatizada pode ser observada na Tabela 3.2. Já as *strings* utilizadas em cada base são apresentadas na Tabela 3.3.

[1] https://scholar.google.it/citations?user=A-CX3y4AAAAJ&hl=en
[2] https://scholar.google.com/citations?user=ra_NUagAAAAJ&hl=fi

Tabela 3.2: *String* de busca base

***String* base**	(“ethic” OR “ethics” OR “ethical” OR “ethically” OR “applied ethics” OR “ethical values” OR “responsible ai” OR “ai ethics”) AND (“design” OR “development” OR “governance” OR “method” OR “framework” OR “tool” OR “process” OR “implementing” OR “implementation” OR “practices” OR “guidelines” OR “principles” OR “user stories”) AND (“artificial intelligence” OR “machine learning” OR “AI” OR “ML”)

Fonte: o Autor

Tabela 3.3: *String* de busca para cada base digital

Base Digital	***String* de Busca**
ACM Digital Library	[[Abstract: ethic] OR [Abstract: ethics] OR [Abstract: ethical] OR [Abstract: “applied ethics”] OR [Abstract: “ethical values”] OR [Abstract: “responsible ai”] OR [Abstract: “ai ethics”]] AND [[Abstract: design] OR [Abstract: development] OR [Abstract: governance] OR [Abstract: method] OR [Abstract: framework] OR [Abstract: tool] OR [Abstract: process] OR [Abstract: implementing] OR [Abstract: implementation] OR [Abstract: practices] OR [Abstract: guidelines] OR [Abstract: principles] OR [Abstract: ‘user stories”]] AND [[Abstract: “artificial intelligence’] OR [Abstract: “machine learning’] OR [Abstract: AI] OR [Abstract: ML]] AND [Publication Date: (01/01/2021 TO *)]
IEEE Xplore	((“Abstract”:ethic OR “Abstract”:ethics OR “Abstract”:ethical OR “Abstract”:“applied ethics” OR “Abstract”:“ethical values” OR “Abstract”:“responsible ai” OR “Abstract”:“‘ai ethics”) AND (“Abstract”:design OR “Abstract”:development OR “Abstract”:governance OR “Abstract”:method OR “Abstract”:framework OR “Abstract”:tool OR “Abstract”:process OR “Abstract”:implementing OR “Abstract”:implementation OR “Abstract”:practices OR “Abstract”:guidelines OR “Abstract”:principles OR “Abstract”: “user stories”) AND (“Abstract”:“artificial intelligence” OR “Abstract”:“machine learning” OR “Abstract”:AI OR “Abstract”:ML))
DBLP	(“ethic” OR “ethics” OR “ethical” OR “ethically” OR “applied ethics” OR “ethical values” OR “responsible ai” OR “ai ethics”) AND (“design” OR “development” OR “governance” OR “method” OR “framework” OR “tool” OR “process” OR “implementing” OR “implementation” OR “practices” OR “guidelines” OR “principles” OR “user stories”) AND (“artificial intelligence” OR “machine learning” OR “AI” OR “ML”)

Fonte: o Autor

Para a *string* de busca das bibliotecas digitais ACM e IEEE Xplore foi utilizada a opção de realizar pesquisa pelo *abstract* em detrimento ao título ou conjunto completo

dos metadados, pois os resultados foram mais concisos e com menor número de falsos positivos. Na base DBLP foi utilizada a *string* genérica. Por fim, nas bases IEEE Xplore e DBLP, foi aplicado o filtro manual sobre a data de publicação.

3.4 Critérios de Seleção

Kitchen & Charters [51] definem o objetivo dos critérios de seleção de estudos como a identificação de estudos primários que fornecem provas diretas relativas à questão de investigação. Além disso, os autores propõem que os critérios de seleção sejam divididos em duas categorias diferentes: critérios de inclusão (IC) e critérios de exclusão (EC). Os dois tipos de critérios são baseados nas questões de pesquisa, mas servem propósitos diferentes. Os critérios de inclusão definem as caraterísticas que um estudo deve reunir para ser selecionado como estudo primário de uma determinada revisão sistemática de literatura. Em contraste, o principal objetivo dos critérios de exclusão é excluir os estudos que satisfazem os critérios de inclusão, mas que não apresentam os aspectos desejados para a revisão. Cinco critérios de inclusão foram definidos para esta revisão, sendo eles:

(IC1) O estudo é um artigo publicado como *full paper*;

(IC2) O texto completo está disponível para acesso;

(IC3) O estudo deve abordar questões práticas em Inteligência Artificial ou *Machine Learning*;

(IC4) O estudo deve ter sido publicado após abril de 2021;

(IC5) O estudo deve ser relacionado à ética em Inteligência Artificial por meio de técnicas, princípios, diretrizes, métodos, processos, *frameworks* ou ferramentas.

Três critérios de exclusão foram definidos para esta revisão, sendo eles:

(EC1) Não discute o tópico de ética em Inteligência Artificial;

(EC2) O estudo não é escrito em Inglês ou Português;

(EC3) Aborda a questão em um contexto diferente do objeto de estudo (por exemplo: *Artificial Moral Agents*).

Os critérios IC1, IC2, IC4, IC5, EC2 e EC3 foram definidos em contraste com os utilizados por Cerqueira [1]. A definição resultou da necessidade de formalização no contexto deste estudo, dada a necessidade de basear a classificação dos estudos primários nos critérios de seleção. Isto é particularmente evidente em critérios como IC1 e IC4, que

foram definidos para explicar a lógica subjacente à remoção de artigos que não tenham sido publicados como um artigo completo ou que tenham sido publicados antes da data especificada.

3.5 *Quality Assessment*

Mesmo com a filtragem de estudos feita pela utilização dos critérios de seleção, é necessário ter critérios para avaliar a qualidade dos estudos, sua validade e se os impactos foram medidos de maneira correta. De acordo com Kitchenham et al. [51], *Quality Assessment* é sobre determinar até que ponto os resultados de um estudo empírico são válidos e livres de enviesamento. Para isso, foi definida a seguinte lista de perguntas para avaliar a qualidade dos estudos:

(QA1) Os resultados são claros e relevantes?

(QA2) O estudo possui alguma limitação ou ameaça à validade?

(QA3) A metodologia é adequada e bem definida?

(QA4) Os objetivos são bem definidos e relevantes?

(QA5) O estudo é baseado em pesquisas (ou é apenas um relatório do tipo "lições aprendidas" baseada em opinião de especialistas)?

(QA6) O resultado apresentado possui valor científico no campo de ética em inteligência artificial?

Este trabalho se propõe em analisar apenas resultados medidos de maneira cientificamente correta, visto que é diretamente ligado ao impacto da tecnologia nos usuários, portanto as questões QA1 à QA5 tratam do rigor científico dos estudos. Já QA6 tem como objetivo atestar a utilização e aplicabilidade do estudo no campo de interesse deste trabalho.

A classificação dos estudos foi realizada conforme os critérios utilizados por Mendes et al. [54], na qual cada pergunta pode ser respondida com "sim" (adiciona 1.0 ponto no valor final), "parcialmente" (adiciona 0.5 pontos no valor final) e "não" (não altera o valor final). Se o estudo tiver o resultado final menor que 2.0, ele é desclassificado. O resultado final encontra-se na Seção 3.9.

3.5.1 Processo de Seleção

Kitchen e Charters [51] propuseram uma abordagem em duas fases para o processo de seleção, mas sugeriram também que uma terceira fase poderia ser útil para garantir o rigor. Nesta revisão, foi utilizada a metodologia de três fases. Estas fases são:

Primeira fase Durante esta fase, todos os estudos identificados por meio do processo de pesquisa foram submetidos a uma avaliação rigorosa para determinar se podiam ser excluídos com confiança. Isto foi feito por meio da análise e aplicação dos critérios de seleção no título, resumo e palavras-chave.

Segunda fase Nesta fase, foi realizada uma análise mais completa de cada um dos estudos selecionados na fase anterior. Esta análise envolveu um maior aprofundamento dos artigos, com a exclusão de alguns com base na aplicação dos critérios de seleção a diferentes aspectos, como a metodologia de investigação, as conclusões e o método de amostragem.

Terceira fase A fase final envolveu uma leitura exaustiva de cada estudo para responder a todas as questões de avaliação da qualidade, quantificar os resultados e excluir os estudos que não atingiam a pontuação miníma.

3.6 Formulário de Extração de Dados

A Tabela 3.4 contém o *template* para o formulário de extração de dados. Cada item está conectado a uma questão de pesquisa (RQ), se possível. Para auxiliar o processo foi utilizada a ferramenta Parsifal[3] e por fim, os dados foram convertidos para formato de planilha, possibilitando a criação de gráficos para a visualização de resultados.

[3]https://parsif.al/

Tabela 3.4: *Template* Formulário de Extração de Dados

Item	**Valor**	**RQ**
ID Estudo	-	-
Fonte	IEEE Xplore, ACM Digital Library ou DBLP	-
Título de publicação	-	-
Autor(es)	-	-
Data de publicação	-	-
Palavras-chave	-	-
Número de citações	-	-
Tipo de questões de pesquisa (Tipo de estudo)	Classificação de questões de pesquisa em artigos de Engenharia de Software por Shaw [55]: método para desenvolvimento, método para análise, análise/avaliação de instância, generalização/caracterização e estudo de viabilidade	-
Questões de pesquisa discutidas no estudo	-	-
Proposta do estudo	técnica, metodologia, método, *framework*, ferramenta, processo, diretriz ou princípio	RQ1 e RQ3
Contexto do estudo	indústria, academia ou ilustração do autor	-
Etapa de Engenharia de Software	SWEBOK define as principais etapas dentro do contexto de Engenharia de Software como [56]: *planning*, *software requirements*; *software design*; *software construction*; *software testing*; *software maintenance and deploy*	RQ2
Recomendações ou achados principais do estudo	-	RQ1, RQ2 e RQ3
Limitações do estudo	-	-

Fonte: o Autor

3.7 Compilação do Protocolo

Ao finalizar a etapa de planejamento o protocolo foi definido, como pode ser observado na Figura 3.2. Para a criação do protocolo foi utilizada a definição de cada etapa juntamente com as suas respectivas tarefas proposta por Kitchenham et al. [51].

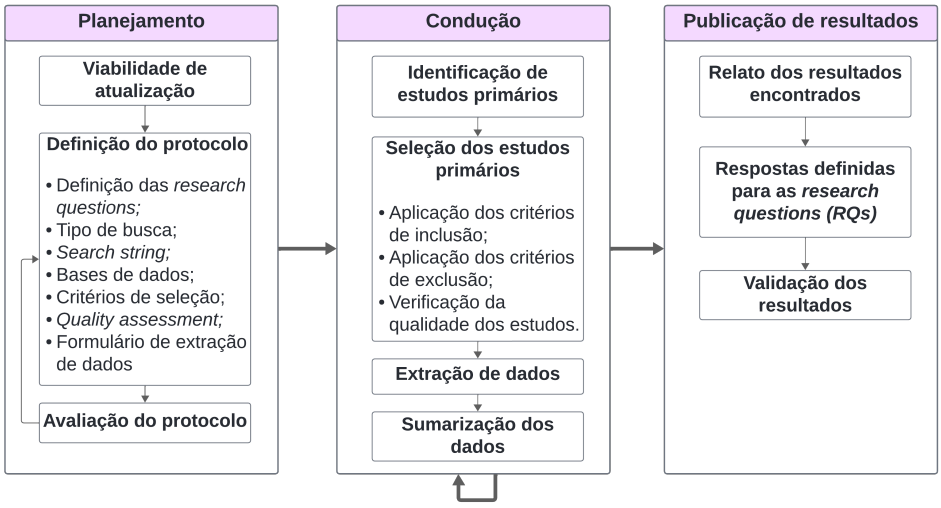


Figura 3.2: Protocolo da atualização da RSL. Fonte: o Autor.

3.8 Condução da RSL

Para a condução da RSL foi utilizado o Parsifal[4], uma ferramenta web, gratuita e *open-source* para apoio à RSL. A ferramenta foi escolhida por facilitar a condução e por ter suas funcionalidades e fluxos baseados nos processos de RSL propostos por Kitchenham et al. [51], que são utilizados neste trabalho. Na Figura 3.4 é apresentada à direita a quantidade de artigos resultantes de cada etapa. A Figura 3.3 mostra o número de artigos por ano.

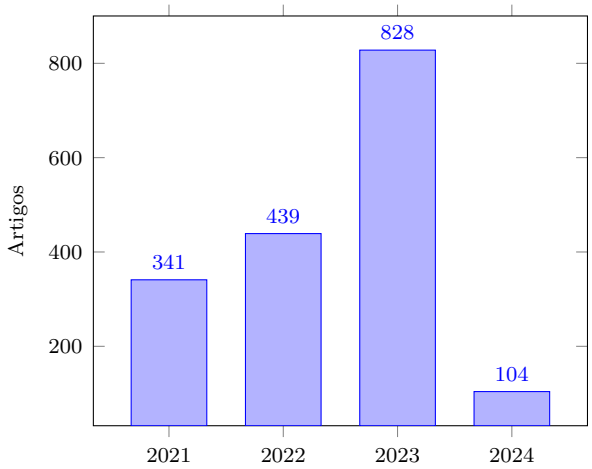


Figura 3.3: Número de artigos por ano. Fonte: o Autor.

[4]https://parsif.al/

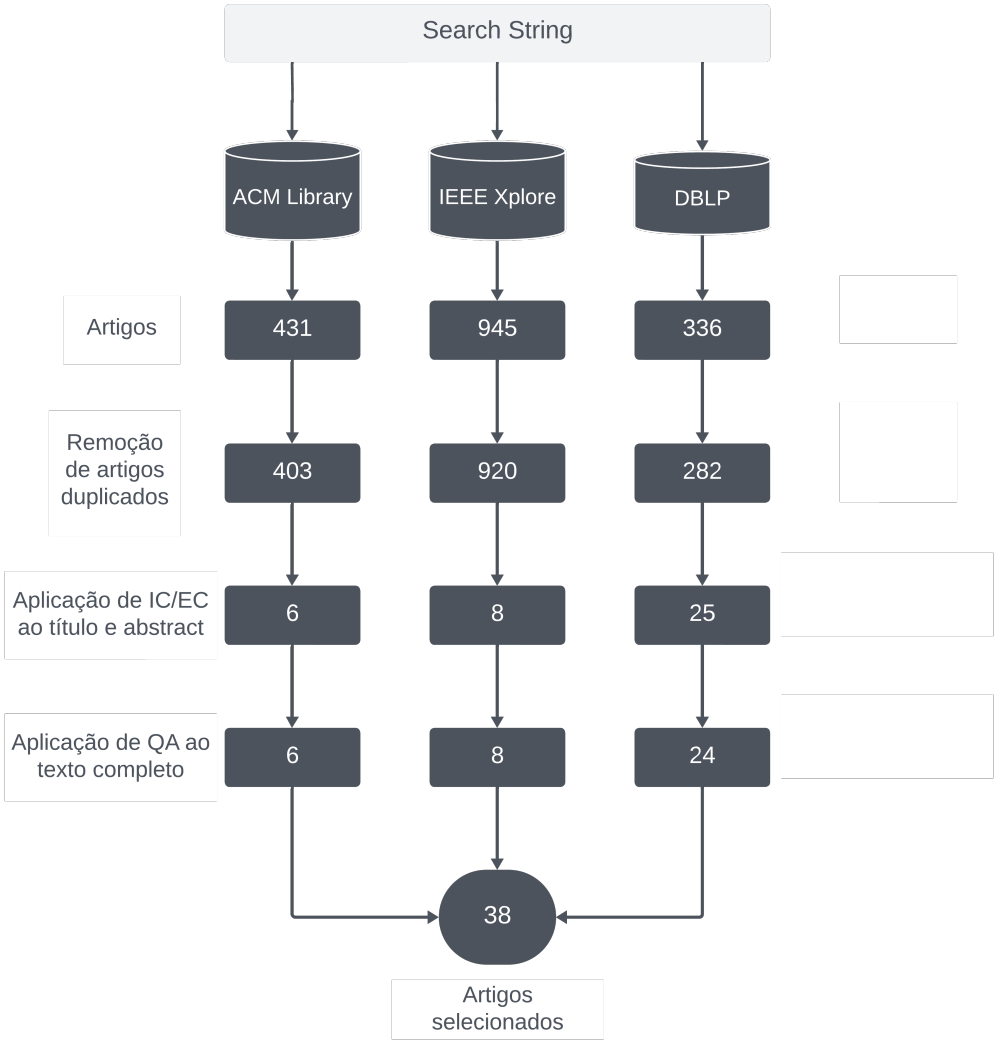


Figura 3.4: Filtragem dos estudos por etapa. Fonte: o Autor.

Os artigos foram coletados até maio de 2024, resultando em um total de 1712 artigos (431 da ACM Digital Library, 945 do IEEE Xplore e 336 da DBLP), conforme apresentado na Figura 3.4. Para dar continuidade a seleção, foram removidos 107 artigos duplicados utilizando a funcão de remoção de duplicatas da ferramenta Parsifal[5]. Isto resultou em 1605 artigos originais para a aplicação dos critérios de inclusão e exclusão por meio do título e *abstract*. Nesta etapa foram excluídos 1566 artigos por violação de critérios de seleção ou fuga do tópico após a leitura do título e *abstract*, resultando em 39 artigos para a verificação de qualidade (6 da ACM Digital Library, 8 do IEEE Xplore e 25 da DBLP). Finalmente foram feitas análises mais profundas, com a leitura completa dos artigos e verificação de qualidade, resultando na remoção de um artigo e 38 estudos primários (6 da ACM Digital Library, 8 do IEEE Xplore e 24 da DBLP). A Figura 3.5 apresenta a

[5]https://parsif.al/

quantidade de artigos identificados por base de dados digital no início das buscas e os selecionados após a finalização da verificação de qualidade.

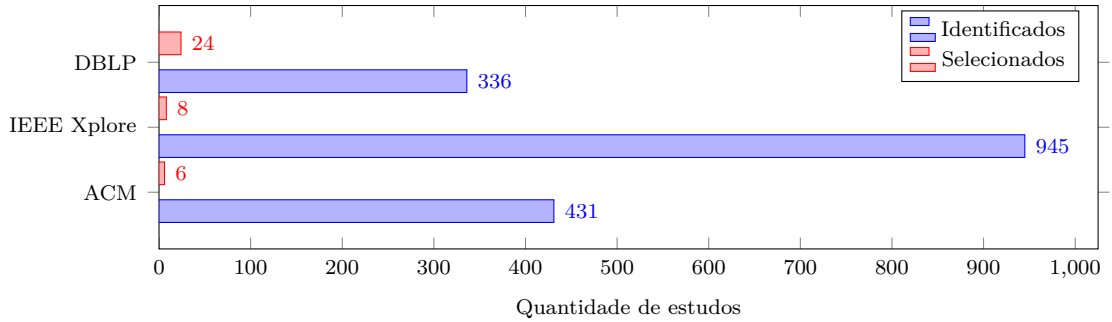


Figura 3.5: Quantidade de estudos identificados e selecionados por base de dados digital. Fonte: o Autor.

3.9 Resultados da RSL

Os 38 estudos primários selecionados são apresentados na Tabela 3.5. A coluna ID representa o identificador dos estudos. A coluna título representa o título dos estudos. A coluna proposta representa o resultado proposto de cada estudo, quer seja metodologia, método, *framework*, ferramenta, princípio, processo ou análise/generalização de propostas. A coluna de fase do ciclo de vida representa em quais etapas a proposta pode ser utilizada com base nas definições do SWEBOK [56]. Por fim, a coluna de RQ representa quais questões de pesquisa o estudo ajuda a responder.

Tabela 3.5: Estudos primários

ID	**Título**	**Proposta**	**Fase do ciclo da vida**	**QA**	**RQ**
S1	A Deployment Model to Extend Ethically Aligned AI Implementation Method ECCOLA	metodologia	*software requirements*; *software design*; *software construction*;	4.0	2
S2	From ethical AI frameworks to tools: a review of approaches	análise	*planning*; *software requirements*; *software design*; *software construction*; *software testing*; *software maintenance and deploy*	5.5	3
S3	A framework for assessing AI ethics with applications to cybersecurity	*framework*	*software requirements*	3.5	1 e 2
S4	All that glitters is not gold: trustworthy and ethical AI principles	*framework*	*software requirements*	3.5	2

S5	A Rapid Review of Responsible AI frameworks: How to guide the development of ethical AI	análise	-	4.0	3
S6	AI Ethics Impact Assessment based on Requirement Engineering	metodologia	*software requirements*	5.0	2
S7	Z-Inspection: A Process to Assess Trustworthy AI	processo	*planning*; *software maintenance and deploy*	5.0	2
S8	Implementing AI Ethics: Making Sense of the Ethical Requirements	princípio	-	2.5	1
S9	How Do Software Companies Deal with Artificial Intelligence Ethics? A Gap Analysis	análise	-	4.5	1
S10	Governance of Ethical and Trustworthy AI Systems: Research Gaps in the ECCOLA Method	método	-	4.5	2
S11	Ethical Requirements Stack: A framework for implementing ethical requirements of AI in software engineering practices	*framework*	*software requirements*	3.0	2 e 3
S12	Implementing AI Ethics in a Software Engineering Project-Based Learning Environment - The Case of WIMMA Lab	*framework*	*planning*; *software requirements*; *software design*; *software construction*; *software testing*; *software maintenance and deploy*	2.5	2
S13	AI Ethics Principles in Practice: Perspectives of Designers and Developers	processo	*planning*; *software requirements*; *software design*; *software construction*	5.5	2
S14	Beyond 100 Ethical Concerns in the Development of Robot-to-Robot Cooperation	*framework*	-	3.5	2
S15	What Is the Cost of AI Ethics? Initial Conceptual Framework and Empirical Insights	*framework*	*planning*; *software requirements*; *software design*; *software construction*	5.0	2
S16	Ethics of AI: A systematic literature review of principles and challenges	análise	-	3.5	1
S17	Ethics by Design for Intelligent and Sustainable Adaptive Systems	*framework*	*software construction*	5.5	2

S18	Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development	ferramenta	*planning*; *software requirements*; *software design*; *software construction*; *software testing*; *software maintenance and deploy*	6.0	2 e 3
S19	A participatory data-centric approach to AI Ethics by Design	processo	*planning*; *software requirements*; *software design*; *software testing*; *software maintenance and deploy*	5.5	2
S20	AI Ethics for Industry 5.0 – From Principles to Practice	*framework*	*software construction*; *software testing*; *software maintenance and deploy*	3.0	2 e 3
S21	Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance	metodologia	*planning*; *software requirements*; *software design*; *software construction*; *software testing*; *software maintenance and deploy*	5.0	2
S22	How Do AI Ethics Principles Work? From Process to Product Point of View	análise	-	4.0	1 e 3
S23	The Different Faces of AI Ethics Across the World: A Principle-to-Practice Gap Analysis	análise	-	5.0	3
S24	Measuring Ethics in AI with AI: A Methodology and Dataset Construction	ferramenta	*software construction*	4.0	-
S25	Governance of Artificial Intelligence – A Framework Towards Ethical AI Applications	*framework*	*planning*; *software requirements*; *software design*; *software construction*; *software testing*; *software maintenance and deploy*	3.5	2
S26	Worldwide AI Ethics: a review of 200 guidelines and recommendationsfor AI governance	ferramenta	-	5.5	1
S27	ECCOLA — A method for implementing ethically aligned AI systems	método	*planning*; *software requirements*; *software design*; *software construction*; *software testing*; *software maintenance and deploy*	5.0	2 e 3

S28	The Importance of an Ethical Framework for Trust Calibration in AI	*framework*	*planning*; *software requirements*; *software design*; *software construction*; *software testing*; *software maintenance and deploy*	4.0	2
S29	The Role of Explainable AI in the Research Field of AI Ethics	análise	-	4.0	1
S30	Towards a Roadmap on Software Engineering for Responsible AI	análise	-	5.0	3
S31	Governance in Ethical, Trustworthy AI Systems: Extension of the ECCOLA Method for AI Ethics Governance Using GARP	método	*planning*; *software requirements*; *software design*; *software construction*; *software testing*; *software maintenance and deploy*	5.0	2
S32	Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems	*framework*	*software construction*	4.5	2
S33	From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience	análise	*planning*; *software requirements*; *software design*; *software construction*; *software testing*; *software maintenance and deploy*	3.5	3
S34	Fairness in Design: A Tool for Guidance in Ethical Artificial Intelligence Design	*framework*	*software requirements*	4.0	2 e 3
S35	Putting AI ethics to work: are the tools fit for purpose?	análise	-	4.5	2
S36	Transparency and explainability of AI systems: From ethical guidelines to requirements	análise	*software requirements*	5.0	1 e 3
S37	Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering	processo	*planning*; *software requirements*; software design; *software construction*; *software testing*; *software maintenance and deploy*	5.0	2
S38	Ethical Guidelines and Principles in the Context of Artificial Intelligence	análise	-	4.0	1

3.9.1 RQ1 - Quais princípios e diretrizes existem na literatura e indústria no contexto de ética em IA?

Diferentes estudos têm como objetivo a caracterização e generalização dos principais princípios e requisitos de ética em IA para apoiar a criação de ferramentas de apoio universais. Para Rees & Muller [57] os principais princípios e requisitos incluem: robustez, licitude, equidade, responsabilidade, segurança e impacto, diversidade e legalidade.

Khan et al. [58] identificaram um conjunto de 21 princípios e suas frequências após a RSL, sendo eles: transparência (17), privacidade (16), responsabilização (15), equidade (14), autonomia (10), explicabilidade (8), justiça (7), não maleficência (7), dignidade humana (6), beneficência (6), responsabilidade (5), segurança (5), segurança de dados (4), sustentabilidade (2), liberdade (1), solidariedade (1), prosperidade (1), efetividade (1), acurácia (1), predizibilidade (1) e interpretabilidade (1). Após a definição dos princípios, Khan et al. [58] definiram os principais desafios da aplicação dos princípios em sistemas reais: falta de conhecimento ético, princípios vagos, princípios muito generalizados, conflitos com a prática, interpretações diferentes de princípios, falta de entendimento técnico e conjunto extra de restrições.

Para entender como princípios e requisitos são utilizados nas organizações, Vakkuri et al. [59] coletaram dados por meio de um *survey* em 249 empresas. As respostas foram analisadas qualitativamente para definir as contribuições empíricas primárias dos dados utilizando as diretrizes de ética em IA da União Europeia. Os autores identificaram um conjunto de requisitos relevantes para a academia mas que não foram discutidos nos dados coletados, sendo eles: bem estar social, diversidade, equidade e não discriminação. Os requisitos de supervisão humana não são abordados de maneira direta, mas certas práticas contribuem com o seu cumprimento. Já os requisitos de robustez técnica e segurança são normalmente abordados por meio de métodos comuns de desenvolvimento de software, testagem e validação. Regulações governamentais definidas são as principais maneiras de abordagem de requisitos como privacidade e governança de dados. Por fim, os requisitos de transparência e responsabilização são abordados com uma visão muito simplista, pois são entendidas como questões entre apenas duas partes.

Balasubramaniam et al. [60] seguiram a mesma linha de pesquisa de Vakkuri et al. [59], aplicando um questionário em dezesseis organizações para entender quais requisitos e princípios tem posição de destaque na indústria. Após a aplicação do questionário, foram identificados dois requisitos principais, transparência e explicabilidade, presentes em 87.5% das organizações. De acordo com as organizações, os dois principais motivos para a importância destes requisitos são: a construção e manutenção de confiança dos usuários e o apoio a segurança que a transparência traz.

Kemell et al. [38] alteraram a visão para entender o ponto de vista do produto em relação a ética em IA. Foi criado um sistema fictício focado no bem-estar e produtividade de uma organização. Para a análise foi selecionado um conjunto de princípios e objetivos do sistema, abrindo espaço para a discussão sobre quais as aplicações de cada princípio. Como resultado deste experimento mental foi criado um grupo contendo características e dificuldades relacionadas a implementação:

- Potencial para sobrecarga de informação devido a quantidade de princípios e conceitos relacionados.
- Sobreposição de princípios, como por exemplo, violações de privacidade estarem contidas em não maleficência.
- Conflito de princípios durante a implementação.
- Existência de princípios específicos de contexto, como solidariedade, que não se encaixa no sistema proposto.
- Relevância de problemas a nível global, visto que assuntos relevantes a nível global como promoção da paz, distribuição de riquezas e responsabilização de uso militar talvez não devam ser abordados em ferramentas focadas na implementação de ética em IA.
- Forte relação entre ética e qualidade no contexto de desenvolvimento de software, como por exemplo a ideia de predizibilidade, conceito abordado nas duas áreas.
- Importância da comunicação, de acordo com diretrizes e princípios, para solução de conflitos e problemas encontrados durante a implementação de requisitos éticos.

Agbese et al. [44] entrevistaram de maneira semi-estruturada dez executivos de engenharia de software Finlandeses para entender como profissionais de gerência entendem, avaliam e implementam os requisitos éticos. Após a realização das entrevistas e análise dos dados coletados, foram definidas cinco contribuições empíricas primárias (PECs), sendo elas:

- A gerência de engenharia de software tem um entendimento fragmentado de requisitos éticos.
- Requisitos éticos possuem valor como requisitos técnicos e regulatórios mas não valor financeiro.
- Requisitos éticos tendem a ser implementados como requisitos legais.

- O valor dos requisitos éticos pode ser aprimorado se os mesmos forem tratados como requisitos de riscos éticos e iniciativas de sustentabilidade financialmente viáveis.
- Requisitos éticos podem ser implementados como requisitos de riscos éticos e sustentabilidade utilizando *frameworks* específicos.

Para identificar o estado da arte de requisitos éticos, Corrêa et al. [9] conduziram uma análise quantitativa e qualitativa de 200 documentos originários de 37 países. Desta análise, foi desenvolvida uma ferramenta de visualização de dados para facilitar a compreensão dos resultados. Na análise quantitativa foi definido um conjunto de princípios mais citados, que é composto por transparência, segurança, confiabilidade, equidade, privacidade, responsabilidade, autonomia, diversidade e não maleficência. Já o conjunto de tipos de instituições que publicam artigos sobre o tema é composto por instituições governamentais, corporações privadas, ONGs, organizações sem fim lucrativos, academia, organizações internacionais, associação de profissionais e associação de indústria.

A análise qualitativa trouxe certos receios em relação ao futuro do tema, como a descoberta que o maior investimento em pesquisas na área é advindo do setor privado, muitas vezes até atrelado a instituições governamentais, abrindo a possibilidade para autoregulação. Outro ponto importante se relaciona com o princípio de diversidade, abordado em diversas diretrizes, mas não é posto em prática, visto que foi encontrado pelos autores que mais de 63% dos artigos é escrito por pessoas que se identificam ao gênero masculino [9].

Os problemas éticos se intensificam ao analisar casos como o sistema de software para apoio a decisões da corte norte americana COMPAS, no qual pessoas negras têm quase duas vezes mais probabilidade de serem rotulados como possíveis reincidentes pela IA [12]. Por meio das considerações acima é identificada a necessidade de compreensão dos sistemas de IA por todo o conjunto de *stakeholders* e do investimento em regulamentações governamentais com o objetivo de abordar possíveis riscos e falhas que afetam, em sua maioria, os mais socialmente vulneráveis.

O termo *Explainable AI* (XAI) se refere a um sistema de IA que pode explicar suas decisões. Para entender como XAI se relaciona com o campo de ética em IA, Vainio-Pekka et al. [61] realizaram um mapeamento sistemático da literatura. Os autores identificaram cinco achados empíricos primários que correspondem ou contradizem as implicações teóricas, sendo eles:

- XAI é um tópico relevante no campo de pesquisa de ética em IA, constituindo 28% dos artigos de pesquisa empírica publicados após 2012.
- Dentro do campo de estudos de XAI, o tipo mais comum de pesquisa empírica é o estudo de novas técnicas que podem solucionar um desafio computacional.

- A perspectiva humana em relação a XAI não é conhecida, não existindo um conhecimento profundo das expectativas de desenvolvedores e usuários sobre os sistemas.
- A implementação de XAI não é estudada com esmero no campo de ética em IA. Portanto, existe uma lacuna de pesquisa na perspectiva gerencial e implicações de negócio envolvendo XAI.
- Pesquisadores de XAI estão mais preocupados em resolver problemas e aplicações do mundo real, em vez de se concentrarem apenas nos aspectos técnicos.

O conjunto final inclui 26 princípios éticos identificados nos estudos. Destes 26 princípios, a transparência (n=4), a equidade (n=4) e a responsabilidade (n=4) foram os mais mencionados. A tabela 3.6 apresenta os 26 princípios éticos identificados na literatura, acompanhados das referências dos estudos relacionados com cada princípio.

Princípios

Foi observado que a um conjunto de princípios éticos, tais como transparência e responsabilidade, foi atribuída uma prioridade mais alta em relação a outros [60]. A implementação desta estratégia de priorização pode ter efeitos adversos se não for realizada corretamente, podendo resultar na negligência de determinados princípios [9]. Nos parágrafos seguintes, serão apresentados os princípios mais frequentes, as suas definições e caraterísticas que prevalecem atualmente no domínio da ética da IA.

Transparência A transparência é uma questão importante no desenvolvimento de sistemas baseados em IA, especialmente em aplicações práticas, e é o princípio mais discutido [9, 58, 59, 60]. Pode ser dito que o princípio está intimamente ligado à confiabilidade e também explicabilidade [60, 61]. O princípio pode ser dividido em duas categorias: transparência técnica e transparência operacional [58]. A primeira diz respeito à transparência do processo técnico, enquanto a segunda está relacionada com a transparência das operações do sistema de IA [58]. Esta distinção permite a aplicação do princípio em todo o processo de desenvolvimento e utilização de sistemas de IA, facilitando uma compreensão mais aprofundada do sistema entre os usuários finais e *stakeholders*, o que, por sua vez, aumenta a confiabilidade do sistema [58]. Transparência operacional é um aspeto crucial dos sistemas de IA, dado o potencial para a tomada de decisões tendenciosas [9]. Para garantir o resultado mais favorável, é também importante ir além da noção simples de transparência. Isto porque o termo é atualmente entendido de uma forma que o apresenta como uma questão apenas entre duas partes, o que constitui uma representação inadequada do conceito [59].

Tabela 3.6: Princípios éticos identificados nos estudos primários

Princípio Ético	**Referência**	**Citado por Cerqueira** [1]?
1. Acurácia	[58]	Não
2. Autonomia	[9, 58]	Sim
3. Bem estar social	[59]	Não
4. Beneficência	[58]	Sim
5. Confiabilidade	[9]	Sim
6. Dignidade humana	[58]	Sim
7. Diversidade	[9, 57, 59]	Não
8. Efetividade	[58]	Não
9. Equidade	[9, 57, 58, 59]	Sim
10. Explicabilidade	[58, 60, 61]	Não
11. Interpretabilidade	[58, 61]	Não
12. Justiça	[58]	Não
13. Legalidade	[57]	Não
14. Não discriminação	[59]	Não
15. Não maleficência	[9, 58]	Sim
16. Predizibilidade	[58]	Não
17. Privacidade	[9, 59]	Sim
18. Prosperidade	[58]	Não
19. Responsabilização	[9, 57, 58, 59]	Sim
20. Robustez	[57, 59]	Não
21. Segurança	[9, 57, 58, 59]	Sim
22. Governança de dados	[57, 58, 59]	Sim
23. Solidariedade	[58]	Sim
24. Supervisão humana	[59]	Sim
25. Sustentabilidade	[58]	Sim
26. Transparência	[9, 58, 59, 60]	Sim

Fonte: o Autor

Equidade Equidade está empatada no primeiro lugar dos princípios mais discutidos na literatura. É um dos princípios mais relevantes no campo da ética da IA, dada a sua correlação com uma série de questões fundamentais, incluindo dados, resultados e implementação [57]. O princípio pode ser agrupado com a não discriminação e a mitigação de vieses. A sua principal caraterística é que, independentemente dos diferentes atributos sensíveis que possam caraterizar um indivíduo, a tomada de decisões algorítmica deve ser conduzida de forma justa e imparcial e não deve iludir as pessoas, prejudicando a sua autonomia [9, 58]. Para isso, os desenvolvedores devem garantir que o processo de tomada de decisões dos sistemas de IA seja mais transparente, com uma identificação clara das entidades responsáveis [58]. O princípio em questão foi classificado de diferentes formas por várias diretrizes. Alguns o classificam como um princípio centrado nos benefícios

relacionados com o uso da IA, enquanto outros o classificam como um princípio centrado nos resultados relacionados com o processo de tomada de decisões dos sistemas de IA [9]. Isto pode resultar em benefícios limitados ao determinar a forma como o princípio pode ser aderido e regulado [57].

Responsabilização ou *Accountability* Os princípios de responsabilidade e prestação de contas (*accountability*) são utilizados para descrever a responsabilidade e a integridade das diferentes partes envolvidas nas operações de IA [61]. Além disso, pode ser definido como um princípio concebido para garantir a justiça, responsabilizando os indivíduos e evitando a ocorrência de danos adicionais [58]. A fim de garantir a responsabilização e a prestação de contas de um sistema de IA, é essencial que as partes interessadas relevantes sejam responsabilizadas pelas decisões e ações do sistema. Isto ajudará a minimizar eventuais problemas de responsabilidade, uma vez que cada indivíduo deve ser responsabilizado pelos seus respectivos papéis no sistema [58]. A importância deste princípio ético pode ser percebida no caso de um problema, visto que o passo inicial no processo de responsabilização é determinar a quem o sistema é responsável e a extensão e limitações dessa responsabilidade [57].

Uso de Princípios Éticos

A aplicação dos princípios éticos difere entre os contextos acadêmico e empresarial. Vakkuri et al. [59] identificaram um subconjunto de princípios relevantes na academia que não são discutidos nas empresas, incluindo bem-estar social, diversidade, equidade e não discriminação. Este estudo também identificou que os princípios aplicados não são abordados de forma satisfatória, sendo apenas utilizadas práticas comuns de desenvolvimento de software e regulamentação governamental [59]. Balasubramaniam et al. descobriram que as empresas tendem a concentrar-se num subconjunto ainda menor de princípios, consistindo principalmente em transparência, explicabilidade e privacidade para construir e manter a confiança dos usuários [59, 60]. No entanto, parece que mesmo estes princípios não estão a sendo aplicados de maneira apropriada.

Análise

Com base nos resultados da RSL, foi identificado que o objetivo mais frequente era o de examinar os princípios e diretrizes propostos e utilizados em ambientes do mundo real com a intenção de identificar problemas, desafios e caraterísticas distintas, ao invés de desenvolver novos princípios [9, 38, 57, 58, 59, 60, 44]. A totalidade dos princípios discutidos na literatura existente não é aplicada no contexto do desenvolvimento de software do mundo real, resultando numa discrepância entre abordagens teóricas e práticas [59, 60, 61]. Este

fato pode ser atribuído à falta de normalização, a abordagens não específicas e à falta de apoio interno [9, 38, 58, 44]. Embora as orientações e os princípios possam servir como um ponto de partida útil para a discussão, eles não são uma solução abrangente para o desafio da ética da IA sem a incorporação de medidas práticas adicionais, uma vez que há ausência de esforços para adotar estratégias eficazes para implementar princípios éticos [59, 61].

3.9.2 RQ2 - Quais são as técnicas, metodologias, métodos, *frameworks*, ferramentas e processos existentes na literatura para apoiar a operacionalização dos requisitos éticos de IA?

O método mais difundido na literatura é o método ECCOLA [45], que tem como objetivo analisar e compreender possíveis problemas éticos presentes em sistemas de IA. O método é composto de um conjunto de 21 cartas divididas em oito temas. Sete destes temas representam princípios éticos, enquanto o tema restante trata da análise de *stakeholder*. O conjunto de temas é baseados nas diretrizes definidas pelo grupo de especialistas de IA da União Europeia (HLEG AI) e IEEE EADv1. Cada carta é dividida em três partes: motivação (porque é importante), o que fazer (como solucionar o problema) e exemplo prático do tópico (para facilitar a aplicação prática). O método ECCOLA é aplicado por meio de um processo modular, de *sprint* em *sprint*, integrando ética de IA em histórias de usuário.

Rousi et al. [62] analisaram a aplicação do método ECCOLA no contexto de cooperação em sistemas multi-robôs, especificamente com o sistema CACDAR. CACDAR é um projeto que tem como foco a colaboração e cooperação entre robôs diversos, composto de um software de prova de conceito e um processo de desenvolvimento de cooperação três mundos (mundo de bloco em duas dimensões, mundo simulado em três dimensões e mundo real). Para identificar os principais temas éticos emergentes, foi realizado um *workshop* com dez especialistas resultando em doze temas:

- Temas que fazem sobreposição com o método ECCOLA incluem responsabilidade, ineficiência de dados, defeitos, empoderamento, privacidade de dados, garantia de segurança, garantia de proteção e ineficiência humana.
- Temas específicos a sistemas multi-robôs mas que podem ser adições ao método ECCOLA: cooperação, confronto ou incompatibilidade por valores ou lógica.
- Temas vagos ou menos relevantes: garantia de preocupações éticas e design.

Antikainen et al. [63] propuseram um modelo de implantação para o método ECCOLA. Este modelo é composto de quatro processos primários que por sua vez são compostos de subprocessos. Tanto processos e subprocessos são executados de maneira cíclica durante as etapas do ciclo de vida de desenvolvimento de software. O modelo tem como objetivo a priorização de diferentes requisitos éticos, participação de *stakeholders* e aplicação cíclica integrada com metodologias ágeis. Isso possibilita a execução de certas atividades, como novas priorizações e avaliações de desenvolvimento. Além do método, também foi proposta uma ferramenta de apoio de visualização gráfica pelos autores chamada de *Ethical Situational Picture*. A ferramenta permite a identificação de características principais de cada requisito dentro do método. Essas características incluem o consenso entre *stakeholders*, grau de importância, mudanças de priorização e quão bem o requisito foi abordado durante o o processo de desenvolvimento de software.

Agbese et al. [64] realizaram uma análise do método ECCOLA sob a perspectiva de governança de IA. O conceito de governança de IA é baseado nos princípios de transparência, explicabilidade e condutas éticas no que diz respeito a sistemas de IA. Para a análise foi utilizado o *framework* proposto pela Comissão de Proteção de Dados Pessoais em Singapura (PDPC), que classifica as práticas de governança em três categorias: governança de dados, governança da informação e governança corporativa. Por fim, foi identificado que as cartas do método ECCOLA abordam por completo a categoria de governança corporativa. Mas para garantir mais robustez, podem ser atualizadas para incorporar governança de dados e de informações.

Com o objetivo de encontrar e preencher lacunas sob o aspecto de governança de IA, Agbese et al. [65] realizaram um estudo de caso no método ECCOLA, que foi posteriormente submetido a uma análise crítica baseada nas práticas do *framework* de governança de informações GARP. O *framework* GARP pode ser definido como uma estrutura que espelha relações interconectadas, fatores e influências dentro de uma instituição. Ele é composto por oito princípios: responsabilidade, transparência, integridade, proteção, conformidade, disponibilidade, retenção e disposição.

A análise produziu duas conclusões fundamentais. A primeira conclusão identificou os princípios que são abordados pelas cartas do método ECCOLA e os que requerem maior atenção. Na segunda os autores apresentaram duas propostas. A primeira é melhorar a representação de seis princípios por meio de referências a práticas pertinentes em cartas existentes no método. A segunda é a introdução de uma nova carta para os princípios de retenção e disposição [65].

Nitta et al. [66] propuseram uma metodologia de avaliação dos impactos éticos de IA com o objetivo de analisar como e onde riscos éticos podem ocorrer pelo uso de sistemas de IA. A metodologia é dividida em três passos:

- Primeiro passo: criação de um diagrama para o sistema de IA contendo os *stakeholders* relacionados a cada componente (dados, modelos, entre outros).
- Segundo passo: extração das características éticas correspondentes a cada interação do sistema utilizando um modelo de ética em IA. Este passo pode ser realizado mecanicamente utilizando as interações extraídas no primeiro passo como entradas.
- Terceiro passo: extração e apresentação em formato de tabela de cada situação contrária as características éticas como um possível fator ou evento de risco.

Utilizando a estrutura de *Value Sensitive Design* (VSD), Boyd [67] propôs um guia de campo para o desenvolvimento de *machine learning* (ML). VSD é uma estrutura de design que utiliza métodos interativos conceituais, técnicos e empíricos pra desenvolver designs que refletem os valores dos *stakeholders* principais de um sistema. O guia possui quatro objetivos principais, sendo eles:

- Aumentar a capacidade do usuário de reconhecer e particularizar mitigações técnicas de problemas éticos conhecidos nos dados de treino.
- Melhorar a percepção total de intervenções técnicas, tanto novas como já existentes.
- Empoderar treinadores, educadores e líderes em ML com informações estruturadas e reestruturáveis sobre intervenções técnicas novas e já existentes relacionadas aos dados de treino.
- Cumprir os objetivos acima minimizando as interrupções nas práticas de engenheiros de ML.

Para identificar as necessidades do guia, foram realizadas quatro etapas com especialistas. Estes incluem entrevistas, particularizações pessoais sem a ferramenta, particularizações com o rascunho da ferramenta e particularizações com o *toolkit*. A integração dos objetivos juntamente com a participação dos especialistas culminou na finalização do guia de campo[6] [67].

Bruschi & Diomede [68] salientaram a necessidade da utilização de uma ferramenta de avaliação de riscos para delimitar ética em IA. Eles destacam que princípios podem estar em conflito, sendo crucial quantificar e qualificar o nível de oposição. Para suprir essa lacuna, os autores propuseram um *framework* de avaliação ética composta por uma matriz tridimensional. Esta matriz informa a probabilidade de ocorrência de um evento, perda estimada por ocorrência e o nível de violação de um dado princípio ético relacionado ao evento.

[6] https://ml-ethics-tool.web.app

Uma das principais maneiras de promover a implementação de requisitos éticos é a integração com práticas padrões de engenharia de software. Seguindo essa premissa Agbese et al. [69] propuseram um *framework* dividido em quatro camadas adaptados às diferentes funções nas organizações de software, sendo elas:

- Temas: gerência de maior nível, elabora requisitos éticos estratégicos para a empresa utilizando *frameworks* ou ferramentas apropriadas.
- Épicos: gerência de nível intermediário, interpreta os requisitos éticos estratégicos cascateando para o restante da equipe e desenvolve um plano de gerência.
- *Features*: gerência a nível operacional, alinha estratégias de negócio com requisitos mandatórios, decompondo requisitos éticos em requisitos éticos operacionais, facilitando a abordagem pelas unidades operacionais específicas da empresa.
- Histórias: individual ou time, interpretam os requisitos éticos operacionais em termos de atividades individuais ou de time e requisitos éticos individuais ou de time.

Agbese et al. [70] propuseram um *framework* para implementação de ética em IA dentro de um ambiente *project-based learning* (PBL). O *framework* consiste no uso de ferramentas definidas pela equipe durante o ciclo completo de um projeto de maneira não-linear, utilizando documentação feita por alunos para ajudar futuros participantes do projeto.

Sanderson et al. [71] realizaram entrevistas com pesquisadores e desenvolvedores da agência nacional de pesquisa da Austrália (CSIRO) que trabalham diretamente com desenvolvimento de IA. As entrevistas tem como objetivo entender aspectos éticos e seus princípios, abordados no dia a dia. Os temas relacionados a ética em IA citados nas entrevistas podem ser divididos em oito áreas: privacidade e segurança (81%), confiabilidade e segurança (90%), transparência e explicabilidade (86%), equidade (48%), contestabilidade (38%), responsabilidade (62%), valores humano-centrados (14%) bem estar humano, social e ambiental (52%).

Foram propostos dois conjuntos de suporte para operacionalização de princípios éticos dentro da organização. O primeiro conjunto trata do suporte organizacional e engloba práticas de definição de processos para implementação, treinamento, criação de políticas internas, utilização de *data planners* para identificação da necessidade do uso de IA e construção de tabelas contendo metadados dos *datasets*. O segundo conjunto aborda a perspectiva dos desenvolvedores, apresentando notas de design específicas para cada princípio [71].

O estudo de Kemell & Vakkuri [72] se baseia em conceitos comuns entre ética em IA e qualidade de software para propor um *framework* conceitual de custo ético de IA. O

framework é composto de três fases atribuídas aos princípios de qualidade. A primeira fase, investimento ético, trata do investimento inicial em incorporar ética na engenharia de software. A segunda fase, manutenção ética, trata do custo de manutenção dos processos definidos durante a primeira fase. Por fim, a terceira fase, receita ética, é relacionada a receita derivada das práticas definidas nas fases anteriores.

O termo "Indústria 5.0" representa uma mudança de foco da Indústria 4.0, orientada a tecnologia, para uma abordagem mais centrada no ser humano, na qual organizações pretendem alcançar mais do que a mera produtividade. Para isso, Ciobanu & Meșniță [73] desenvolveram um *framework* de ética em IA baseado no modelo Variáveis, Critérios, Indicadores e Observáveis (VCIO), uma abordagem multidimensional que trata de princípios éticos de IA em diversos níveis de produtores e consumidores. O *framework* possui duas camadas principais: *AI embedded ethics by design* (AI EED) e *AI desired state configuration* (AI DSC).

A primeira camada é composta de um grupo de processos nos quais as organizações podem utilizar proativamente antes do lançamento do produto. Os processos são: criação de *dashboards* de treinamento, modelos treinados para endereçamento de princípios éticos, princípios definidos pelo modelo VCIO e testes de modelo para mitigação de riscos antes de irem para produção. A segunda camada tem como foco a etapa de pós-implementação do produto, propondo o gerenciamento humano por meio de monitoramento, medição de critérios específicos e atualização constante [73].

Gerdes [74] apresentou um processo de *Value-Sensitive Design* (VSD) aplicado a sistemas de IA trazendo atenção a valores específicos. Estes valores podem estar relacionados com a falta de privacidade que pode resultar da utilização de dados para fins de treinamento, bem como com situações que podem ocorrer durante a fase de desenvolvimento e o ciclo de vida do desenvolvimento de software.

O método é aplicado durante o estágio de implantação, deixando de lado problemas envolvendo treinamento e verificação de modelos. Essa escolha se dá por Kelleher e Tierney [75], que determinaram que, durante o desenvolvimento da IA, equipes gastam cerca de 80% do seu tempo total na fase de preparação dos dados.

A abordagem proposta se apoia também nas dificuldades que os praticantes de IA encontram por estarem em um campo que exige a compreensão de dados multidisciplinares. Portanto é proposta a criação de atividades participativas e *workshops* com especialistas nos devidos campos de aplicação do sistema [74].

Os autores também propuseram atividades de investigação de valores como parte da elicitação de requisitos que podem ser realizadas por meio de entrevistas semi-estruturadas e ferramentas propostas na literatura. Por fim, a autora conclui que deve haver um monitoramento cauteloso da performance de sistemas de IA por meio de ferramentas de

verificação de qualidade e considerações de alterações causadas por ferramentas de IA no fluxo de trabalho [74].

Mäntymäki et al. [76] desenvolveram um modelo ampulheta de governança organizacional composto de três camadas. A camada inicial, chamada de camada de ambiente, é constituída por entradas advindas do contexto do ambiente organizacional, forças e fatores externos a sua área de influência. Esta camada é subdividia em três categorias: regulações normativas, princípios e diretrizes, e pressão de *stakeholders*. A segunda camada, chamada de camada organizacional, engloba dois temas inter-relacionados: o alinhamento estratégico e o alinhamento de valores. O alinhamento estratégico define a direção geral e as expectativas relacionadas com o sistema de IA, enquanto o alinhamento de valores exige a definição dos valores e da ética da IA a que a organização irá aderir. A terceira e última camada é chamada de camada do sistema de IA. Esta camada refere ao nível operacional da governança, no qual requisitos externos e diretrizes internas são implementados.

Lachenmaier et al. [77] propuseram um *framework* de governança para identificar e mitigar problemas éticos que podem surgir pelo uso de aplicações de IA. O *framework* considera a situação específica e os casos de utilização da entidade jurídica individual em questão. Ele é composta de doze áreas de governança e seis fatores de design, sendo eles:

- Áreas: privacidade de dados, monitoramento, gerenciamento de riscos, criação e funcionamento de soluções de IA, potencial e gerenciamento de inovações, segurança de TI, gerenciamento de conhecimento empresarial, padrões externos e fornecedores, perspectiva do usuário no uso de IA, *stakeholders*, estratégia e responsabilidade. Cada área deve ser trabalhada para o funcionamento correto do *framework*.

- Fatores de design: indústria, dados pessoais, objetos focais, criticalidade, impacto e abastecimento. O *framework* pode ser adaptado para organizações com base nos fatores de design propostos.

Com o objetivo de desenvolver um processo para verificar a confiabilidade em diferentes áreas de IA, Zicari et al. [78] propuseram o processo *Z-inspection* para auditoria. O processo é composto de três fases: *setup*, *assess* e *resolve*. A fase de *setup* é dividida em quatro subetapas, sendo elas:

- Catálogo de questões: definição de questões para serem respondidas antes processo para identificar as expectativas em relação ao sistema.

- Limites e contexto: identificar o contexto e os limites em que o sistema está inserido, tal como seu ecossistema (conjunto de setores e partes da sociedade, níveis sociais de organização e *stakeholders* em um contexto político e econômico).

- Como lidar com propriedade intelectual: como lidar com propriedade intelectual dentro do contexto do sistema de IA.
- Definição do intervalo temporal para a verificação: definir riscos presentes, futuros e a longo prazo.

A fase de *assess* é composta por [78]:

- Cenários socio-técnicos: cenários úteis para descrição de atores, expectativas, tecnologias, contextos e objetivos do sistema e atores.
- Descrição de problemas e tensões éticas: descrever e classificar se problemas éticos representam tensões éticas e descrevê-los. As tensões éticas podem ser classificadas em dilemas reais (conflito entre duas ou mais tarefas, obrigações ou valores), dilemas em prática (tensão existe por capacidade tecnológica ou restrições) e dilemas falsos (situações nas quais existem outras opções além do conflito entre valores).
- Áreas de investigação confiáveis de IA: utilizar a lista de problemas éticos para identificar quais áreas requerem inspeção.
- Mapeamento de problemas ético para áreas confiáveis de investigação: podem ser utilizados vários métodos para mapeamento e priorização entre membros do time, sempre procurando evitar a introdução de vieses.
- Execução: se o time de inspeção não assinou um acordo de não divulgação, a análise do código não é feita e a IA é verificada como caixa-preta. Se o acordo for assinado, uma investigação *top-down* é realizada, verificando dados de treino, teste, modelo e saídas.
- Pré-verificação: opcional, sendo possível a realização de uma pré-verificação não aprofundada em casos específicos.
- Criação de caminhos: descreve a dinâmica da inspeção e se difere em cada caso. Diferentes caminhos de investigação podem ser criados para verificar problemas éticos, técnicos e legais.
- Procura de caminhos: para a execução de caminhos, os especialistas podem definir um conjunto com o qual começar ou começar aleatoriamente, descobrindo as partes faltantes.
- Problemas éticos são reavaliados: ao final de cada caminho é feito um *feedback* contendo conjuntos de valores para de indicadores mensuráveis e textos para indicadores não mensuráveis. Após o *feedback* do último caminho, o time pode reanalisar os cenários socio-técnicos até chegar a um consenso.

Por fim, a fase de *resolve* é composta por [78]:

- Definição de pontuação: etapa opcional na qual pode ser utilizado um esquema de marcação para quantificar o nível de confiança e risco.
- Abordar e resolver tensões: após a lista de problemas éticos ser definida é feita uma priorização, identificando quais os desafios mais importantes neste caso de uso. Os autores propõem abordar estas questões por meio de um *framework* constituído por quatro elementos: aprofundar o entendimento das capacidades tecnológicas e limitações relevantes a problemas éticos, aplicar evidências para resolver tensões, construir trabalhos de engajamento público para entender as perspectivas de diferentes membros da sociedade, e identificar qual o tipo de dilema.
- Recomendações: dependendo da lista de problemas éticos classificados, o time de inspetores pode propor um conjunto de recomendações.
- *Trade-offs*: resultados da investigação podem ser relevantes aos *stakeholders* responsáveis pela tomada de decisões finais para o uso apropriado ou não do sistema.
- Manutenção ética: monitoramento do sistema de IA para confirmar que cumpre os requisitos de confiabilidade durante seu período ativo.

Para mensurar equidade em um sistema de IA, Agarwal et al. [79] propuseram um *framework* para cálculo de equidade e um índice de enviesamento para cada atributo protegido do *dataset*. As representações matemáticas e notações são:

$$\begin{aligned} a &= Y\text{: valor de treino} \\ b &= Y\text{': valor predito} \\ c &= S \in 0, 1\text{: indicador binário de classe protegida} \end{aligned} \tag{3.1}$$

Foram propostas cinco equações para cálculo de métricas, sendo as duas principais:

$$P(Y = 1|S = 0) - P(Y = 1|S = 1) \tag{3.2}$$

$$\frac{P(Y = 1|S = 0)}{P(Y = 1|S = 1)} \tag{3.3}$$

A aplicação do *framework* se dá por um conjunto de passos sequenciais, sendo eles [79]:

1. Identificar os atributos protegidos do *dataset*.
2. Identificar as classes privilegiadas e não privilegiadas do *dataset*.
3. Definir a banda de tolerância para cada atributo.

4. Computar os valores das métricas definidas para cada atributo protegido de treino e verificar se estão na banda de tolerância.
5. Treinar o modelo utilizando o *dataset* de treino.
6. Rodar o modelo no *dataset* de teste e registrar o resultado.
7. Computar os valores das métricas definidas para cada atributo protegido e verificar se estão na banda de tolerância.
8. Criar um gráfico com os valores de cada métrica para ajudar na visualização dos resultados.
9. Calcular o índice de enviesamento para cada atributo protegido e a pontuação de equidade total do sistema.
10. Baseado nos resultados anteriores, o certificado de equidade pode ser publicado.

Zhang et al. [80] desenvolveram um *framework* metodológico chamado *Fairness in Design* (FID) que tem como objetivo facilitar a identificação e mitigação de riscos relacionados a equidade em sistemas de IA. A metodologia é composta por um conjunto de dez cartas consolidadas por dez princípios, divididos em duas categorias: equidade grupal e individual. O fluxo simplificado se dá pela aplicação das cartas relacionadas ao domínio do sistema, seguido pela escolha das métricas de equidade, simulando problemas e soluções de cada *stakeholder* do projeto. Para finalizar, é realizada uma avaliação de importância dos princípios de equidade para o sistema por meio de uma escala *likert*.

Ayling & Chapman [21] realizaram um estudo com o objetivo de utilizar a avaliação de impacto e auditoria para desenvolver uma tipologia de análise documental comparativa das ferramentas propostas na literatura. Os autores usaram análise quantitativa de conteúdo para elicitar terminologias e abordagens frequentemente aplicadas. Por fim, o *dataset* com 169 documentos foi filtrado para conter apenas documentos que possuíam ferramentas de apoio. Os resultados chave encontrados pelas autoras são:

- O foco foi alterado de dados para modelos entre 2017 e 2020, saindo de tópicos como *big data* para modelos e algoritmos nesse intervalo temporal.
- Os tipos de *stakeholders* primários são agrupados durante a fase de desenvolvimento de software, enquanto as saídas das ferramentas são usados pela gerência.
- Existe pouca participação em processos de verificação e auditoria por certos grupos de *stakeholders*, como usuários e clientes.

- Das ferramentas analisadas, apenas uma não foi desenvolvida para auto-verificação interna.
- Questionários e checklists são muito utilizadas em ferramentas de verificação de impacto, mas menos encontradas em ferramentas de auditoria.
- A saída das ferramentas analisadas podem providenciar documentação para supervisão de atores externos, mas não existem processos ou requisitos para publicação dos seus resultados.
- Um terço das ferramentas de verificação de impacto focam em processos de aquisição de sistemas de IA por fornecedores terceirizados.

Squadrone et al. [81] propuseram um *framework* de ética por design que utiliza redes neurais para evitar problemas técnicos e discriminatórios em sistemas existentes. O *framework* propõe a injeção de regras éticas durante o treinamento de sistemas de IA, permitindo que a IA selecione o melhor resultado baseando-se no conjunto de dados e regras éticas. Princípios éticos abstratos podem ser aplicados por meio de regras éticas que restringem características individuais, como raça e gênero, determinando assim o grau ético da decisão.

De acordo com Rees & Muller [57], a falta de conhecimento sobre requisitos éticos faz com que a aplicação deles seja nula. Para contornar esse problema, os autores fizeram uma proposta de um conjunto mínimo de requisitos éticos para sistemas de IA:

- Foco em diversidade: étnica, cultural, socio-econômica, religiosa, de gênero, idade, metodológica e de habilidade.
- Desenvolver sistemas considerando possíveis legislações futuras sobre responsabilidade de IA.
- Verificação regular da conformidade ética e legal de ferramentas de IA.
- Sumarizar as decisões éticas do sistema com termos para leigos pois isso demonstra transparência, equidade e rastreabilidade.

Schmid & Wiesche [82] analisaram um *framework* de impacto ético de uma montadora de carros para calibrar a confiança em sistemas de IA. O *framework* é composto de cinco elementos: *failure models*, ocorrências, detectabilidade, severidade e impacto ético. Para análise, foram estudadas a calibração de confiança e o envolvimento de confiabilidade de IA durante as fases de engenharia de requisitos, design, implementação, verificação e operação. Observou-se que, antes da fase de verificação e validação, os funcionários da empresa demonstraram falta de confiança no sistema de IA, dada a possibilidade de falha,

mesmo que mínima. No entanto, após esta fase, os níveis de confiança aumentaram, o que indica que a calibração e a avaliação dos riscos efetuadas pelo *framework* tiveram um impacto positivo.

Lu et al. [83] conduziram uma revisão multivocal de literatura para encontrar padrões de governança, padrões de processo e padrões de produto relacionados a IA responsável (RAI). Foram identificados 24 padrões de governança, divididos em nível de indústria, organização e time: marca de confiança, modelo de maturidade RAI, supervisão independente, certificação RAI, *sandbox* regulatório, regulação RAI, construção de código, padrões RAI, comitê de ética, código de ética, verificação de riscos éticos, compromisso da liderança com RAI, relatórios padronizados, contrato de responsabilização por função, treinamento em ética e *bill of materials* de RAI.

Os padrões de processo foram divididos em cinco grupos [83]: engenharia de requisitos, design, implementação, testes e operação. Estes processos são: verificação de aptidão de IA, requisitos éticos verificáveis, requisitos de dados durante o ciclo de vida, histórias éticas de usuário, co-arquitetura multinível, carta de previsão, modelagem de design para ética, simulação ética a nível de sistema, design de interface humano-centrada para XAI, governança de RAI para APIs, governança de RAI por APIs, construção ética com reuso, testes de aceitação éticos, verificação ética de casos de teste, *deploy* contínuo para RAI, co-versionamento multinível e verificação de riscos extensível, adaptável e dinâmica [83].

Por fim, os padrões de produto são [83]: registro de *bill of materials*, credenciais éticas verificáveis, registro de co-versionamento, aprendizagem federada, alternador de modo IA, decisor multi-modelo, redundância homogênea, validador ético contínuo, *sandbox* ético, base de conhecimento ética, registro de incentivos, caixa preta ética, auditor de visão global e gêmeo digital ético. A Tabela 3.7 sumariza as técnicas, metodologias, métodos, *frameworks*, ferramentas, processos e ou ferramentas identificados na RSL.

O conjunto final é constituído por 24 propostas de soluções práticas identificadas na literatura. Estas 24 propostas estão divididas em cinco categorias, sendo elas: processos (n=3), métodos (n=3), metodologias (n=3), *frameworks* (n=14) e ferramentas (n=1). A frequência de cada categoria é ilustrada de forma gráfica na Figura 3.6. A Tabela 3.7 apresenta as 24 propostas identificadas na literatura, acompanhadas de uma breve descrição de cada uma e das respectivas citações.

Tabela 3.7: Técnicas, metodologias, métodos, *frameworks*, ferramentas, processos e ou ferramentas identificados na RSL

Tipo	**Proposta**	**Referência**
Ferramentas	Tipologia de análise documental comparativa das ferramentas propostas na literatura	[21]
Frameworks	Conjunto mínimo de requisitos éticos para sistemas em IA	[57]
	Aplicação do método ECCOLA no contexto de cooperação em sistemas multi-robôs	[62]
	Guia de campo para o desenvolvimento de *Machine Learning*	[67]
	Framework de avaliação ética por meio de matrizes	[68]
	Framework de camadas para diferentes cargos em organizações de software	[69]
	Framework para implementação de ética em IA dentro de um ambiente *project-based learning*	[70]
	Conjuntos de suporte para operacionalização de princípios éticos de IA	[71]
	Framework conceitual de custo ético de IA	[72]
	Framework de ética em IA baseada no modelo variáveis, critérios, indicadores e observáveis (VCIO)	[73]
	Framework de governança para identificar e mitigar problemas éticos	[77]
	Framework para cálculo de equidade de um sistema	[79]
	Fairness in Design	[80]
	Framework de ética por design que utiliza redes neurais	[81]
	Framework de impacto ético	[82]
Metodologias	Modelo de implantação para o método Eccola	[63]
	Avaliação de impactos éticos de IA baseado em ER	[66]
	Modelo ampulheta de governança organizacional	[76]
Métodos	Eccola	[45]
	Lacunas no método Eccola	[64]
	Análise do método Eccola com base na *framework* GARP	[65]
Processos	*Value-Sensitive Design* aplicado a IA	[74]
	Z-inspection	[78]
	Catálogo de padrões de processo	[83]

Fonte: o Autor

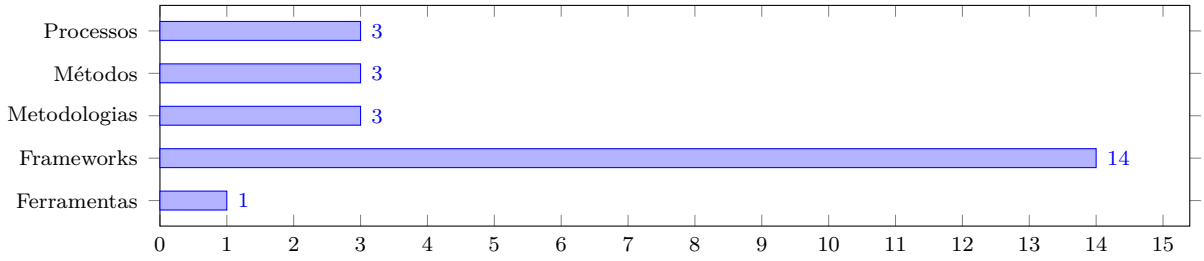


Figura 3.6: Propostas de soluções práticas para a aplicação de ética em IA identificadas na literatura. Fonte: o Autor.

Achados

Os resultados deste estudo indicam que a maioria das propostas assumem a forma de *frameworks*, com mais de metade dentro desta categoria. Dentro desta, as propostas abrangem um conjunto diversificado de tópicos, incluindo governança, *ethics-by-design* e guias de campo [67, 77, 81]. Além disso, a investigação revelou uma tendência para os métodos e metodologias gravitarem em torno do método ECCOLA [45]. Isto levou à proposta de novas formas de o implementar, à identificação de potenciais lacunas e à análise da sua eficácia [63, 64, 65]. A investigação também revelou que os processos tendem a identificar aplicações inovadoras de abordagens estabelecidas, como a auditoria e a monitorização, a fim de garantir que sistemas de IA sejam desenvolvidos em conformidade com normas éticas [74, 78].

As propostas descritas acima abrangem uma série de tópicos, categorias e abordagens, criando um ecossistema diversificado no âmbito da ética da IA. Isto demonstra a necessidade de uma variedade de soluções práticas para abordar eficazmente as diferentes questões, uma vez que nenhuma abordagem atual única pode abranger todos os aspectos do assunto. Ao adotar esta variedade, os *stakeholders* podem navegar mais efetivamente nas complexidades das ferramentas, *frameworks*, métodos, metodologias e processos modernos e contribuir para soluções mais abrangentes e adaptáveis, bem como integrar uma série de soluções práticas de forma eficaz.

Análise

É notável o aumento do número de estudos que propõem maneiras para operacionalizar os requisitos éticos de IA se comparado com os encontrados por Cerqueira [1]. No entanto, como já foi destacado por Cerqueira [1], ainda não é possível confirmar se todas as lacunas existentes foram resolvidas. Propostas como a *Z-inspection* [78] são aplicadas ao longo do ciclo de vida do desenvolvimento, mas apenas abordam um tópico do conjunto total de princípios. Este facto realça a necessidade de utilizar um conjunto de soluções para

desenvolver um sistema de IA que seja considerado minimamente ético. Os tópicos de design, governança e avaliação estão entre os mais discutidos na literatura, indicando que as fases inicial e final do ciclo de vida do desenvolvimento são de grande importância.

3.9.3 RQ3 - Como possibilitar a implementação de princípios éticos de IA durante o ciclo de desenvolvimento de software?

Tidjon & Khomh [24] fizeram a análise de 100 princípios e suas implementações para entender quais são as lacunas existentes entre princípios e implementações e como solucioná-las. Foi identificado um grupo de lacunas principais e suas causas, sendo elas:

- Falta de ferramentas de implementação para princípios de ética em IA. As ferramentas desenvolvidas focam em princípios como equidade, não-discriminação e enviesamento e muitas vezes não abordam todos os aspectos relacionados aos princípios em questão.
- Falta de padrões efetivos.
- Falta de cursos práticos de ética em IA, sejam em universidades e companhias de treinamento.
- Implementação fraca de princípios de ética em IA dentro da governança de organizações, como supervisão imprópria, problemas durante decisões processuais, entre outros.
- Enviesamento humano, visto que as implementações éticas não possuem uma transição desejável entre os princípios abstratos éticos e algoritmos e ferramentas de IA. Isso se dá principalmente por percepções individuais na implementação e dificuldade de tradução de valores complexos em ferramentas.
- Pouca diversidade na comunidade de IA.
- Complexidade e interconexão de decisões e processos aprendidos por dados.
- Problemas em parcerias público-privadas e pesquisas financiadas pela indústria, pois limitam a liberdade dos pesquisadores exigindo que o foco das pesquisas seja nas necessidades privadas e não na solução de problemas do mundo real.
- Falta de métricas para avaliar/implementar princípios éticos de IA.

Para solucionar os problemas encontrados, os autores propuseram cinco soluções [24]: 1) Deve haver diversidade e inclusividade nos membros envolvidos no design e implementação de ferramentas éticas, evitando apenas pessoas técnicas; 2) Educação e conhecimento em valores éticos, culturais, métodos e práticas devem ser incentivados para garantir que times desenvolvam uma perspectiva ética; 3) Times de desenvolvimento devem alinhar projetos com *EthicsOps* para garantir que soluções são construídas para seguir princípios éticos durante design, desenvolvimento e implantação; 4) Aplicar leis e padrões éticos durante a AI governança e processos de engenharia; e 5) Parcerias entre setores público/-privado e universidades devem ser incentivadas para acelerar a pesquisa, desenvolvimento e transferência de conhecimento na área de ética em IA.

Georgieva et al. [84] realizaram um mapeamento dos requisitos de confiabilidade HLEG no ciclo de vida de um serviço digital ou produto baseado em IA para identificar a sua aplicabilidade prática. Os requisitos utilizados foram: supervisão humana, robustez técnica e segurança, privacidade e governança de dados, transparência, diversidade e não discriminação (equidade), responsabilidade e bem estar ambiental e social.

Com esse mapeamento, foi encontrado um conjunto de resultados práticos e implicações [84]. Para desenvolvedores é observada a importância da tradução de requisitos éticos em nível operacional, seja por procedimentos e restrições ou por *ethics-by-design* posteriormente testada e validada. Em relação a gerência, mesmo que exista a promoção de princípios relevantes ainda não é claro como traduzi-los para prática, dificultando sua implementação. Uma potencial solução é o uso de *Design for Values* (DfV), uma abordagem metodológica de design que tem como foco fazer com que valores façam parte do design, pesquisa e desenvolvimento tecnológico.

Kemell et al. [38] tiveram como objetivo entender o ponto de vista do produto em relação aos princípios de ética em IA. A metodologia do estudo consiste na utilização de diretrizes e aplica-las a um projeto ficcional. Os autores identificaram quatro barreiras que devem ser ultrapassadas para permitir a implementação de requisitos éticos, sendo elas: 1) Falta de obrigação de empresas de software em fazer o bem, sendo mais responsáveis pelos acionistas do que para com a sociedade; 2) Pouca tradição no que se diz em ser um bom desenvolvedor se comparado, por exemplo, com um bom médico; 3) Falta de métodos comprovados para traduzir princípios a prática; e 4) Falta de mecanismos robustos de responsabilidade legal e profissional.

Barletta et al. [23] fizeram uma revisão em um conjunto de *frameworks* de ética em IA. Como resultado, foi encontrado que em sua maioria, *frameworks* são desenvolvidas para a fase de elicitação de requisitos. Sobre a origem, foi encontrado que *frameworks* são propostas por instâncias heterogêneas, públicas e privadas, acrescentando na democratização da IA (diversidade de perspectiva), mas apresenta a falta de consenso e uniformização.

Poucos *frameworks* abordam todas as fases do ciclo de vida e provêm apoio prático a quem deseja desenvolver, testar e realizar a implantação de aplicações. Não foi encontrada pelo estudo um *framework* completo, de uso simples, organizado e uniforme para apoiar os *stakeholders* durante todo o ciclo de vida. Portanto se torna necessária a utilização de um conjunto de *frameworks*, processos e ferramentas para o desenvolvimento ético de IA.

O *framework* proposto por Agbese et al. [69] aborda os três principais níveis de gerência e o nível de operacionalização, seja de maneira individual ou em times de desenvolvimento. Para facilitar a implementação de requisitos éticos, o *framework* sugere a criação dos mesmos a partir da gerência de alto nível, tal qual Temas, cascateando para a gerência intermediária, fragmentando os Temas em Épicos. Seguindo a cascata, os Épicos são posteriormente divididos em *Features* pela gerência operacional e por fim são transformados em histórias éticas de usuário a nível de times.

Para preencher lacunas existentes na operacionalização de RAI, Lu et al. [85] desenvolveram um *roadmap* de Engenharia de Software para RAI. O *roadmap* é dividido em três perspectivas diferentes, sendo elas:

- Perspectiva de governança: estruturas e processos desenvolvidos para garantir que sistemas de IA são compatíveis com regulações e responsabilidades éticas. É dividido em níveis industriais, organizacionais e de time, com cada nível contendo uma série de práticas.

- Perspectiva de processo: conjunto de melhores práticas incorporadas a processos de desenvolvimento. Os processos incluem engenharia de requisitos, design, implementação, verificação e operação. As práticas são específicas por fases do processo, como histórias éticas de usuário compondo a fase elicitação que se encontra no processo de engenharia de requisitos.

- Perspectiva de sistema: práticas para possibilitar a implementação de RAI a nível de arquitetura e sistema, como estilo e padrões de arquitetura e práticas mais pontuais, como aprendizado federado, computação aproximada, remoção de dependências, entre outras.

O guia de campo[7] desenvolvido por Boyd [67] para o desenvolvimento ético de *Machine Learning* (ML) é uma ferramenta que apresenta um conjunto de estratégias para mitigação de problemas éticos. O guia define as situações para o uso de cada estratégia, os requisitos necessários e um resumo do procedimento, permitindo a filtragem por objetivos, tipo de dados, problema ético, campo de pesquisa e estágio de desenvolvimento.

O método ECCOLA [45], apresentado na Subseção 3.9.2, é composto de 21 cartas divididas em oito temas. As cartas são: análise de *stakeholder*, tipos de transparência,

[7] https://ml-ethics-tool.web.app

explicabilidade, comunicação, documentação de *trade-offs*, rastreabilidade, confiabilidade do sistema, privacidade e dados, qualidade de dados, acesso a dados, atividade humana, supervisão humana, segurança de sistema, proteção de sistema, acessibilidade, participação de *stakeholder*, impacto ambiental, efeitos sociais, auditoria, capacidade de reparação e minimização de impactos negativos. Na sua utilização, um conjunto relevante de cartas é aplicado cada *sprint*, sendo documentado, revisado e avaliado para manutenção de rastreabilidade e verificação da execução correta das ações planejadas.

O *framework* proposto por Zhang et al. [80] foi apresentada na Subseção 3.9.2, mas o fluxo de trabalho foi simplificado. O fluxo de trabalho completo se dá por:

- Escolha do domínio da aplicação que definirá o ambiente para a sessão.
- Cada membro da equipe seleciona um tipo de carta que melhor descreve o domínio da sua aplicação, sendo eles: sistemas críticos, uso industrial e comercial, escritório e casa, exploratório e criativo, aplicações colaborativas e aplicações sócio-técnicas.
- A equipe define o conjunto de *stakeholders* do domínio da aplicação.
- Cada membro da equipe retira uma carta de princípios de equidade e a aplica no domínio.
- Um por um, os membros do time aplicam a métrica de sua carta ao domínio e estimulam os problemas e soluções potenciais do conjunto de *stakeholders*.
- O time compila o conjunto de respostas dos membros e as randomiza para leitura anônima.
- A equipe realiza a leitura, discussão e avaliação das respostas, definindo o grau de importância por meio de uma escala *likert*.
- Após a finalização do processo, o time pode conduzir uma nova análise de *stakeholders* ou apenas finalizar a sessão.

O questionário aplicado por Balasubramaniam et al. [60] identificou os princípios de transparência e explicabilidade como os mais utilizados nas organizações estudadas, como visto na Subseção 3.9.2. Com essa informação, os autores realizaram um estudo empírico para avaliar os resultados com profissionais da área. Foi identificado pelos autores um conjunto de boas práticas, sendo duas diretamente relacionadas a implementação do princípio de explicabilidade. A primeira define a utilização de um modelo de componentes de explicabilidade para identificar e analisar as necessidades do conjunto de *stakeholders* relacionadas ao requisito. Por fim, a segunda aborda a utilização de um *template* para

representar requisitos específicos de explicabilidade de maneira estruturada. As três práticas finais são: realização de *workshops* multidisciplinares, definição clara dos objetivos do sistema de IA na perspectiva de usuários e outros *stakeholders* e consideração de riscos potenciais e consequências negativas do sistema.

O *framework* proposto por Ciobanu & Meșniță [73] é baseado no modelo variáveis, critérios, indicadores e observáveis (VCIO), consiste de duas camadas, a primeira relacionada a implementação e a segunda a manutenção. Os processos de interesse para a implementação de requisitos éticos são:

- Desenvolvedores devem ter sempre uma interface de IA em formato *dashboard* para o treinamento do modelo.
- Cada desenvolvedor das partes produtoras e consumidoras do modelo podem definir os princípios pelo modelo VCIO.
- Dentro da interface de *dashboard*, os desenvolvedores podem criar modelos de treinamento para cada princípio definido.
- Cada modelo de treinamento pode ser testado antes de ir para produção. Durante a execução dos testes devem ser realizadas a análise de riscos e ações para mitigar os riscos encontrados.

Para entender quais são as principais abordagens de ética em IA, Prem [86] fez uma análise sistemática de 100 *frameworks*, modelos de processo e ferramentas. As principais abordagens aplicadas, como algoritmo e código, tendem a focar em áreas mais específicas, como explicabilidade e equidade. Em relação a construção de uma infraestrutura de ética em IA, o autor propõe uma abordagem de conjunto, contendo algoritmos, para pontos específicos, ferramentas e *frameworks*. A Tabela 3.8 apresenta as 12 principais sugestões identificadas na literatura, acompanhadas das suas respectivas referencias.

Tabela 3.8: Sugestões para implementação de requisitos e princípios éticos

Sugestões para implementar requisitos e princípios éticos	**Referência**
Identificação e soluções para as lacunas: falta de ferramentas, padrões e cursos práticos, implementação imprópria, enviesamento humano, falta de diversidade na comunidade de IA, complexidade de decisões, problemas em financiamentos feitos pela indústria e falta de métricas de avaliação	[24]
Conjunto de resultados práticos: utilização de procedimentos, restrições ou modelos de *ethics-by-design* para tradução de requisitos éticos a nível operacional e utilização de DfV para ajudar com essas tarefas	[84]
Infraestrutura composta por algoritmos, ferramentas e *frameworks* para abordar o conjunto completo de requisitos éticos	[86]
Barreiras a serem superadas na implementação: falta de comprometimento em fazer o bem por parte das empresas, pouca tradição no que se diz em ser um bom desenvolvedor, falta de métodos para traduzir princípios na prática e falta de mecanismos robustos de responsabilidade	[38]
Revisão de *frameworks*	[23]
Método em cascata para a implementação de requisitos éticos nas organizações	[69]
Roadmap de Engenharia de Software para RAI	[85]
Guia de campo para desenvolvimento ético de Machine Learning	[67]
Método iterativo para aplicação de ética em IA	[45]
Framework de ética para o design de sistemas de IA	[80]
Conjunto de boas práticas: modelo de componentes de explicabilidade para *stakeholders*, utilização de um *template* para requisitos de explicabilidade, organização de *workshops*, entendimento do processo apoiado pelo sistema, definir claramente o sistema para *stakeholders* e considerar possíveis riscos	[60]
Processos para a implementação de princípios e requisitos éticos	[73]

Fonte: o Autor

Análise

A aplicação dos princípios éticos de IA enfrenta uma série de desafios significativos. Um dos principais obstáculos é a ausência de uma obrigação formal de as empresas de software priorizarem as práticas éticas, aliada à falta de uma tradição ética no desenvolvimento de software em comparação com domínios estabelecidos como a medicina [38]. Além disso, existe uma lacuna entre os princípios éticos e a sua aplicação prática, incluindo ferramentas insuficientes para a implementação, falta de normas eficazes e recursos educativos limitados em matéria de ética da IA [24].

Para abordar estas lacunas, foram propostos diversos *frameworks* e metodologias. Por

exemplo, soluções como as desenvolvidas por Agbese et al. [69] e Lu et al. [85] procuram integrar os requisitos éticos em diferentes níveis de gestão e desenvolvimento. Diversas ferramentas práticas foram desenvolvidas para ajudar na incorporação da ética nos sistemas de IA. Estas incluem o método ECCOLA [45], o Guia de Campo de Boyd [67] e o modelo VCIO de Ciobanu & Meșniță [73], que fornecem abordagens estruturadas para incorporar a ética nos sistemas de IA, oferecendo estratégias práticas para abordar considerações éticas durante as fases de design e desenvolvimento.

No entanto, estes esforços ainda não conseguiram atingir o nível de padronização desejado. Tal como observado por Barletta et al. [23], um número considerável de *frameworks* tem como único objetivo a fase de elicitação de requisitos e não cobre a totalidade do ciclo de vida de desenvolvimento de IA. Prem [86] sugere uma abordagem de conjunto para enfrentar este desafio, que combina um grupo de soluções práticas compostas por metodologias, métodos, ferramentas e *frameworks* para construir uma infraestrutura de ética em IA.

Em conclusão, a implementação dos princípios éticos de IA é um processo complexo. Embora uma variedade de soluções práticas possa ser aplicada a aspectos específicos do desenvolvimento de sistemas baseados em IA, é necessário adotar uma abordagem mais unificada para tratar na sua totalidade. Por isso, para garantir o desenvolvimento ético dos sistemas baseados em IA, é essencial adotar uma abordagem que integre uma série de soluções para construir a chamada infraestrutura de ética em IA. Como não é viável, num futuro próximo, desenvolver uma solução completa que aborde todas as potenciais questões éticas de um sistema baseado na IA, esta abordagem integrada será fundamental para o avanço do desenvolvimento de IAs éticas.

Além disso, foi verificado que os métodos, processos e sugestões propostos na literatura são extremamente diversificados quanto ao tipo de abordagem, fase do ciclo de vida e objetivos. Comparando este resultado com o de Cerqueira [1], que encontrou uma tendência para a utilização e adaptação de processos de engenharia de requisitos, torna-se claro que o campo tem sofrido uma transformação, englobando um maior número de soluções para mitigar os riscos e problemas éticos presentes nos sistemas de IA.

3.10 Discussão

Nesta seção são discutidos os resultados obtidos por esta RSL, apresentando as conclusões e esclarecendo as informações consideradas mais importantes. Os resultados também serão comparados aos obtidos por Cerqueira [1] com o objetivo de identificar as principais diferenças.

Um grupo de princípios éticos foi identificado como frequente nos estudos revisados. Este grupo é constituído por 25 princípios e pode ser encontrado na Tabela 3.6. Comparando este grupo com estudos primários que analisaram diretrizes e princípios [57, 58], nota-se que os princípios prioritários tendem a permanecer os mesmos, passando por pequenas alterações. Esta afirmação é apoiada pelo estudo conduzido por Balasubramaniam et al. [60], que identificaram transparência e explicabilidade como os princípios mais comuns em organizações.

Entre os estudos primários, verificou-se que a análise e a validação de princípios e diretrizes foram feitas em detrimento de novas propostas [57, 58]. Isto é confirmado pelo fato de nenhum dos estudos primários ter proposto um novo conjunto de princípios ou diretrizes. Isto sugere que os principais princípios éticos foram abordados na literatura e que existe um certo grau de maturidade. Por outro lado, a implementação de requisitos éticos nos sistemas de IA tem sido lenta, quer devido à falta de normalização, a abordagens não específicas ou à falta de apoio interno [9, 38, 58, 44]. Portanto é necessário aproximar a prática da teoria com soluções que preencham as lacunas encontradas.

Dos 38 estudos analisados, 24 propõem algum tipo de solução para o desenvolvimento de sistemas éticos de IA. Este número de soluções apresenta um crescimento em comparação com a revisão de Cerqueira [1] e mostra que o campo está a amadurecer, embora lentamente, como mencionado no parágrafo anterior. No entanto, é importante reconhecer que as propostas de soluções práticas, por si só, não preencherão as lacunas.

Em termos das características das soluções propostas, Kelleher & Tierney [75] constataram que desenvolvedores de IA passam 80% do seu tempo total na fase de preparação dos dados. Em consonância com a afirmação anterior, esta análise identificou uma tendência de priorização das etapas iniciais e finais do ciclo de vida de desenvolvimento. O estudo publicado por Tidjon & Khomh [24] definem que, para possibilitar a implementação de sistemas éticos de IA, é necessário entender o conjunto de problemas que afetam o campo. Um dos problemas identificados é o desenvolvimento monofocal das ferramentas. Ferramentas essas que em certos casos não abordam nem o contexto completo do único princípio em questão. Estes constatações demostram que, atualmente, é necessário implementar uma infraestrutura constituída por um conjunto de soluções para o desenvolvimento ético de IA ao longo do ciclo de vida, introduzindo mais um empecilho para a operacionalização de requisitos éticos.

3.10.1 Comparação de resultados

Durante a revisão, Cerqueira [1] encontrou 1018 estudos em resultado da sua pesquisa automatizada e 33 ensaios primários após a filtragem final. Nesta atualização foi encontrado um aumento de 57,6% no número de estudos, sendo 1605 após a filtragem de

estudos duplicados e 38 estudos primários. Isto mostra que o tópico da ética da IA está a tornar-se mais amplamente pesquisado. A Figura 3.3 colabora com essa afirmação, visto que um padrão de crescimento é observado na mesma ao decorrer dos anos.

Na revisão publicada por Cerqueira [1], foi identificada uma inclinação para com o uso e adaptação de processos de Engenharia de Requisitos no contexto de ética de IA. Em contrapartida, neste estudo foi encontrado que o conjunto de métodos, processos e propostas encontrados na literatura é extremamente diversificado em termos de tipo de abordagem, fase do ciclo de vida e objetivos. Essa afirmação evidencia a transformação heterogênea que o campo está passando, englobando um maior número de soluções para mitigar os riscos e as questões éticas dos sistemas de IA. Espera-se que esta diversificação dentro do campo traga um impacto positivo no futuro da ética de IA.

A maturidade das soluções propostas é outro ponto que contradiz as conclusões de Cerqueira [1]. Soluções como a *Fairness in Design* e a *Z-inspection* [78, 80], apesar de não abordarem todo o conjunto de requisitos éticos, são consideradas maduras naquilo que propõem, apresentando formas claras de implementação e os seus objetivos finais. No entanto, em conformidade com as conclusões de Cerqueira [1], tende a haver uma falta de evidências substanciais e de testes suficientes para garantir a sua eficácia na operacionalização da ética de IA.

O conjunto de princípios encontrados nos estudos primários contém 25 princípios. Dentre estes 25, quatorze estão presentes nos principais princípios definidos no estudo de Cerqueira [1], sendo eles: transparência, equidade, não maleficência, responsabilidade, privacidade, beneficência, liberdade e autonomia, confiabilidade, sustentabilidade, dignidade, solidariedade, dados, supervisão e segurança.

3.11 Ameaças à Validade

Esta seção descreve as possíveis ameaças a validade desta atualização de revisão sistemática de literatura e as respectivas estratégias de mitigação utilizadas. Para analisar as ameaças foram utilizados os conceitos de validade de construto, interna e externa definidas por Wohlin [87].

Validade interna tem como foco relações causais, em outras palavras, identificar se o tratamento causou o resultado [87]. Em atualizações sistemáticas de literatura pode ser referida como o grau de rigor do processo de revisão. Para mitigação, esta RSL foi feita seguindo o processo definido por Kitchenham et al. [47, 48].

Tal como Cerqueira [1], uma das dificuldades encontradas durante a execução deste estudo foi a relação encontrada entre os estudos revisados e as questões de pesquisa. Vários estudos definem um conjunto de princípios éticos, respondendo a RQ1. Em adição,

propõe uma solução prática, respondendo a RQ2. Por fim, abordam maneiras para a implementação de requisitos éticos, respondendo a RQ3. A mitigação desta ameaça foi feita por meio de definição de preferência em cada questão de pesquisa. Para responder a RQ2 foram priorizados os estudos que analisam ou propuseram alguma solução prática para a implementação de requisitos éticos. Para a RQ1 foi dada a preferência para estudos que propõem ou analisam um conjunto de princípios éticos em IA. Em relação a RQ3, foram priorizados estudos que analisam ou propõem alguma solução para a implementação de requisitos éticos, seja por meio da mitigação de problemas encontrados ou por estratégias. Estudos que propõem soluções práticas juntamente com soluções para implementação foram divididos de maneira diferente, tendo cada um de seus resultados abordados por diferentes perguntas.

Validade externa tem como foco a generalização. Se existe uma relação causal entre o construto e o efeito, o resultado do estudo pode ser generalizado fora do seu escopo [87]. As questões de pesquisa definidas podem não cobrir a área total de ética em IA. Portanto, não é possível que sejam encontradas respostas para questões não especificadas neste estudo. Como este estudo fez uma atualização da RSL de Cerqueira [1], para mitigação do risco às questões de pesquisa se mantiveram as mesmas.

Validade de construto foca na relação entre teoria e observação, garantindo que o tratamento e resultado refletem bem o construto do efeito [87]. O processo de seleção dos estudos primários pode afetar a qualidade dos dados coletados para a síntese. Além disso, é impossível garantir que todos os estudos primários relevantes tenham sido selecionados para esta revisão. Para mitigação de ambos os riscos, a *string* de busca da pesquisa automatizada passou por um processo de atualização até cumprir os critérios definidos. Por fim, o resultado foi analisado por três pesquisadores.

Durante o processo de extração de dados, os estudos primários foram classificados e resumidos pelo autor, introduzindo a variável de subjetividade. Para contornar este problema, o resultado foi revisado por outros dois pesquisadores.

Capítulo 4

Proposta da Ferramenta

Este Capítulo descreve a metodologia utilizada no desenvolvimento de uma ferramenta para automatizar a geração de histórias de usuário a partir de um conjunto de requisitos éticos, desde as fases iniciais, como a obtenção do conjunto de dados, passando pela preparação dos dados, até à fase final de modelagem e avaliação da qualidade do modelo. Para identificar as características da ferramenta, foram utilizados os resultados da revisão sistemática de literatura do Capítulo 3. O código da ferramenta pode ser encontrado no repositório do Github[1].

4.1 Motivação da Ferramenta

A revisão sistemática da literatura, detalhada no Capítulo 3, apresentou um conjunto de características que embasaram a proposta da ferramenta desenvolvida neste trabalho. Em primeiro lugar, é importante reconhecer que os princípios éticos tendem a ser constantes [57, 58, 60], e ainda assim as implementações existentes são lentas devido à falta de padronização e às abordagens não específicas, o que torna difícil incorporar os princípios éticos ao processo de desenvolvimento de software [9, 38, 58, 44]. Outro problema identificado é o desenvolvimento monofocal de ferramentas para apoiar a implementação dos princípios éticos na prática. Em alguns casos, as ferramentas disponíveis nem sequer abordam o contexto completo do princípio em questão. Essas descobertas demonstram que, é necessário implementar uma infraestrutura que consiste em um conjunto de soluções para o desenvolvimento ético da IA durante todo o ciclo de vida do software. A falta de ferramentas práticas introduz mais um obstáculo à operacionalização dos requisitos éticos.

Portanto, com o intuito de mitigar estes problemas, a ferramenta deve atender a alguns requisitos essenciais. Primeiramente, é fundamental que a ferramenta seja projetada

[1] https://github.com/joaorossi15/requirements-to-us

para se integrar intuitivamente aos fluxos de trabalho dos times de desenvolvimento de software já existentes, permitindo uma adoção sem dificuldades significativas. Além disso, é fundamental que ela seja baseada em padrões claros e consistentes, facilitando a incorporação dos princípios éticos de maneira uniforme e menos suscetível a variações, superando a falta de padronização que torna as implementações lentas e complexas. Além disso, a ferramenta deve tentar integrar o conjunto de princípios éticos de maneira abrangente e detalhada, evitando uma abordagem monofocal e rasa. Por fim, é importante que, no caso da criação de uma infraestrutura de desenvolvimento ético de IA, a ferramenta seja facilmente acoplada.

Kelleher & Tierney [75] constataram que desenvolvedores de IA passam 80% do seu tempo total nas fases iniciais do projeto. Portanto, é possível deduzir que uma ferramenta que atenda aos requisitos mencionados acima e seja aplicada nos estágios iniciais do projeto pode representar uma solução que tende a apoiar as equipes de desenvolvimento nessa fase. Dessa forma, a ferramenta opera traduzindo requisitos éticos de IA em histórias éticas de usuário, facilitando assim a implementação prática de princípios éticos em sistemas de IA e preenchendo a lacuna entre a teoria e a prática. A ferramenta permite que as equipes de desenvolvimento traduzam os princípios éticos em tarefas transparentes, permitindo assim que os princípios sejam abordados durante todo o ciclo de vida do sistema e que sejam discutidos durante os estágios iniciais, onde o trabalho é mais concentrado. Além disso, ela aborda o desenvolvimento monofocal de princípios éticos, permitindo que os vários princípios do conjunto definido sejam abordados por meio de sua integração com as histórias de usuário. Com estas características definidas, a ferramenta pode então ser formalizada.

4.2 Visão Geral

A ferramenta é formalizada como um sistema baseado em LLM que converte requisitos éticos de IA de alto nível em histórias éticas de usuário, com o objetivo de auxiliar equipes de desenvolvimento de software na incorporação prática de princípios éticos em seus sistemas de IA. A ferramenta utiliza as técnicas de *retrieval-augmented generation* [88, 89] e *fine-tuning* [37] para gerar predições. Ao adicionar RAGs ao sistema, a ferramenta combina o uso de um modelo de linguagem com o mecanismo de recuperação de informações em bancos de dados e documentos específicos [88, 89]. Isso permite a busca de referências em padrões éticos, diretrizes legais e casos de uso relevantes para aumento de precisão e contextualização das gerações do modelo. Os diagramas de desenvolvimento e uso da ferramenta podem ser observados na Figura 4.1.

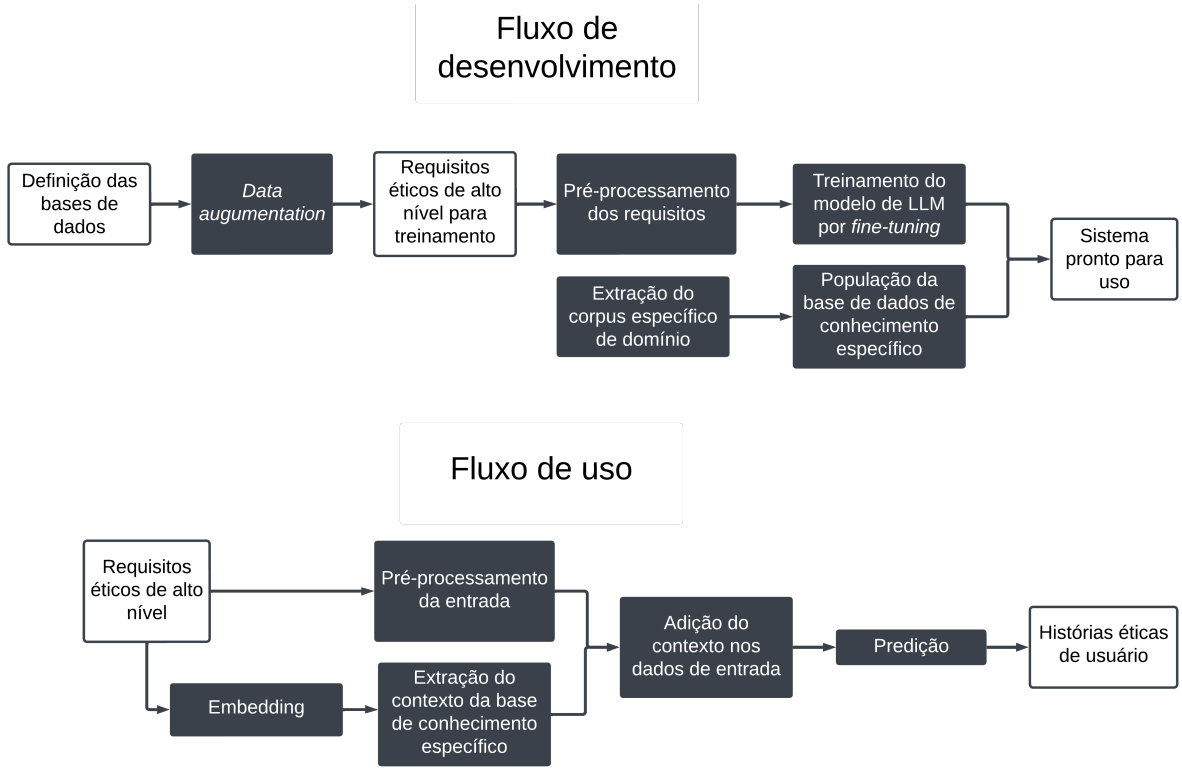
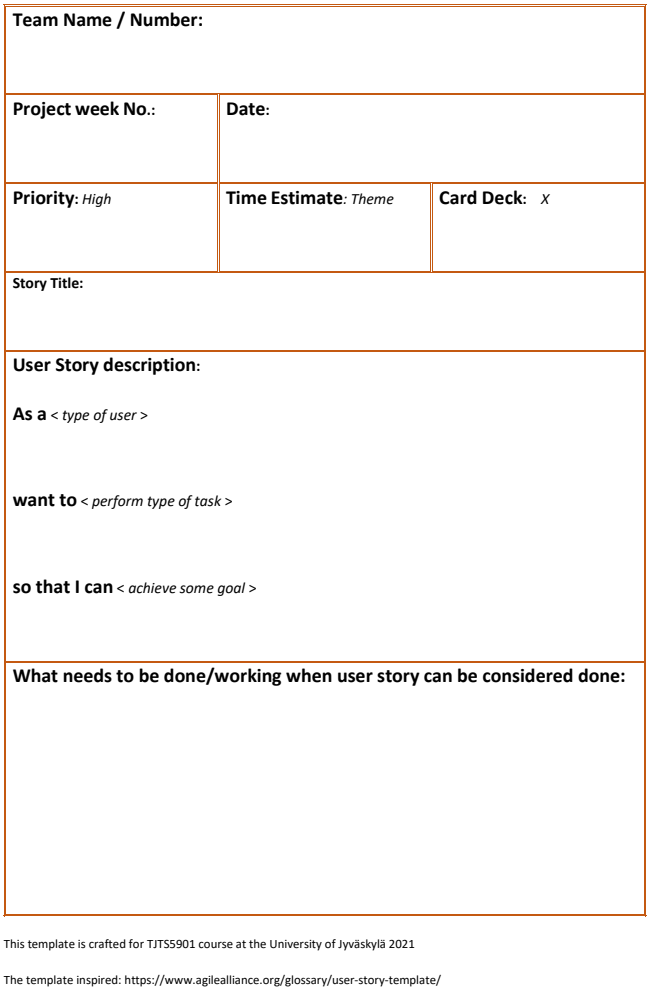


Figura 4.1: Diagramas de desenvolvimento e uso da ferramenta. Fonte: o Autor.

As entradas da ferramenta são requisitos de alto nível escritos em linguagem natural, como pode ser observado nas bases de dados *PROMISE* [90] e *Passenger Flow* [91]. Nestas bases de dados, os requisitos são descritos em um formato simples e legível, permitindo que os usuários especifiquem suas necessidades com clareza, sem a utilização de jargões técnicos ou especificações complexas. Para o padrão de saída das histórias de usuário foi utilizado o estudo de Halme et al. [8]. Neste estudo, os autores apresentaram a utilização de histórias éticas de usuário como meio de promover a integração da ética em IA nas práticas convencionais de engenharia de software. Os autores categorizaram as histórias éticas de usuário como aquelas projetadas para ajudar a abordar e formalizar questões éticas na engenharia de software do ponto de vista de um contexto ético [8].

Em um sistema baseado em IA, a aplicação de histórias éticas de usuário é uma prática benéfica, pois facilita a implementação de princípios éticos, explorando a eficácia de uma das técnicas mais utilizadas na engenharia de requisitos. Isso permite a tradução de considerações éticas em especificações tangíveis, especialmente em modelos ágeis [8]. Portanto, a escolha do estudo se dá por seu pioneirismo ao propor um modelo específico para histórias éticas de usuário. A Figura 4.2 apresenta o *template* definido pelos autores. Informações como nome do time, semana do projeto, data, prioridade e estimação serão preenchidas pelo usuário.



Team Name / Number:		
Project week No.:	**Date**:	
Priority: *High*	**Time Estimate**: *Theme*	**Card Deck**: *X*
Story Title:		
User Story description: **As a** < *type of user* > **want to** < *perform type of task* > **so that I can** < *achieve some goal* >		
What needs to be done/working when user story can be considered done:		

This template is crafted for TJTS5901 course at the University of Jyväskylä 2021

The template inspired: https://www.agilealliance.org/glossary/user-story-template/

Figura 4.2: *Template* de histórias éticas de usuário. Fonte: Halme et al. [8].

4.3 *Dataset*

A obtenção dos dados no contexto de desenvolvimento de soluções éticas em IA assume uma complexidade adicional se comparada aos softwares convencionais, pois os dados são menos acessíveis. O *dataset* utilizado se encontra neste Trabalho[2] e é chamado de *EthicalRequirements4AI*. Esse dataset foi construído por meio da adição de duas bases de dados: *PROMISE* [90] e *Passenger Flow* [91]. A base de dados *PROMISE* [90] é composta por requisitos funcionais e requisitos não funcionais para sistemas de software

[2] https://zenodo.org/records/10938484

convencionais, não relacionados à IA. Já a base de dados *Passenger Flow* [91] é composta por requisitos éticos de IA.

É importante notar que a coluna que contém os requisitos apresentados no formato de história de usuário não está incluída no *dataset*. Por este motivo, esta coluna foi acrescentada ao *dataset* para o treinamento do modelo. Para a geração da coluna de histórias éticas de usuário, foi decidido utilizar um processo de *data augumentation* por meio de LLMs. De acordo com Ding et al. [92], *data augumentation* envolve a adoção de métodos destinados a reforçar a eficácia do modelo por meio do aumento dos dados de treino por meio de dados gerados sinteticamente, evitando assim a necessidade de esforços adicionais de coleta de dados. Ao relacionar este processo com a utilização de LLMs, Ding et al. [92] explicam que:

> Do ponto de vista dos dados, a utilização de LLMs para aumentar os dados representa uma estratégia viável para ultrapassar as limitações relacionadas a coleta de dados, facilitando a criação de conjuntos de dados sintéticos de alta qualidade que, em certos casos, podem revelar-se de maior valor do que os dados selecionados por humanos (Ding et al. [92], 2024, p. 1, tradução nossa).

Com o objetivo de garantir a qualidade das histórias geradas automaticamente, foi realizada a transcrição manual de um subconjunto de dez requisitos, convertendo-os em histórias éticas de usuário. Esse processo ocorreu por meio da leitura detalhada dos requisitos éticos e da aplicação do **template** proposto por Halme et al. [8]. Inicialmente, os autores analisaram os dados disponíveis no **dataset**, selecionando dez requisitos éticos. Em seguida, cada requisito foi transcrito manualmente, seguindo uma abordagem estruturada que incluía a escrita de um título claro e conciso, a formulação de uma descrição detalhada do requisito em um contexto específico e a definição dos critérios de aceitação (work). Esse processo garantiu que as histórias resultantes mantivessem o grau desejado de alinhamento com os princípios éticos, ao mesmo tempo em que serviram como referência para a filtragem dos resultados da geração automática de histórias de usuário.

A seleção dos requisitos a serem utilizados foi uma das etapas mais críticas deste processo. Para isso, os autores consideraram sua familiaridade com os objetivos de cada requisito, garantindo uma adaptação mais precisa ao formato desejado. Esse processo manual proporcionou um controle mais rigoroso sobre a fidelidade das histórias aos requisitos e princípios éticos, além de estabelecer um conjunto inicial de exemplos como referência. No entanto, a abordagem manual também apresenta algumas limitações, como a possibilidade de vieses interpretativos por parte dos autores e a restrição na diversidade de cenários abordados, uma vez que os requisitos foram escolhidos com base na familiaridade com o tema.

Após a avaliação e validação das histórias geradas manualmente, os requisitos restantes foram submetidos ao processo de **data augmentation** por meio de LLMs. Esse

processo utilizou como entrada cada requisito ético definido no **dataset**, juntamente com os utilizados para a geração manual, ampliando a base de exemplos disponíveis sem comprometer a relevância e a coerência das histórias resultantes. Como falado anteriormente, a inclusão das histórias manuais permitiu a comparação com os resultados gerados pelos LLMs, facilitando a filtragem e seleção das histórias para compor o **dataset** final. Dessa forma, a combinação entre histórias manuais e geração automatizada possibilitou a criação de um conjunto mais robusto e diversificado de histórias de usuário alinhadas às diretrizes éticas. Abaixo pode ser observada uma das histórias geradas manualmente.

> **Requisito Ético**: Terminal area, booking documents and booking app contains the help information (for different purposes broken down).
>
> **História Ética de Usuário**
> **Title**: Help Information
>
> **Description**: As a user, i want help information to be clearly available and organized by purpose in the terminal area, booking documents, and booking app, so that i can easily find the guidance for my specific situation without confusion or frustration.
>
> **Work**:
>
> 1. Help information is prominently displayed in the terminal area, booking documents, and app interfaces.
> 2. Help topics are categorized by purpose (booking assistance, cancellation policies, payment issues, accessibility features).
> 3. Help content is accessible in multiple formats (text, audio, video).
> 4. Information is regularly reviewed and updated to ensure accuracy and relevance.

4.4 Preparação dos Dados

O processo de pré-processamento de dados envolve a limpeza e a preparação do texto para etapas subsequentes do fluxo de *machine learning*. Os tipos de pré-processamento vão de acordo com o objetivo final do sistema, não sendo uma fórmula universal [93]. Neste estudo, os algoritmos de pré-processamento foram desenvolvidos utilizando a linguagem de programação Python e estão disponíveis no repositório do Github[3].

Com o objetivo de preparar os dados, foram exploradas as técnicas descritas na literatura e foi identificado um conjunto de técnicas principais de pré-processamento [94, 95]. Dentro deste conjunto de técnicas, foram selecionadas três principais para utilização neste modelo. A primeira técnica é composta da transformação dos caracteres maiúsculos em

[3] https://github.com/joaorossi15/requirements-to-us

caracteres minúsculos com o objetivo de normalizar as entradas [95], e é apresentado no *Listing* 1. A segunda técnica é a remoção de espaços redundantes para garantir a formatação correta [95] e o código relacionado a ela pode ser observado no *Listing* 2. De acordo com Palomino & Aider [94], "a tokenização de um fluxo de caracteres em uma sequência de elementos semelhantes a palavras é um dos componentes mais críticos do pré-processamento de texto". Portanto, as entradas foram transformadas em tokens, como se pode observar no *Listing* 3.

- Transformação do conjunto de caracteres da entrada em caracteres minúsculos: Substituição de caracteres maiúsculos por minúsculos para normalizar a formatação dos textos de entrada.

Listing 1: Função de Transformação em Minúsculo

```
1    def text_to_lower(text: str) -> str:
2        return text.lower()
```

Fonte: o Autor

- Remoção de espaços redundantes: Remoção de espaços desnecessários para normalizar a formatação dos textos de entrada.

Listing 2: Função de Remoção de Espaços Redundantes

```
1    def pad_multiple_spaces(text: str) -> str:
2        return re.sub( r"\s+"," ",text)
```

Fonte: o Autor

- Tokenização dos caracteres de entrada: O processo de tokenização é de extrema importância para o processamento de linguagem natural. Modelos de LLM utilizam tokens que representam unidades semânticas básicas para compreender melhor a semântica do texto [96].

Listing 3: Função de Tokenização

```
1    def tokenize_dataset(model_name: str):
2        t = transformers.AutoTokenizer.from_pretrained(model_name,
         ↪  use_fast=True)
3        t.pad_token = t.eos_token
```

```
        data_collator =
        ↪ transformers.DataCollatorForLanguageModeling(t,
        ↪ mlm=False)

        return data_collator, t
```

Fonte: o Autor

O fluxo de aplicação de pré-processamento segue a seguinte linha: transformação do conjunto de caracteres da entrada em letra minúscula, remoção de espaços redundantes e finalmente a tokenização dos caracteres. O Trecho de Código 4 encapsula as etapas de transformação e remoção de espaços, sendo seguida pela aplicação da função de tokenização.

Listing 4: Aplicação das Etapas de Pré-processamento

```
def prepare_dataset(df: pd.DataFrame) -> pd.DataFrame:
    df['text'] = df['text'].apply(text_to_lower)
    df['ethical_us'] = df['ethical_us'].apply(text_to_lower)
    df['text'] = df['text'].apply(pad_multiple_spaces)
    df['ethical_us'] = df['ethical_us'].apply(pad_multiple_spaces)
    instruction = construct_instruction()
    df['data'] = df.apply((lambda row: apply_in_row(row, instruction)),
    ↪ axis=1)

    return df

df = prepare_dataset(df)
data_collator, t = tokenize_dataset(model_name)
```

Fonte: o Autor

A construção de um modelo de *deep learning* requer a utilização de uma quantidade considerável de dados. No entanto, a utilização de dados que não tenham sido corretamente separados pode resultar em um desafio notável no treinamento, conhecido como *overfitting*. Para Valdenegro-Toro & Sabatelli [97], overfitting pode ser definido como a falta de generalização em um modelo de *machine learning*.

Portanto, após o pré-processamento, os dados foram divididos em dois conjuntos separados: um conjunto de treino e um conjunto de teste. Isto foi realizado por meio

da utilização da técnica de validação *hold-out*. Yadav & Shuckla [98] definem a técnica *hold-out* como:

> *Hold-out validation* foi proposta para eliminar o problema de *overfitting* que existia na validação de re-substituição. Os dados são divididos em duas partes que não se sobrepõem e estas duas partes são utilizadas para treinar e testar, respectivamente. A parte que é utilizada para o teste é a parte de *hold-out*. Tem este nome porque retiramos essa parte para teste e o modelo aprende utilizando a parte restante dos dados (Yadav & Shuckla [98], 2016, p. 79, tradução nossa).

A proporção de dados atribuídos a teste e treino pode variar. Normalmente, é utilizada uma divisão de 20% de teste e 80% de treino, que pode ser ajustada de acordo com o modelo específico. No caso de uma porcentagem de 10%, existe um risco de *overfitting* [98]. Portanto, foi selecionada uma porcentagem de 20% para os dados de teste neste estudo. No Trecho de Código 5 é observada esta divisão dentro do código do modelo.

Listing 5: Divisão de Dados de Treino e Teste

```
def train_test_split(df: pd.DataFrame) -> pd.DataFrame:
    df = datasets.Dataset.from_pandas(df)
    df = df.train_test_split(test_size=0.2)

    return df
```

Fonte: o Autor

4.5 Desenvolvimento do Modelo

O processo de desenvolvimento do modelo foi composto de um conjunto de etapas que, em ordem, contribuíram para a definição do modelo utilizado, treinamento por *fine-tuning* e hiperparametrização. Em cada etapa, os resultados foram submetidos a uma avaliação, o que permitiu a aplicação de modificações em caso necessário e em tempo real. A primeira etapa consistiu na escolha do modelo utilizado, e para isso, foi utilizada a plataforma Hugging Face[4], que possui um grande catálogo de modelos pré-treinados para diversas finalidades, incluindo PLN.

A fim de identificar o modelo mais adequado, foi estabelecida uma série de modelos para efeitos de avaliação, utilizando o *dataset* deste projeto. Este conjunto inclui os seguintes modelos: Falcon40B [99], Falcon7B [99], BERT [100], RoBERTa [101] e Mistral7B [102]. Os modelos foram avaliados de acordo com os parâmetros definidos na Tabela 4.1

[4] https://huggingface.co/models

e o modelo Mistral7B [102] demonstrou o melhor desempenho, produzindo as respostas mais otimizadas sobre o tema. Cada parâmetro possui um conjunto de valores associado que foi utilizado para alcançar os resultados. O processo de treinamento por *fine-tuning* dos modelos foi realizado na plataforma Google Collab[5]. Os parâmetros e os valores foram definidos por serem considerados os básicos e padrões de um modelo de IA [103, 104]. Os valores para *batch size* e *number of epochs* foram ajustados para serem inferiores aos valores padrão devido aos recursos computacionais disponíveis para treinar os modelos. No Trecho de Código 6 podem ser observadas as configurações do processo de *fine-tuning* do modelo.

Tabela 4.1: Parâmetros definidos para o teste dos modelos

Parâmetro	**Valores Testados**	**Valores Finais**
Learning Rate	1e-2, 1e-3, 1e-4 e 1e-5	**1e-3**
Number of Epochs	3, 4, 5 e 6	**3**
Batch Size	4, 5 e 6	**4**
Weight Decay	1e-1, 1e-2 e 1e-3	**1e-1**

Fonte: o Autor

Listing 6: Aplicação do processo de *fine-tuning*

```
    def train_model(model, lr, batch_size, num_epochs, tokenized_data,
    ↪  collator):
        ... # QLORA steps

        training_args = transformers.TrainingArguments(
            output_dir= "checkpoints_output",
            learning_rate=lr,
            per_device_train_batch_size=batch_size,
            per_device_eval_batch_size=batch_size,
            num_train_epochs=num_epochs,
            weight_decay=0.01,
            logging_strategy="epoch",
            evaluation_strategy="epoch",
            save_strategy="epoch",
            load_best_model_at_end=True,
            warmup_steps=2,
            fp16=True,
            optim="paged_adamw_8bit",
```

[5] https://colab.research.google.com/

```
        )

        trainer = transformers.Trainer(
            model=model,
            train_dataset=tokenized_data["train"],
            eval_dataset=tokenized_data["test"],
            args=training_args,
            data_collator=collator
        )

        # train model
        model.config.use_cache = False  # silence the warnings
        trainer.train()

        # renable warnings
        model.config.use_cache = True

        return model
```

Fonte: o Autor

O método de *fine-tuning* foi utilizado para o treinamento do modelo. No entanto, é necessário levar em consideração que o treino é um processo custoso, uma vez que muitos modelos requerem mais de 30 GB de memória GPU. Uma das metodologias mais comuns para contornar este problema é a quantização. No entanto, estas técnicas só são eficazes durante a fase de inferência e são pouco úteis durante a fase de treino [105].

Para preencher esta lacuna foi proposto o método QLORA, que utiliza uma técnica para quantizar um modelo pré-treinado para quatro bits e, em seguida, adiciona um pequeno conjunto de pesos adaptáveis [105]. Estes pesos são ajustados por gradientes durante a *backpropagation* por meio dos pesos quantizados [105]. O QLORA introduz uma combinação de três componentes. O primeiro é a quantização utilizando NormalFloat de quatro bits, que é teoricamente o tipo ótimo para dados normalmente distribuídos [105]. O segundo é a quantização dupla, que funciona ao quantificar as constantes de quantização, economizando uma média de cerca de 0,37 bits por parâmetro [105]. Por fim, são introduzidos os otimizadores de paginação, que usam a memória unificada NVIDIA para evitar os picos de memória de pontos de controle de gradiente ocorridos ao processar um *mini-batch* de sequência longa [105]. O Trecho de Código 7 ilustra a forma como o método é utilizado no código do modelo.

Listing 7: Aplicação da Técnica QLORA

```python
def train_model(model, lr, batch_size, num_epochs, tokenized_data,
    collator):
    model.train() # training state
    model.gradient_checkpointing_enable()
    model = peft.prepare_model_for_kbit_training(model) # turn into
        qlora

    # lora config
    config = peft.LoraConfig(
        r=32,
        lora_alpha=64,
        target_modules=["q_proj"],
        lora_dropout=.1,
        bias="none",
        task_type="CAUSAL_LM"
    )
    config.inference_mode = False

    model = peft.get_peft_model(model, config) # model in lora style

    ... # training steps
```

Fonte: o Autor

4.6 *Retrieval-Augmented Generation*

Para melhorar o resultado das previsões foi utilizado o método de *Retrieval-Augmented Generation* (RAG). O método RAG combina o uso de um modelo de linguagem com o mecanismo de recuperação de informações em bancos de dados e documentos específicos, normalmente utilizando bancos de dados vetoriais ou baseados em grafo, dependendo da aplicação [88, 89]. O funcionamento do sistema se dá pela busca de documentos ou informações relevantes na base de dados usando um módulo de recuperação, e depois a LLM utiliza essas informações para criar respostas informadas e contextuais [88, 89]. Com essa combinação, as *queries* se tornam mais contextualizadas, facilitando a geração de texto pelo modelo.

O método mais comum de uso é chamado de *Naive RAG*, a primeira versão proposta, é composta de três fases [89]. A fase inicial é chamada de fase de indexação, durante o qual os documentos são recebidos e seus dados brutos são extraídos, transformados em um

formato de texto uniforme e divididos em unidades menores e de mais fácil processamento, chamadas de blocos de texto [89]. Por fim, essas unidades são então codificadas em representaç ões vetoriais usando um modelo de incorporação e armazenadas em um banco de dados vetorial. Na segunda fase, chamada de fase de recuperação, o sistema RAG utiliza o mesmo modelo usado durante a indexação para transformar a solicitação em um vetor. Em seguida, é verificado o grau de similaridade desse vetor com os blocos de texto indexados. O sistema escolhe os N blocos principais que melhor correspondem à solicitação e são incorporados no *prompt* [89]. A terceira e última fase é chamada de fase de geração. Nesta fase, a *query* apresentada e os documentos selecionados são sintetizados em um *prompt* coerente e repassado ao modelo para a geração da resposta [89]. No Trecho de Código 8 podem ser observadas as funções que representam a fase de indexação, coordenando a criação da base de dados vetorial contendo os documentos de conhecimento específico para o RAG.

Listing 8: Construção da base de dados para o RAG

```
    def load_documents(dir: str):
        doc_loader = PyPDFDirectoryLoader(dir)
        return doc_loader.load()

    def split_text(docs: list[Document]):
        text_splitter = RecursiveCharacterTextSplitter(
            separators=["\n\n", "\\n", "\n", ".",],
            chunk_size=500,
            chunk_overlap=150,
            length_function=len,
        )
        chunks = text_splitter.split_documents(docs)
        print(f"Split {len(docs)} documents into {len(chunks)} chunks.")

        return chunks

    def chroma(chunks: list[Document], path: str):
        if os.path.exists(path):
            shutil.rmtree(path)

        db = Chroma.from_documents(
                chunks,
                HuggingFaceEmbeddings(model_name=
```

```
                'sentence-transformers/all-mpnet-base-v2'
                ),
                persist_directory=path)
        db.persist()
        print(f'Saved {len(chunks)} chunks to {path}')

    def generate_store(path: str, chroma_path: str):
        documents = load_documents(path)
        chunks = split_text(documents)
        chroma(chunks, chroma_path)
```

Fonte: o Autor

4.7 Utilização da Ferramenta

Para utilizar a ferramenta proposta, é necessário, inicialmente, acessar o repositório disponível no github e realizar o download do repositório completo para o ambiente local. Após o download, o usuário deve configurar o ambiente de desenvolvimento, seguindo as instruções fornecidas no repositório, e hospedar a aplicação localmente. Se preferir, é possível apenas subir o arquivo **main.ipynb** para uma instância do **Google Colab** e rodar as duas primeiras células presentes. Em seguida, é preciso popular a base de dados vetorial utilizada durante o processo de **Retrieval-Augmented Generation** (RAG), garantindo que o sistema tenha acesso às informações necessárias para processar os requisitos éticos e gerar histórias de usuário adequadas. Para popular a base, basta adicionar os arquivos de texto desejados na pasta **/rag-data**. Esse processo inicial de configuração é essencial para garantir o funcionamento correto da ferramenta dentro do contexto desejado pelo usuário. Após estas etapas, o usuário pode utilizar o sistema por meio do arquivo **main.ipynb**.

O funcionamento da ferramenta segue o fluxo ilustrado na Figura 4.3. O usuário deve fornecer um ou mais requisitos éticos como entrada para o sistema. A partir disso, dois processos ocorrem simultaneamente: a tokenização e a conversão dos requisitos em vetores. A tokenização transforma os requisitos em tokens específicos, que serão utilizados diretamente na construção do **prompt** do modelo, enquanto a conversão vetorial é empregada no processo de **Retrieve** do módulo **RAG**. Após a etapa de recuperação (**Retrieve**), o contexto relevante é extraído do banco de dados e combinado com o requisito original para a geração do **prompt**. Finalmente, esse **prompt** enriquecido é processado pelo modelo

de linguagem, que gera como saída a história ética de usuário correspondente. A seguir, apresenta-se um exemplo de história gerada pelo sistema.

> **Title**: User Consent for Decision Making
>
> **Description**: As a user, i want the AI to obtain my explicit consent before making decisions on my behalf so that i maintain control over my decisions and ensure that my autonomy is respected.
>
> **Work**:
>
> 1. Implement a mechanism for users to provide explicit consent for AI to make decisions on their behalf.
> 2. Ensure that the AI does not initiate any action without obtaining prior consent from the user.
> 3. Provide clear communication to users about what decisions the AI will make and when it requires consent.
> 4. Develop a user interface that allows users to easily revoke or modify their consent settings at any time.

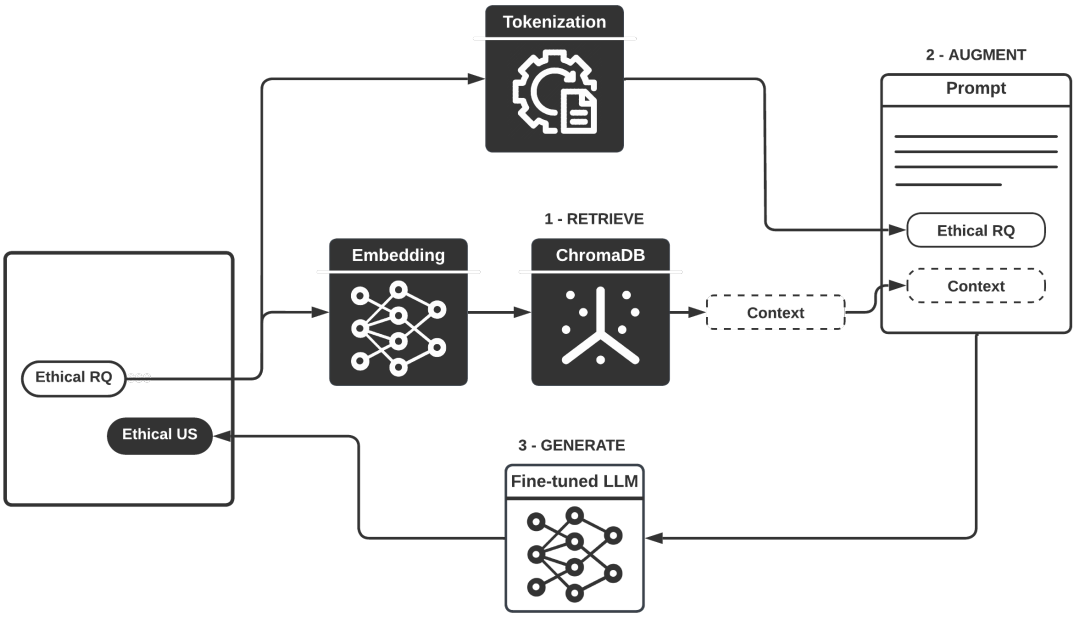


Figura 4.3: Fluxo Completo da Ferramenta. Fonte: o Autor.

4.8 Validação da Ferramenta

A validação da ferramenta foi realizada através de um questionário aplicado para 30 participantes, incluindo estudantes de graduação e pós-graduação, com o objetivo de avaliar a capacidade da ferramenta de transformar requisitos éticos de IA em histórias

de usuário de maneira eficiente. O conjunto de questões do formulário é apresentado na Tabela 4.2.

Tabela 4.2: Questões do Formulário

ID	Questão
Q1	Qual a sua idade?
Q2	Qual é a sua experiência com atividades de desenvolvimento de software?
Q3	Qual papel você desempenha na sua equipe de desenvolvimento de software?
Q4	Familiaridade com: [Requisitos de software]
Q5	Familiaridade com: [Histórias de usuário]
Q6	Familiaridade com: [Requisitos éticos]
Q7	Familiaridade com: [Histórias éticas de usuário]
Q8	Familiaridade com: [Princípios éticos de IA]
Q9	Familiaridade com: [Precisão: Sistemas de IA devem produzir resultados confiáveis.]
Q10	Familiaridade com: [Autonomia: Sistemas de IA devem ser projetados para a autonomia humana.]
Q11	Familiaridade com: [Bem-estar social: Promover o bem-estar humano, social e ambiental.]
Q12	Familiaridade com: [Beneficência: Sistemas de IA devem promover o bem.]
Q13	Familiaridade com: [Confiabilidade: Sistemas de IA devem ser confiáveis em sua operação.]
Q14	Familiaridade com: [Dignidade humana: Sistemas de IA não devem comprometer valores humanos.]
Q15	Familiaridade com: [Diversidade: Sistemas de IA devem promover a diversidade.]
Q16	Familiaridade com: [Eficácia: Sistemas de IA devem ser eficazes social e tecnicamente.]
Q17	Familiaridade com: [Equidade: Sistemas de IA não devem discriminar.]
Q18	Familiaridade com: [Explicabilidade: Compreender como a IA toma decisões autônomas.]
Q19	Familiaridade com: [Interpretabilidade: Os processos de tomada de decisão da IA devem ser interpretáveis.]
Q20	Familiaridade com: [Justiça: Sistemas de IA devem promover justiça e eliminar discriminações.]
Q21	Familiaridade com: [Legalidade: Sistemas de IA devem cumprir padrões legais.]
Q22	Familiaridade com: [Não discriminação: Sistemas de IA não devem agir de forma discriminatória.]

Tabela 4.2: Questões do Formulário

ID	Questão
Q23	Familiaridade com: [Não maleficência: Evitar o uso indevido de sistemas de IA.]
Q24	Familiaridade com: [Previsibilidade: As implicações de longo prazo da IA devem ser previsíveis.]
Q25	Familiaridade com: [Privacidade: Garantia da privacidade de usuários e dados.]
Q26	Familiaridade com: [Prosperidade: Compartilhar a prosperidade impulsionada pela IA.]
Q27	Familiaridade com: [Responsabilidade: Garantir justiça atribuindo responsabilidades.]
Q28	Familiaridade com: [Robustez: Sistemas de IA devem lidar com ataques e manter a sua funcionalidade.]
Q29	Familiaridade com: [Segurança: Sistemas de IA devem evitar causar danos.]
Q30	Familiaridade com: [Governança de dados: Garantia da segurança dos dados em sistemas de IA.]
Q31	Familiaridade com: [Solidariedade: Compartilhar produtividade e resultados de sistemas de IA.]
Q32	Familiaridade com: [Supervisão humana: Apoiar a tomada de decisão humana.]
Q33	Familiaridade com: [Sustentabilidade: Sistemas de IA devem ser eticamente sustentáveis.]
Q34	Familiaridade com: [Transparência: Os processos de tomada de decisão da IA devem ser transparentes.]
Q35	Percepção: [A ferramenta facilita a compreensão dos requisitos éticos.]
Q36	Percepção: [As histórias éticas geradas foram escritas de forma clara e coerente.]
Q37	Percepção: [Os requisitos éticos foram representados adequadamente nas histórias de usuário.]
Q38	Percepção: [As histórias de usuário geradas estão alinhadas com todos os princípios éticos relacionados.]
Q39	Percepção: [A ferramenta gera histórias de usuário de maneira mais eficiente do que manualmente.]
Q40	Percepção: [Eu recomendaria essa ferramenta a outros estudantes ou desenvolvedores de software.]
Q41	Percepção: [As histórias de usuário geradas pela ferramenta são fáceis de entender.]

Tabela 4.2: Questões do Formulário

ID	Questão
Q42	Percepção: [A ferramenta é uma adição valiosa ao processo de geração de histórias éticas de usuário.]
Q43	Descreva sua percepção sobre a ferramenta (aspectos positivos, negativos ou ambos).
Q44	Quais são suas sugestões para melhorar a ferramenta?

Fonte: o Autor.

Os participantes, que tinham predominantemente menos de 25 anos de idade, apresentavam uma variedade de níveis de experiência em desenvolvimento de software, com a maioria tendo entre um e quatro anos de experiência. Na sua maioria, eram desenvolvedores (n=26), enquanto outros eram analistas de negócios (n=2), testadores (n=1) e engenheiros de requisitos (n=1). Mais detalhes sobre o perfil dos participantes podem ser encontrados na Tabela 4.3. Embora a maioria dos participantes tenha demonstrado familiaridade com requisitos de software e histórias de usuários, a compreensão dos princípios éticos de IA apresentou uma variação considerável. Alguns participantes indicaram uma exposição prévia mínima, enquanto outros demonstraram maior compreensão. No entanto, houve um grau discernível de familiaridade com alguns princípios éticos, como por exemplo acurácia, a autonomia e a sustentabilidade.

Tabela 4.3: Perfil (n=30)

Age	%
Menos de 21	20
21-25	66.7
26-30	13.3
Anos de Experiência	%
Menos de 1	13.3
1-2	50
3-4	30
5-7	6.7
Papel	%
Analista de Negócios	6.7
Engenheiro de Requisitos	3.3
Testador	3.3
Engenheiro de Software	86.7

Fonte: o Autor.

O questionário utilizou uma abordagem de métodos mistos, integrando questões quantitativas e qualitativas para avaliar o desempenho da ferramenta em várias dimensões. Quanto às questões objetivas, as respostas forneceram uma visão abrangente das percepções dos participantes sobre o desempenho da ferramenta. A avaliação utilizou um formato de escala likert, em que os participantes classificaram o seu nível de concordância com afirmações específicas sobre a clareza e a eficácia da ferramenta. Os percentuais de cada afirmação foram calculados, destacando tanto os pontos fortes da ferramenta como as potenciais áreas de melhoria. Depois de os participantes terem preenchido o questionário, os resultados foram submetidos a uma análise, tendo sido calculado o percentual de cada categoria para cada pergunta, a fim de facilitar uma compreensão mais abrangente da ferramenta. As questões e os seus identificadores são definidos na Tabela 4.4.

Tabela 4.4: Identificadores de cada Questão

ID	**Item**
QP1	Recomendaria esta ferramenta a outros estudantes ou programadores de software
QP2	A ferramenta é um complemento importante para o processo de criação de histórias éticas de usuário
QP3	As histórias de usuário geradas pela ferramenta são de fácil compreensão
QP4	A ferramenta gera histórias de usuário de forma mais eficiente do que manualmente
QP5	As histórias éticas geradas foram escritas de forma clara e coerente
QP6	A ferramenta facilita a compreensão dos requisitos éticos
QP7	Os requisitos éticos foram adequadamente representados nas histórias de usuário
QP8	As histórias de usuário geradas estão em conformidade com todos os princípios éticos relacionados

Fonte: o Autor.

Os participantes demonstraram um elevado nível de consenso quanto à recomendação e utilidade geral da ferramenta, com 93.3% dos respondentes estando entre "Concordo"e "Concordo totalmente", atribuído à disposição de a recomendar a outros estudantes e programadores. Do mesmo modo, a ferramenta foi considerada útil na criação de histórias éticas de usuário, indicando a sua aplicação prática em cenários do mundo real. A clareza foi identificada como outro ponto forte, com os participantes considerando que as histórias de usuário geradas eram fáceis de entender e escritas de forma clara. Além disso, a ferramenta foi reconhecida pela sua eficácia devido a sua capacidade de gerar histórias de usuário de uma forma mais eficiente em termos de tempo do que manualmente. Também

foi demonstrada a eficácia em ajudar os usuário a compreender os requisitos éticos. No entanto, o alinhamento das histórias de usuário com os requisitos éticos, embora ainda possuindo a maioria de respostas entre "Concordo"e "Concordo totalmente", apresentou um pequeno declínio, sugerindo potencial de melhoria para garantir uma representação completa. O resultado mais baixo está relacionado à integração de todos os princípios éticos, com 47% dos resultados (14 respondentes) selecionando a opção "Neutro", o que indica que alguns participantes consideraram que certos aspectos éticos não foram corretamente abordados. Isto pode ser atribuído a um desequilíbrio dos dados de treino, o que era esperado, dado que o *dataset* é recente e está sendo atualizado. Estes resultados destacam a eficácia da ferramenta, ao mesmo tempo que identificam áreas que podem ser melhoradas para aumentar a sua abrangência e alinhamento com as regras éticas. Os resultados são ilustrados na Tabela 4.5 e na Figura 4.4.

Tabela 4.5: Quantidade de Respostas para cada Questão

ID	**Discordo totalmente**	**Discordo**	**Neutro**	**Concordo**	**Concordo totalmente**
QP1	0	0	2	6	22
QP2	0	0	1	11	18
QP3	0	0	2	10	18
QP4	0	0	2	10	18
QP5	0	0	2	12	16
QP6	0	0	2	15	13
QP7	0	0	3	14	13
QP8	2	2	14	11	1

Fonte: o Autor.

O feedback qualitativo destacou o carácter inovador e prático da ferramenta. Os participantes apreciaram a sua facilidade de uso e clareza, tendo vários observado o seu potencial para poupar tempo e melhorar a eficiência do fluxo de trabalho. As sugestões de melhoria incluíram a criação de uma interface gráfica, o aperfeiçoamento da formatação dos resultados para uma melhor legibilidade e o aumento da inclusão de alguns princípios éticos, especialmente os menos considerados uma prioridade no domínio, como a sustentabilidade. Abaixo podem ser vistas três respostas das questões abertas. Na Tabela 4.6, são apresentadas as palavras-chave identificadas nas respostas qualitativas, as quais refletem as opiniões dos participantes acerca da ferramenta, juntamente com os contextos em que tais citações foram empregadas. Essa análise permite compreender as percepções predominantes sobre a utilidade, funcionalidades e impactos da ferramenta, destacando tanto os aspectos positivos quanto as possíveis limitações apontadas pelos usuários.

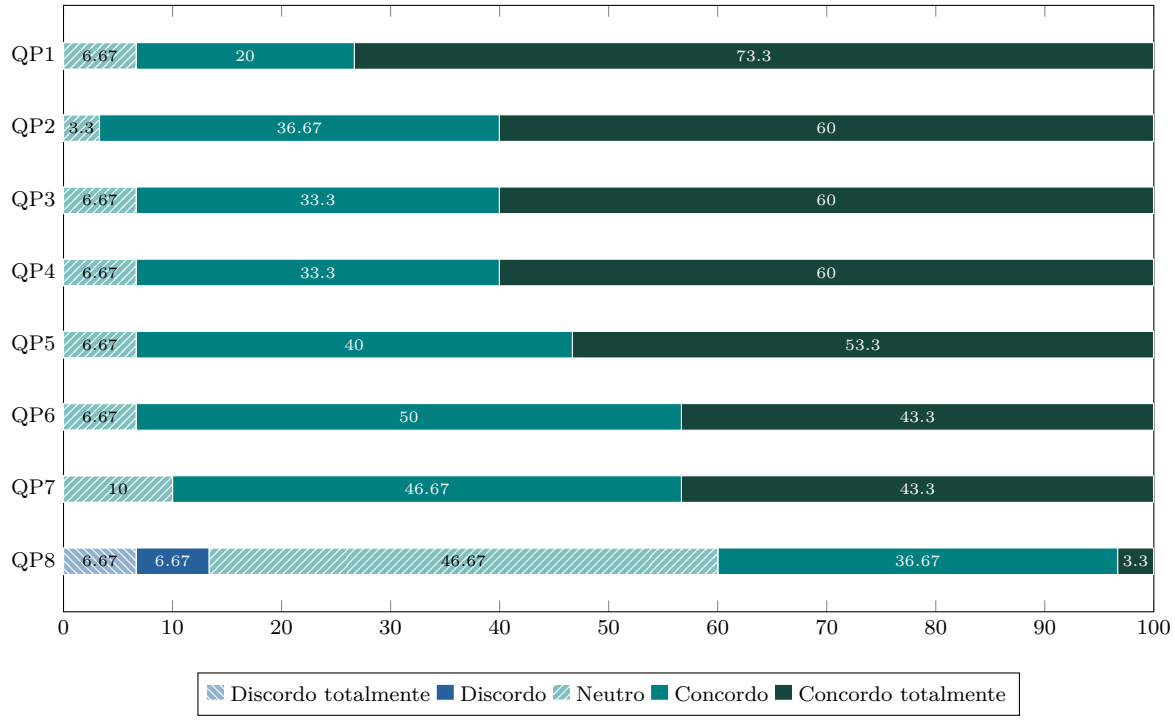


Figura 4.4: Percepções dos entrevistados. Fonte: o Autor.

"A ferramenta tem uma utilidade muito grande, visto que auxilia o trabalho do programador no sentido de definição dos requisitos éticos. O que antes levava um tempo muito grande para ser implementado, agora pode ser feito em um período menor, de forma que o programador não precisa "se preocupar"tanto com essa parte, será necessário apenas revisar os requisitos gerados pela ferramenta. [...]"

"Achei a ferramenta bem útil, especialmente para facilitar a criação de histórias de usuários com foco em ética. Ela ajuda a entender os requisitos éticos de maneira clara e acelera bastante o processo, que seria bem mais demorado se fosse feito manualmente. Além disso, as histórias são escritas de forma simples e fácil de entender, o que é ótimo para compartilhar com a equipe."

"[...] A ferramenta é muito importante não só para ajudar a diminuir os custos de fabricação de um software, mas para ajudar desenvolvedores a limitar os vieses em seus programas, possibilitando a criação de aplicações mais justas e que podem ser usadas por mais usuários [...]"

Embora a ferramenta tenha sido frequentemente descrita como "Útil", "Precisa"e "Relevante", a presença da palavra-chave "Limitada"levanta questões importantes sobre a capacidade da ferramenta de abordar plenamente todos os princípios éticos. Os participantes observaram que certos princípios éticos, como a sustentabilidade, não foram consisten-

Palavra-chave	Significado Geral/Contexto
Útil	A ferramenta é frequentemente descrita como útil para gerar histórias de usuários éticas, economizando tempo e esforço dos desenvolvedores.
Prática	Muitos usuários consideraram a ferramenta prática, pois simplifica o processo de criação de histórias de usuários éticas.
Inovadora	A ferramenta é frequentemente elogiada por sua abordagem inovadora para tratar considerações éticas no desenvolvimento de software.
Precisa	Os usuários destacaram a precisão da ferramenta em gerar histórias de usuários éticas coerentes e relevantes.
Fácil de usar	A ferramenta é descrita como amigável, com uma interface simples e saídas claras que são fáceis de entender.
Eficiente	Muitos respondentes notaram a eficiência da ferramenta em automatizar tarefas, acelerando o processo de desenvolvimento de software.
Relevante	A ferramenta é considerada relevante, especialmente no contexto do aumento da atenção às considerações éticas no desenvolvimento de software.
Clara	As saídas da ferramenta são descritas como claras e de fácil entendimento, facilitando a colaboração entre os membros da equipe.
Importante	Os usuários enfatizaram a importância da ferramenta para abordar requisitos éticos, que muitas vezes são negligenciados em projetos de software.
Limitada	Alguns usuários notaram limitações, como a necessidade de um banco de dados mais amplo ou a necessidade de revisão humana para refinar as saídas da ferramenta.

Tabela 4.6: Palavras-chave mais usadas que representam opiniões sobre a ferramenta

temente integrados nas histórias de usuário geradas. Isso sugere que a ferramenta pode exigir refinamentos adicionais para garantir uma cobertura abrangente das considerações éticas. Em segundo lugar, as palavras-chave "Fácil de usar"e "Eficiente"destacam o design amigável da ferramenta e sua capacidade de economizar tempo. No entanto, ainda não está claro como essas características se traduzem na adoção real por equipes de desenvol-

vimento. Por exemplo, a facilidade de uso da ferramenta encoraja mais desenvolvedores a integrar considerações éticas em seus fluxos de trabalho? Além disso, como a eficiência da ferramenta se compara aos métodos manuais em termos de tempo e alocação de recursos? A exploração dessas questões pode fornecer dicas valiosas sobre a aplicabilidade prática da ferramenta e seu potencial para simplificar a integração de requisitos éticos em projetos de desenvolvimento de software.

Por fim, a palavra-chave "Limitada"reforça a necessidade de abordar as limitações atuais da ferramenta. Os participantes identificaram lacunas na capacidade da ferramenta de representar plenamente os requisitos éticos, especialmente em casos onde certos princípios éticos não foram adequadamente integrados. Isso levanta a questão de como essas limitações podem ser mitigadas. Por exemplo, a expansão do conjunto de dados para incluir cenários éticos mais diversos poderia melhorar o desempenho da ferramenta? Além disso, como a integração de **frameworks** ou diretrizes éticas adicionais poderia aprimorar a capacidade da ferramenta de gerar histórias de usuário mais abrangentes? Abordar essas questões poderia levar a melhorias significativas na funcionalidade da ferramenta, garantindo que ela atenda às necessidades em evolução de desenvolvedores e partes interessadas no campo da IA ética.

Essas questões em aberto destacam a importância de pesquisas e desenvolvimentos contínuos para refinar a ferramenta e garantir sua eficácia em preencher a lacuna entre os princípios éticos teóricos e o desenvolvimento prático de software. Ao abordar essas questões, trabalhos futuros podem construir sobre os pontos fortes atuais da ferramenta enquanto resolvem suas limitações, contribuindo, em última análise, para o desenvolvimento e a implantação responsáveis de sistemas de IA.

4.8.1 Ameaças da Validação

O processo de validação da ferramenta desenvolvida para traduzir requisitos éticos de IA em histórias de usuários éticas esteve sujeito a diversas potenciais ameaças à validade. A validade interna refere-se à medida em que os resultados obtidos podem ser atribuídos à utilização da própria ferramenta, excluindo influências externas. Uma ameaça relevante foi a variação do conhecimento prévio dos participantes sobre o campo de ética em IA. Embora a maioria dos participantes tivesse experiência em desenvolvimento de software, o nível de familiaridade com os princípios éticos variou, o que pode ter afetado a capacidade de avaliar os resultados da ferramenta de forma adequada. Além disso, a utilização de dados obtidos através de questionários pode ter introduzido enviesamentos, nos quais os participantes podem ter dado à ferramenta classificações mais positivas do que seria justificável.

A validade externa é utilizada para descrever a possibilidade de generalizar os resultados obtidos num estudo específico a outras situações e contextos. Embora o grupo de participantes tenha sido composto maioritariamente por estudantes de graduação e pós-graduação, muitos dos quais já trabalham na indústria de desenvolvimento de software, o que reduz parcialmente esta limitação. No entanto, seria positivo aumentar a diversidade de experiências profissionais e de níveis de senioridade para melhorar a representatividade. Além disso, o tamanho relativamente pequeno da amostra, com apenas 30 participantes, restringe a abrangência dos resultados. Para fortalecer essa generalização, seria recomendável incluir um número maior de participantes, com perfis profissionais mais diversos.

Por fim, a validade de construto envolve garantir que a avaliação mede o que se pretende medir. O instrumento de pesquisa utilizado para validação incluiu questões em escala Likert para avaliar a clareza, a eficácia e a relevância da ferramenta. No entanto, certos construtos, como a adequação da representação dos princípios éticos, podem ter sido interpretados de maneira diferente pelos participantes, o que pode levar a um possível desalinhamento entre os resultados pretendidos e os resultados obtidos. Além disso, a dependência das percepções subjetivas dos participantes pode não capturar totalmente a capacidade técnica da ferramenta.

4.9 Discussões

As conclusões iniciais deste estudo indicam uma falta de ferramentas adequadas para a aplicação da ética em IA em sistemas reais. Esta conclusão é exposta no Capítulo 2, que apresenta em detalhe os resultados da revisão sistemática da literatura realizada. De forma a colaborar com aplicação da ética em IA, foi proposta uma ferramenta baseada em LLM que realiza a tradução de requisitos éticos em histórias éticas de usuário. Foram identificadas diversas ferramentas que utilizam LLMs no contexto de histórias de usuário. Zhang et al. [106] exploram o desenvolvimento de uma ferramenta que automatiza a melhoria da qualidade de histórias de usuário por meio da integração de agentes baseados em LLM em ambientes reais. Luitel et al. [46] desenvolveram uma maneira de utilizar um sistema baseado em *Large Language Models* (LLMs) para gerar previsões contextualizadas para o preenchimento de requisitos de software. Estes estudos serviram como base para a ferramenta.

Neste contexto, a ferramenta serve como um meio prático de ligação entre a ética em IA e a engenharia de requisitos, permitindo a automatização deste processo. Halme et al. [8] propuseram a noção da utilização de histórias éticas de usuário como meio de facilitar a integração da ética em IA nas práticas convencionais de engenharia de software. E com

base nessa proposta o modelo foi desenvolvido. Durante o desenvolvimento da ferramenta, ficou evidente que a disponibilidade limitada de dados relacionados à ética em IA foi um desafio para o treinamento do modelo. Para resolver este problema, foi utilizado o processo de *data augmentation* em conjunto com a técnica de *retrieval-augmented generation*. Após a aplicação de RAG, os resultados gerados pelo modelo se tornaram mais relevantes, facilitando assim a utilização do sistema em casos reais.

A ferramenta foi bem recebida, tendo a maioria dos participantes indicado a vontade de a recomendar a outros. Estes resultados demonstram o potencial da ferramenta como um recurso importante para os programadores e pesquisadores na criação de histórias de usuário eticamente alinhadas, ao mesmo tempo que fornece informações concretas para um aperfeiçoamento posterior. As pontuações altas relacionadas com a clareza e a eficiência demonstram que a ferramenta aborda efetivamente os desafios comuns na tradução de princípios éticos em histórias éticas de usuário, reduzindo significativamente o tempo e o esforço manual normalmente necessários para este processo. No entanto, há espaço para melhorias em termos de garantir que todos os princípios éticos relevantes são totalmente incorporados no resultado final.

Capítulo 5

Conclusão

Neste trabalho, foi realizado uma revisão de literatura sobre ética em IA para investigar os princípios éticos e qual a relação deles com a Engenharia de Requisitos e como os princípios éticos são implementados durante o ciclo de desenvolvimento de software. Esse estudo também propõe a ferramenta *Requirements to US*, que utiliza histórias de usuário, princípios éticos em IA e modelos de LLMs para promover a integração da ética em IA durante a fase de Engenharia de Requisitos. Na construção da ferramenta foi utilizada a metodologia *Design Science Research* para entender o problema, desenhar o projeto piloto, desenvolver o protótipo e, posteriormente, avaliá-lo.

Na primeira etapa deste trabalho, foi realizada uma atualização da revisão sistemática de literatura publicada por Cerqueira et al. [16] para entender o estado atual da ética em IA. Para verificar a viabilidade de uma atualização, foi utilizado o *framework* proposto por Garnet et al. [7], que apontou para a necessidade de uma atualização da revisão. O protocolo foi definido com base no estudo de Kitchenham et al. [51], e, após sua aplicação, foram selecionados 38 estudos primários.

Como resultado, foi verificado um aumento do número de princípios éticos definidos na literatura, de 14, previamente identificados por Cerqueira [1], para 26. Por outro lado, a análise e validação de princípios e orientações parece ter sido efetuada em detrimento de novas propostas, o que pode indicar uma potencial mudança do foco para estudos futuros. Foi verificado também que os métodos, processos e sugestões propostos na literatura são extremamente diversificados em termos do tipo de abordagem, da fase do ciclo de vida do software e dos seus objetivos. Isto indica que, apesar das atuais limitações da ética da IA, o campo está a desenvolver e a abordar um maior número de soluções para mitigar os riscos e os problemas éticos presentes nos sistemas de IA.

Na segunda e terceira etapa, foram realizadas as especificações da arquitetura e funcionalidades da ferramenta para automatizar a geração de histórias éticas de usuários a partir de uma base de dados de requisitos éticos. Após a definição do *dataset*, os dados

foram submetidos a um *pipeline* de pré-processamento, do qual os caracteres foram transformados em letras minúsculas, os espaços redundantes foram removidos e os caracteres de entrada foram tokenizados. O processo de treinamento foi feito utilizando o principio de *fine-tuning*, definido como um segundo treino supervisionado utilizando um conjunto de dados extremamente reduzido para contextualizar as previsões do modelo [37]. Para permitir o processo de *fine-tuning*, foi utilizado o método QLORA, composto por um conjunto de três componentes que servem para reduzir o tamanho do modelo, facilitando assim o seu treino em máquinas com uma capacidade limitada de memória vídeo[105]. Após o processo de *fine-tuning*, foi realizada uma avaliação comparativa dos resultados preditos entre os modelos: Falcon40B [99], Falcon7B [99], BERT [100], RoBERTa [101] e Mistral7B [102]. O modelo que apresentou o melhor desempenho, Mitral7B, foi selecionado para uso. A integração de RAG permitiu que os modelos acessassem e utilizassem fontes externas atualizadas e relevantes, enriquecendo as narrativas e garantindo mais coerência nas predições, gerando histórias éticas de usuário mais completas, e não apenas descrições. Além disso, a recuperação de informações ajudou a reduzir alucinações, proporcionando respostas mais precisas e justificadas. Essa abordagem garantiu que as histórias fossem mais relevantes, precisas e alinhadas aos princípios éticos estabelecidos, tal como as diretrizes utilizadas para o banco de dados vetorial, aumentando a confiança nas predições do modelo.

A validação da ferramenta demonstrou resultados promissores ao demonstrar a sua capacidade de traduzir requisitos éticos abstratos em histórias éticas de usuário claras e alinhadas com os princípios éticos estabelecidos na literatura. A análise quantitativa e qualitativa realizada com os participantes indicou uma resposta positiva quanto à clareza, aplicabilidade e relevância das histórias éticas de usuário geradas. Estes resultados confirmam a viabilidade da utilização da ferramenta como um suporte efetivo para o desenvolvimento de sistemas baseados em inteligência artificial, mais especificamente nas fases iniciais do ciclo de vida do software.

No entanto, algumas limitações observadas durante a validação apontam para oportunidades de melhoria e trabalhos futuros. A diversidade limitada dos perfis dos participantes e o tamanho reduzido da amostra restringem a generalização dos resultados obtidos. Estudos futuros podem ampliar esta abordagem, incluindo profissionais de diferentes setores e níveis de experiência, de modo a avaliar a ferramenta em contextos mais amplos e diversos.

Outra perspectiva relevante para trabalhos futuros é a integração contínua da ferramenta em processos de desenvolvimento ágil. A aplicação em ambientes reais e a análise de projetos poderão evidenciar como a geração automatizada de histórias de usuário afeta a qualidade do software desenvolvido e a dinâmica interna de equipes de desenvolvimento.

Além disso, a realização de experimentos comparativos com outras abordagens de engenharia de requisitos poderá fortalecer ainda mais a aplicabilidade e a eficiência da solução.

Para aprimorar a ferramenta, recomenda-se a integração de modelos de linguagem mais avançados e a realização de treinamentos contínuos para melhorar a precisão das histórias geradas. A personalização de suas funcionalidades com base em contextos específicos de aplicação também pode aumentar sua adaptabilidade, principalmente devido a facilidade dada pelo RAG. Além disso, o desenvolvimento de uma interface interativa pode facilitar seu uso por equipes de desenvolvedores. Assim, este trabalho oferece uma base sólida para pesquisas futuras, incentivando avanços contínuos na busca por uma tecnologia mais ética e humanizada.

Referências Bibliográficas

[1] Cerqueira, J.: *Exploring Ethical Requirements Elicitation for Applications in the Context of AI.* Tese de Mestrado, Universidade de Brasília, 2021. xi, 1, 4, 6, 7, 11, 12, 32, 33, 34, 37, 39, 53, 67, 74, 75, 76, 77, 104

[2] Mazzeschi, M.: *Machine Learning vs. Deep Learning.* https://pub.towardsai.net/machine-learning-vs-deep-learning-783a87e00126, 2021. Accessado: 13/07/2024. xi, 13

[3] Lecun, Y., L. Bottou, Y. Bengio e P. Haffner: *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 86(11):2278–2324, 1998. xi, 14

[4] Geeks, Geeks for: *Introduction to Recurrent Neural Network.* https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/, 2024. Accessado: 19/06/2024. xi, 15

[5] Zhang, A., Z. C. Lipton, M. Li e A. J. Smola: *Dive into Deep Learning.* arXiv preprint arXiv:2106.11342, 2021. xi, 16

[6] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser e I. Polosukhin: *Attention is All you Need.* Em Guyon, I., U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan e R. Garnett (editores): *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. xi, 16, 17, 18

[7] Garner, P., S. Hopewell, J. Chandler, H. MacLehose, E. A. Akl, J. Beyene, S. Chang, R. Churchill, K. Dearness, G. Guyatt, C. Lefebvre, B. Liles, R. Marshall, L. M. García, C. Mavergames, M. Nasser, A. Qaseem, M. Sampson, K. Soares-Weiser, Y Takwoingi, L. Thabane, M. Trivella, P. Tugwell, E. Welsh, E. C. Wilson e H. J. Schünemann: *When and how to update systematic reviews: consensus and checklist.* BMJ, 2016. https://www.bmj.com/content/354/bmj.i3507. xi, 32, 33, 34, 35, 104

[8] Halme, E., M. Jantunen, V. Vakkuri, K. Kemell e P. Abrahamsson: *Making ethics practical: User stories as a way of implementing ethical consideration in Software Engineering.* Information and Software Technology, 2023. https://www.sciencedirect.com/science/article/abs/pii/S0950584923002343. xi, 5, 30, 31, 35, 80, 81, 82, 102

[9] Corrêa, N. K., C. Galvão, J. W. Santos, C. Del Pino, E. P. Pinto, C. Barbosa, D. Massmann, R. Mambrini, L. Galvão, E. Terem e N. de Oliveira: *Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance.* Patterns, 4(10):100857, 2023, ISSN 2666-3899. https://www.sciencedirect.com/science/article/pii/S2666389923002416. 1, 2, 51, 52, 53, 54, 55, 75, 78

[10] Jobin, A., M. Ienca e E. Vayena: *The global landscape of AI ethics guidelines.* Nature Machine Intelligence, 1(9):389–399, Sep 2019, ISSN 2522-5839. https://doi.org/10.1038/s42256-019-0088-2. 1, 2, 19

[11] Vakkuri, V., K. Kemell, J. Kultanen e P. Abrahamsson: *The Current State of Industrial Practice in Artificial Intelligence Ethics.* IEEE Software, 37(4):50–57, 2020. 1

[12] Angwin, J., J. Larson, S. Mattu e L. Kirchner: *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.*, 2016. 1, 51

[13] Marchal, N., R. Xu, R. Elasmar, I. Gabriel, B. Goldberg e W. Isaac: *Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data*, 2024. https://arxiv.org/abs/2406.13843. 1

[14] Masood, M., M. Nawaz, K. M. Malik, A. Javed e A. Irtaza: *Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward*, 2021. https://arxiv.org/abs/2103.00484. 1

[15] Cerqueira, J., H. Tives e E. Canedo: *Ethical Guidelines and Principles in the Context of Artificial Intelligence.* Em *Proceedings of the XVII Brazilian Symposium on Information Systems*, SBSI '21, New York, NY, USA, 2021. Association for Computing Machinery, ISBN 9781450384919. https://doi.org/10.1145/3466933.3466969. 2, 37

[16] Cerqueira, J., A. Azevedo, H. Tives e E. Canedo: *Guide for Artificial Intelligence Ethical Requirements Elicitation - RE4AI Ethical Guide.* Em *55th Hawaii International Conference on System Sciences, HICSS 2022, Virtual Event / Maui, Hawaii, USA, January 4-7, 2022*, páginas 1–10. ScholarSpace, 2022. http://hdl.handle.net/10125/80015. 2, 19, 30, 31, 104

[17] Dictionary, Cambridge: *Process.* Em *Cambridge Dictionary.* Cambridge University Press & Assessment, 2024. https://dictionary.cambridge.org/us/dictionary/english/process. 3

[18] Reiss, S. P.: *Software tools and environments.* ACM Comput. Surv., 28(1):281–284, março 1996, ISSN 0360-0300. https://doi.org/10.1145/234313.234423. 3

[19] Isman, A.: *Technology and technique: An educational perspective.* Turkish Online Journal of Educational Technology, 11:207–213, abril 2012. 3

[20] Cronholm, S. e P. J. Ågerfalk: *On the Concept of Method in Information Systems Development.* Em *Linköping Electronic Articles in computer and information science Vol. 4*, 1999. https://api.semanticscholar.org/CorpusID:14539129. 3

[21] Ayling, J. e A. Chapman: *Putting AI ethics to work: are the tools fit for purpose?* AI and Ethics, 2(3):405–429, Aug 2022, ISSN 2730-5961. https://doi.org/10.1007/s43681-021-00084-x. 3, 63, 66

[22] Morley, J., L. Floridi, L. Kinsey e A. Elhalal: *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, Aug 2020, ISSN 1471-5546. https://doi.org/10.1007/s11948-019-00165-5. 4

[23] Barletta, V. S., D. Caivano, D. Gigante e A. Ragone: *A Rapid Review of Responsible AI frameworks: How to guide the development of ethical AI.* Em *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, EASE '23. ACM, junho 2023. http://dx.doi.org/10.1145/3593434.3593478. 4, 69, 73, 74

[24] Tidjon, L. N. e F. Khomh: *The Different Faces of AI Ethics Across the World: A Principle-Implementation Gap Analysis*, 2022. 4, 68, 69, 73, 75

[25] Ruparelia, N. B.: *Software development lifecycle models.* SIGSOFT Softw. Eng. Notes, 35(3):8–13, maio 2010, ISSN 0163-5948. https://doi.org/10.1145/1764810.1764814. 5

[26] De Lucia, A., A. Qusef *et al.*: *Requirements engineering in agile software development.* Journal of emerging technologies in web intelligence, 2(3):212–220, 2010. 5

[27] Lucassen, G., F. Dalpiaz, J. M. Van der Werf e S. Brinkkemper: *The Use and Effectiveness of User Stories in Practice.* Em *Requirements Engineering: Foundation for Software Quality*, páginas 205–222, março 2016, ISBN 978-3-319-30281-2. 5, 29

[28] Vaishnavi, V. K. e W. Kuechler: *Design science research methods and patterns: innovating information and communication technology.* Crc Press, 2015. 7

[29] Dignum, V.: *Responsible Artificial Intelligence: how to Develop and Use AI in a Responsible Way.* Springer Cham, 2020. 10, 11, 12

[30] Russell, S. J., P. Norvig e E. Davis: *Artificial Intelligence: A Modern Approach.* Prentice Hall series in artificial intelligence. Prentice Hall, 3a edição, 2010, ISBN 9780136042594. https://books.google.com.br/books?id=8jZBksh-bUMC. 10, 11, 12

[31] Li, H.: *Proofs for the Four Fundamental Equations of the Backpropagation and Algorithms in Feedforward Neural Networks*, outubro 2023. 13

[32] Chung, H., S. J. Lee e J. G. Park: *Deep neural network using trainable activation functions.* Em *2016 International Joint Conference on Neural Networks (IJCNN)*, páginas 348–352, 2016. 13

[33] Shrestha, A. e A. Mahmood: *Review of Deep Learning Algorithms and Architectures.* IEEE Access, 7:53040–53065, 2019. 14

[34] Chowdhury, G. G.: *Natural language processing*. Annual Review of Information Science and Technology, 37(1):51–89, janeiro 2003, ISSN 0066-4200. 15

[35] Salehinejad, H., S. Sankar, J. Barfett, E. Colak e S. Valaee: *Recent Advances in Recurrent Neural Networks*, 2018. 15, 16

[36] Le, Q. V., N. Jaitly e G. E. Hinton: *A Simple Way to Initialize Recurrent Networks of Rectified Linear Units*, 2015. 16

[37] Douglas, M. R.: *Large Language Models*, 2023. 18, 79, 105

[38] Kemell, K., V. Vakkuri e F. Sohrab: *How Do AI Ethics Principles Work? From Process to Product Point of View*. Em Rantanen, Minna M., Salla Westerstrand, Otto Sahlgren e Jani Koskinen (editores): *Conference on Technology Ethics – Tethics, October 18–19, 2023, Turku, Finland.*, CEUR Workshop Proceedings, Saksa, 2023. CEUR-WS.org. Conference on Technology Ethics, Tethics ; Conference date: 18-10-2023 Through 19-10-2023. 19, 50, 54, 55, 69, 73, 75, 78

[39] Ryan, M. e B. C. Stahl: *Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications*. J. Inf. Commun. Ethics Soc., 19:61–86, 2020. https://api.semanticscholar.org/CorpusID:219512435. 19, 20, 21, 22, 23, 24, 25, 26, 27, 28

[40] Bozyiğit, F., Ö. Aktaş e D. Kılınç: *Linking software requirements and conceptual models: A systematic literature review*. Engineering Science and Technology, an International Journal, 24(1):71–82, 2021, ISSN 2215-0986. https://www.sciencedirect.com/science/article/pii/S2215098620342580. 29

[41] Escalona, M. J. e N. Koch: *Requirements engineering for web applications: a comparative study*. J. Web Eng., 2(3):193–212, feb 2003, ISSN 1540-9589. 29

[42] Kassab, M.: *The changing landscape of requirements engineering practices over the past decade*. Em *2015 IEEE Fifth International Workshop on Empirical Requirements Engineering (EmpiRE)*, páginas 1–8, 2015. 29

[43] Guizzardi, R., G. Amaral, G. Guizzardi e J. Mylopoulos: *Ethical Requirements for AI Systems*. Em Goutte, Cyril e Xiaodan Zhu (editores): *Advances in Artificial Intelligence*, páginas 251–256, Cham, 2020. Springer International Publishing, ISBN 978-3-030-47358-7. 29

[44] Agbese, M., R. Mohanani, A. A. Khan e P. Abrahamsson: *Implementing AI Ethics: Making Sense of the Ethical Requirements*. Em *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, EASE '23, página 62–71, New York, NY, USA, 2023. Association for Computing Machinery, ISBN 9798400700446. https://doi.org/10.1145/3593434.3593453. 29, 35, 37, 50, 54, 55, 75, 78

[45] Vakkuri, V., K. Kemell, M. Jantunen, E. Halme e P. Abrahamsson: *ECCOLA — A method for implementing ethically aligned AI systems*. Journal of Systems and Software, 182:111067, 2021, ISSN 0164-1212. https://www.sciencedirect.com/science/article/pii/S0164121221001643. 29, 30, 31, 55, 66, 67, 70, 73, 74

[46] Luitel, D., S. Hassani e M. Sabetzadeh: *Improving requirements completeness: automated assistance through large language models*. Requirements Engineering, 29(1):73–95, Mar 2024, ISSN 1432-010X. https://doi.org/10.1007/s00766-024-00416-3. 30, 31, 102

[47] Kitchenham, B. A., P. Brereton, D. Budgen, M. Turner, J. Bailey e S. G. Linkman: *Systematic literature reviews in software engineering - A systematic literature review*. Inf. Softw. Technol., 51(1):7–15, 2009. 32, 76

[48] Kitchenham, B. A. e S. Charters: *Guidelines for performing systematic literature reviews in software engineering*. Keele University, UK, 9:1–65, 2007. 32, 76

[49] Mendes, E., C. Wohlin, K. Felizardo e M. Kalinowski: *When to update systematic literature reviews in software engineering*. Journal of Systems and Software, 167:110607, 2020, ISSN 0164-1212. https://www.sciencedirect.com/science/article/pii/S0164121220300856. 33

[50] Ferrara, C., F. Casillo, C. Gravino, A. De Lucia e F. Palomba: *ReFAIR: Toward a Context-Aware Recommender for Fairness Requirements Engineering*. ICSE 2024, 2023. https://fpalomba.github.io/pdf/Conferencs/C81.pdf. 35

[51] Kitchenham, B. A., P. Brereton e D. Budgen: *Evidence-based Software Engineering and Systematic Reviews*. CRC Press, 2016. 36, 37, 39, 40, 41, 42, 43, 104

[52] Brereton, P., B. A Kitchenham, D. Budgen, M. Turner e M. Khalil: *Lessons from applying the systematic literature review process within the software engineering domain*. J. Syst. Softw., 80(4):571–583, 2007. https://doi.org/10.1016/j.jss.2006.07.009. 37

[53] Vakkuri, V., K. Kemell, J. Tolvanen, M. Jantunen, E. Halme e P. Abrahamsson: *How Do Software Companies Deal with Artificial Intelligence Ethics? A Gap Analysis*. Em *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*, EASE '22, página 100–109, New York, NY, USA, 2022. Association for Computing Machinery, ISBN 9781450396134. https://doi.org/10.1145/3530019.3530030. 37

[54] Mendes, F.: *Insights from personality and decision-making in software engineering context*. Tese de Doutoramento, University of Oulu, 2021. https://oulurepo.oulu.fi/handle/10024/36818. 40

[55] Shaw, M.: *Writing good software engineering research papers*. Em *25th International Conference on Software Engineering, 2003. Proceedings.*, páginas 726–736, 2003. 42

[56] Bourque, P., R. Fairley e IEEE Computer Society: *Guide to the Software Engineering Body of Knowledge (SWEBOK(R)): Version 3.0*. IEEE Computer Society Press, Washington, DC, USA, 3a edição, 2014, ISBN 0769551661. 42, 45

[57] Rees, C. e B. Muller: *All that glitters is not gold: trustworthy and ethical AI principles*. AI Ethics 3, 1241–1254 (2023), 2022. 49, 53, 54, 64, 66, 75, 78

[58] Khan, A. A., S. Badshah, P. Liang, B. Khan, M. Waseem, M. Niazi e M. A. Akbar: *Ethics of AI: A Systematic Literature Review of Principles and Challenges*, 2021. 49, 52, 53, 54, 55, 75, 78

[59] Vakkuri, V., K. Kemell, J. Tolvanen, M. Jantunen, E. Halme e P. Abrahamsson: *How Do Software Companies Deal with Artificial Intelligence Ethics? A Gap Analysis*. Em *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*, EASE '22, página 100–109, New York, NY, USA, 2022. Association for Computing Machinery, ISBN 9781450396134. https://doi.org/10.1145/3530019.3530030. 49, 52, 53, 54, 55

[60] Balasubramaniam, N., M. Kauppinen, A. Rannisto, K. Hiekkanen e S. Kujala: *Transparency and explainability of AI systems: From ethical guidelines to requirements*. Information and Software Technology, 159:107197, 2023, ISSN 0950-5849. https://www.sciencedirect.com/science/article/pii/S0950584923000514. 49, 52, 53, 54, 71, 73, 75, 78

[61] Vainio-Pekka, H., M. Ori Otse Agbese, M. Jantunen, V. Vakkuri, T. Mikkonen, R. Rousi e P. Abrahamsson: *The Role of Explainable AI in the Research Field of AI Ethics*. ACM Trans. Interact. Intell. Syst., 13(4), dec 2023, ISSN 2160-6455. https://doi.org/10.1145/3599974. 51, 52, 53, 54, 55

[62] Rousi, R., V. Vakkuri, P. Daubaris, S. Linkola, H. Samani, N. Mäkitalo, E. Halme, M. Agbese, R. Mohanani, T. Mikkonen e P. Abrahamsson: *Beyond 100 Ethical Concerns in the Development of Robot-to-Robot Cooperation*. Em *2022 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, páginas 420–426, 2022. 55, 66

[63] Antikainen, J., M. Agbese, H. Alanen, E. Halme, H. Isomaki, M. Jantunen, K. Kemell, R. Rousi, H. Vainio-Pekka e V. Vakkuri: *A Deployment Model to Extend Ethically Aligned AI Implementation Method ECCOLA*. Em *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE, setembro 2021. http://dx.doi.org/10.1109/REW53955.2021.00043. 56, 66, 67

[64] Agbese, M., H. Alanen, J. Antikainen, E. Halme, H. Isomaki, M. Jantunen, K. Kemell, R. Rousi, H. Vainio-Pekka e V. Vakkuri: *Governance of Ethical and Trustworthy Al Systems: Research Gaps in the ECCOLA Method*. Em *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE, setembro 2021. http://dx.doi.org/10.1109/REW53955.2021.00042. 56, 66, 67

[65] Agbese, M., H. Alanen, J. Antikainen, E. Halme, H. Isomaki, M. Jantunen, K. Kemell, Rebekah Rousi, Heidi Vainio-Pekka e V. Vakkuri: *Governance in Ethical and Trustworthy AI Systems: Extension of the ECCOLA Method for AI Ethics Governance Using GARP*. e-Informatica Software Engineering Journal, 17(1):230101, setembro 2023. https://www.e-informatyka.pl/index.php/einformatica/volumes/volume-2023/issue-1/article-1/, Available online: 28 Sep. 2022. 56, 66, 67

[66] Nitta, I., K. Ohashi, S. Shiga e S. Onodera: *AI Ethics Impact Assessment based on Requirement Engineering*. Em *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, páginas 152–161, 2022. 56, 66

[67] Boyd, K.: *Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development*. Em *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, página 2069–2082, New York, NY, USA, 2022. Association for Computing Machinery, ISBN 9781450393522. https://doi.org/10.1145/3531146.3534626. 57, 66, 67, 70, 73, 74

[68] Bruschi, D. e N. Diomede: *A framework for assessing AI ethics with applications to cybersecurity*. AI and Ethics, 3(1):65–72, Feb 2023, ISSN 2730-5961. https://doi.org/10.1007/s43681-022-00162-8. 57, 66

[69] Agbese, M., R. Mohanani, A. A. Khan e P. Abrahamsson: *Ethical Requirements Stack: A framework for implementing ethical requirements of AI in software engineering practices*. Em *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, EASE '23, página 326–328, New York, NY, USA, 2023. Association for Computing Machinery, ISBN 9798400700446. https://doi.org/10.1145/3593434.3593489. 58, 66, 70, 73, 74

[70] Agbese, M., M. Rintamaki, R. Mohanani e P. Abrahamsson: *Implementing AI Ethics in a Software Engineering Project-Based Learning Environment - The Case of WIMMA Lab*. Em Carroll, Noel, Anh Nguyen-Duc, Xiaofeng Wang e Viktoria Stray (editores): *Software Business*, páginas 278–284, Cham, 2022. Springer International Publishing, ISBN 978-3-031-20706-8. 58, 66

[71] Sanderson, C., D. Douglas, Q. Lu, E. Schleiger, J. Whittle, J. Lacey, G. Newnham, S. Hajkowicz, C. Robinson e D. Hansen: *AI Ethics Principles in Practice: Perspectives of Designers and Developers*. IEEE Transactions on Technology and Society, 4(2):171–187, junho 2023, ISSN 2637-6415. http://dx.doi.org/10.1109/TTS.2023.3257303. 58, 66

[72] Kemell, K. e V. Vakkuri: *What Is the Cost of AI Ethics? Initial Conceptual Framework and Empirical Insights*. Em Hyrynsalmi, Sami, Jürgen Münch, Kari Smolander e Jorge Melegati (editores): *Software Business*, páginas 247–262, Cham, 2024. Springer Nature Switzerland, ISBN 978-3-031-53227-6. 58, 66

[73] Ciobanu, A. C. e G. Meșniță: *AI Ethics for Industry 5.0 – From Principles to Practice*, 2022. 59, 66, 72, 73, 74

[74] Gerdes, A.: *A participatory data-centric approach to AI Ethics by Design*. Applied Artificial Intelligence, 36(1):2009222, 2022. https://doi.org/10.1080/08839514.2021.2009222. 59, 60, 66, 67

[75] Kelleher, J. D. e B. Tierney: *Data Science*. The MIT Press, 2018. 59, 75, 79

[76] Mäntymäki, M., M. Minkkinen, T. Birkstedt e M. Viljanen: *Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance*, 2023. 60, 66

[77] Lachenmaier, J. F., M. Werling e D. Morar: *Governance of Artificial Intelligence – A Framework Towards Ethical AI Applications*, 2023. https://ceur-ws.org/Vol-3630/LWDA2023-paper6.pdf. 60, 66, 67

[78] Zicari, R. V., J. Brodersen, J. Brusseau, B. Düdder, T. Eichhorn, T. Ivanov, G. Kararigas, P. Kringen, M. McCullough, F. Möslein, N. Mushtaq, G. Roig, N. Stürtz, K. Tolle, J. J. Tithi, I. van Halem e M. Westerlund: *Z-Inspection®: A Process to Assess Trustworthy AI*. IEEE Transactions on Technology and Society, 2(2):83–97, 2021. 60, 61, 62, 66, 67, 76

[79] Agarwal, A., H. Agarwal e N. Agarwal: *Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems*. AI and Ethics, 3(1):267–279, março 2022, ISSN 2730-5961. http://dx.doi.org/10.1007/s43681-022-00147-7. 62, 66

[80] Shu, Y., J. Zhang e H. Yu: *Fairness in Design: A Tool for Guidance in Ethical Artificial Intelligence Design*. Em Meiselwitz, Gabriele (editor): *Social Computing and Social Media: Experience Design and Social Network Analysis*, páginas 500–510, Cham, 2021. Springer International Publishing, ISBN 978-3-030-77626-8. 63, 66, 71, 73, 76

[81] Squadrone, L., D. Croce e R. Basili: *Ethics by Design for Intelligent and Sustainable Adaptive Systems*. Em Dovier, Agostino, Angelo Montanari e Andrea Orlandini (editores): *AIxIA 2022 – Advances in Artificial Intelligence*, páginas 154–167, Cham, 2023. Springer International Publishing, ISBN 978-3-031-27181-6. 64, 66, 67

[82] Schmid, A. e M. Wiesche: *The Importance of an Ethical Framework for Trust Calibration in AI*. IEEE Intelligent Systems, 38(6):27–34, 2023. 64, 66

[83] Lu, Q., L. Zhu, X. Xu, J. Whittle, D. Zowghi e A. Jacquet: *Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering*, 2023. 65, 66

[84] Georgieva, I., C. Lazo, T. Timan e A. F. van Veenstra: *From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience*. AI and Ethics, 2:1–15, janeiro 2022. 69, 73

[85] Lu, Q., L. Zhu, X. Xu, J. Whittle e Z. Xing: *Towards a Roadmap on Software Engineering for Responsible AI*, 2022. 70, 73, 74

[86] Prem, E.: *From ethical AI frameworks to tools: a review of approaches*. AI and Ethics, 3(3):699–716, Aug 2023, ISSN 2730-5961. https://doi.org/10.1007/s43681-023-00258-9. 72, 73, 74

[87] Wohlin, C., P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell e A. Wesslén: *Planning*, páginas 89–116. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, ISBN 978-3-642-29044-2. https://doi.org/10.1007/978-3-642-29044-2_8. 76, 77

[88] Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel e D. Kiela: *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, 2021. https://arxiv.org/abs/2005.11401. 79, 89

[89] Gao, Y., Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang e H. Wang: *Retrieval-Augmented Generation for Large Language Models: A Survey*, 2024. https://arxiv.org/abs/2312.10997. 79, 89, 90

[90] Canedo, E. e B. Mendes: *Software Requirements Classification Using Machine Learning Algorithms*. Entropy, 22(9):1057, 2020. https://doi.org/10.3390/e22091057. 80, 81

[91] Halme, E., M. Agbese, H. Alanen, J. Antikainen, M. Jantunen, A. A. Khan, K. Kemell, V. Vakkuri e P. Abrahamsson: *Implementation of Ethically Aligned Design with Ethical User stories in SMART terminal Digitalization project: Use case Passenger Flow*. CoRR, abs/2111.06116, 2021. https://arxiv.org/abs/2111.06116. 80, 81, 82

[92] Ding, B., C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W Xia, J. Hu, A. T. Luu e S. Joty: *Data Augmentation using Large Language Models: Data Perspectives, Learning Paradigms and Challenges*, 2024. https://arxiv.org/abs/2403.02990. 82

[93] Haddi, E., X. Liu e Y. Shi: *The Role of Text Pre-processing in Sentiment Analysis*. Procedia Computer Science, 17:26–32, 2013, ISSN 1877-0509. https://www.sciencedirect.com/science/article/pii/S1877050913001385, First International Conference on Information Technology and Quantitative Management. 83

[94] Palomino, M. A. e F. Aider: *Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis*. Applied Sciences, 12(17), 2022, ISSN 2076-3417. https://www.mdpi.com/2076-3417/12/17/8765. 83, 84

[95] Angiani, G., L. Ferrari, T. Fontanini, P. Fornacciari, E. Iotti, F. Magliani e S. Manicardi: *A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter*. Em *International Workshop on Knowledge Discovery on the Web*, 2016. https://api.semanticscholar.org/CorpusID:8111009. 83, 84

[96] Toraman, C., E. H. Yilmaz, F. Şahinuç e O. Ozcelik: *Impact of Tokenization on Language Models: An Analysis for Turkish*. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(4):1–21, março 2023, ISSN 2375-4702. http://dx.doi.org/10.1145/3578707. 84

[97] Valdenegro-Toro, M. e M. Sabatelli: *Machine Learning Students Overfit to Overfitting*, 2022. https://arxiv.org/abs/2209.03032. 85

[98] Yadav, S. e S. Shukla: *Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification*. Em *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, páginas 78–83, 2016. 86

[99] Almazrouei, E., H. Alobeidli, A. Alshamsi, A. Cappelli, Ruxandra Cojocaru, Merouane D., E. Goffinet, D. Heslow, J. Launay, Q. Malartic, B. Noune, B. Pannier e G. Penedo: *Falcon-40B: an open large language model with state-of-the-art performance*, 2023. 86, 105

[100] Devlin, J., M. Chang, K. Lee e K. Toutanova: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019. https://arxiv.org/abs/1810.04805. 86, 105

[101] Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer e V. Stoyanov: *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019. https://arxiv.org/abs/1907.11692. 86, 105

[102] Q., Jiang. A., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix e W. El Sayed: *Mistral 7B*, 2023. https://arxiv.org/abs/2310.06825. 86, 87, 105

[103] Yu, T. e H. Zhu: *Hyper-Parameter Optimization: A Review of Algorithms and Applications*, 2020. https://arxiv.org/abs/2003.05689. 87

[104] Li, H., P. Chaudhari, H. Yang, M. Lam, A. Ravichandran, R. Bhotika e S. Soatto: *Rethinking the Hyperparameters for Fine-tuning*, 2020. https://arxiv.org/abs/2002.11770. 87

[105] Dettmers, T., A. Pagnoni, A. Holtzman e L. Zettlemoyer: *QLoRA: Efficient Finetuning of Quantized LLMs*, 2023. https://arxiv.org/abs/2305.14314. 88, 105

[106] Zhang, Z., M. Rayhan, T. Herda, M. Goisauf e P. Abrahamsson: *LLM-based agents for automating the enhancement of user story quality: An early report*, 2024. https://arxiv.org/abs/2403.09442. 102